E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP18;TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2026)04-1078-12

论文引用格式:Mao K Q, Yao J, Yu G S and Ding D D. 2026. Memory-speed-quality co-optimized real-time neural loop filtering for mobile devices. Journal of Image and Graphics, 31(4):1078-1089(毛可权, 姚杰, 余国生, 丁丹丹. 2026. 内存—速度—质量协同优化的移动端实时神经网络环路滤波方法. 中国图象图形学报, 31(4):1078-1089)[DOI:10.11834/jig.250324]

内存—速度—质量协同优化的移动端实时神经网络环路滤波方法

毛可权[1],姚杰[2],余国生[2],丁丹丹[1*]

1. 杭州师范大学信息科学与技术学院,杭州　311121; 2. 平头哥(上海)半导体技术有限公司,上海　201203

摘　要:**目的**　基于神经网络的方法在视频编码中展现出显著优势,极大提升了压缩效率。随着神经网络加速器成为移动设备标配,如何实现面向移动端的高效、轻量化部署成为亟待解决的问题。本文旨在提出一种既能提升编码效率,又具备高部署友好性的轻量级神经网络环路滤波方法,以解决现有模型在移动端部署中面临的高内存占用、推理速度低以及算子兼容性差等关键问题,从而推动神经网络编码技术在移动端的落地应用。**方法**　本文提出了一种基于结构重参数化与特征压缩融合模块的轻量级环路滤波方法。采用U-Net架构并结合重参数化卷积模块,在保证模型性能的同时大幅降低模型复杂度。此外,针对U-Net跳跃连接造成的内存瓶颈,设计了特征压缩融合模块,通过多尺度特征的空间对齐与通道压缩,生成更紧凑的特征表示,在降低内存占用的同时,增强滤波质量。**结果**　通过特征压缩融合模块,将模型峰值内存阶段的特征图内存占用降低93.75%,显著减少内存消耗。将所设计的滤波模型嵌入AV1(AOMedia Video 1)编码器中,实验结果表明,相比传统的AV1基准环路滤波方法,所提出的模型在帧内编码和帧间编码模式下分别获得了−5.09%和−4.32%的平均BD-Rate增益。进一步地,将所设计的模型部署在骁龙8 Gen1的神经网络处理器上,针对尺寸为1 920×1 080像素的高清视频,推理速度可达77帧/s,峰值内存仅55 MB。**结论**　通过降低模型复杂度和所设计的特征压缩融合模块,显著降低了内存占用,实现了推理速度以及重建质量的协同优化,能够满足移动端实时处理需求。

关键词:环路滤波;特征压缩;移动端部署;AV1;结构重参数化;神经网络模型优化;低内存占用;低延迟

Memory-speed-quality co-optimized real-time neural loop filtering for mobile devices

Mao Kequan[1], Yao Jie[2], Yu Guosheng[2], Ding Dandan[1*]

1. School of Information Science and Technology, Hangzhou Normal University, Hangzhou　311121, China;

2. T-Head Semiconductor Co., Ltd., Shanghai　201203, China

Abstract: **Objective** Driven by standards such as moving picture expert group (MPEG), H.26x, audio video coding standard (AVS), and AOMedia video (AVx), video coding technology has evolved significantly over the past 3 decades. These standards commonly adopt a block-based hybrid framework that improves compression efficiency through spatial/temporal prediction, adaptive entropy coding, and rate-distortion optimization. However, lossy compression inevitably introduces artifacts such as blurring, blocking, and ringing. Traditional encoders address these issues using hand-crafted fil-

收稿日期:2025-08-06;**修回日期**:2025-10-16;**预印本日期**:2025-10-23

*通信作者:丁丹丹　dandanding@hznu.edu.cn

基金项目:国家自然科学基金项目(62571174)

Supported by: National Natural Science Foundation of China(62571174)

ters, such as a deblocking filter, a sample adaptive offset, an adaptive loop filter, a constrained directional enhancement filter, and loop restoration. Deep neural networks, offering superior performance due to their nonlinear representation power, have been recently introduced to replace rule-based filters. However, existing models often require large parameter counts and substantial memory, thereby limiting their deployment on mobile devices. This work addresses these challenges by introducing a compact, deployment-friendly neural network loop filter tailored for mobile and edge environments. **Method** We propose a lightweight and deployment-friendly approach built on a U-Net backbone to tackle the challenges of high memory consumption, slow inference speed, and poor hardware compatibility in neural network-based loop filtering. The network adopts an encoder-decoder structure with progressive downsampling to reduce interblock complexity and overall computational cost. Structural reparameterization is employed to optimize the network further for deployment: During training, multibranch convolutions enhance feature representation capacity, whereas at inference, they are merged into single-branch standard convolutions to minimize operator dependencies, simplify execution, and reduce intermediate memory usage. A key challenge of U-Net architectures is the memory overhead caused by long skip connections, given that feature maps from each encoder stage must be stored for later use. We address this bottleneck by designing a multiscale feature fusion and compression module. First, encoder features from all stages are compressed via pointwise convolutions and spatially aligned at the smallest resolution using downsampling. Then, these features are concatenated along the channel dimension and undergo a second compression stage to generate a compact and expressive representation. In the decoder, this representation is reconstructed to its original resolution and further enhanced using reparameterized convolutional blocks. This design significantly reduces peak memory consumption while preserving feature fidelity and improving filtering quality, effectively achieving a joint optimization of memory footprint, inference speed, and reconstruction performance. The entire network is trained end-to-end by taking image residuals as input and using mean squared error as the loss function. Finally, the trained model is embedded into the AOMedia Video 1(AV1) encoder pipeline and combined with an R-D cost optimization strategy to select the optimal filtered frames adaptively during the loop filtering stage, thereby improving overall coding efficiency under real-world video coding scenarios. **Result** Extensive experiments conducted within the AV1encoding framework demonstrate that the proposed method achieves a −5.09% BD-rate (Bjøntegaard delta rate) with 77 frames/s at 1 080 P in intraframe prediction and a −4.32% BD-rate in interframe prediction, thereby providing a quantitative evaluation of its performance. In particular, under intraframe prediction, our model attains an average BD-rate reduction of −5.09%. This finding indicates that our model outperforms most competing methods, such as variable-filter-size residue learning convolutional neural network (VRCNN) (−3.48%), MobileNetV2 (−2.66%), and RepSR (reparameterization super resolution) (−3.81%). Moreover, it achieves −5.52% and −6.75% in ClassA1 (4 K) and ClassA3 (720 P), respectively, thereby demonstrating strong performance across both high- and mid-resolution content. Under interframe prediction, the proposed method maintains competitive performance with an average BD-rate reduction of −4.32%, closely matching the average BD-rate reduction of deep semi-smooth newton network (DeepSN-Net) (−4.35%) while exhibiting stable results across test sequences. For instance, in ClassA1, our model attains a larger gain (−4.61%) than all other baselines. In ClassA3, where DeepSN-Net achieves the highest gain (−6.23%), our model still performs comparably (−5.77%) with low fluctuation across resolutions. Moreover, the model's peak memory consumption during the M_{sc3} (memory of skip connection 3) stage is reduced from 27.69 MB to 1.73 MB, indicating a 93.75% decrease, by adopting the feature compression strategy with compression coefficients $\alpha_1 = \alpha_2 = \alpha_3 = \beta = 0.5$. This substantial reduction is critical for deployment on resource-constrained platforms. When deployed on a Snapdragon 8 Gen1 neural processing unit (NPU), the model processes 1 080 P video at 77 frames/s with a peak memory usage of only 55 MB. This finding confirms the model's suitability for real-time mobile inference under strict memory and latency constraints. **Conclusion** The lightweight neural network loop filtering method proposed in this work effectively balances compression efficiency, memory footprint, and inference speed, thereby delivering an integrated solution tailored for mobile and edge device deployment. Compared with the existing state-of-the-art neural network video coding models, our approach achieves competitive or superior coding gains while drastically reducing resource consumption, particularly in peak memory usage and inference latency. This combination of high performance and deployment friendliness makes the proposed method highly suitable for real-time video processing scenarios on mobile devices equipped with neural network accelerators. Our results validate the practical

viability of structurally reparameterized convolution modules combined with feature fusion compression in mitigating memory bottlenecks and accelerating inference without sacrificing reconstruction quality. This work provides a promising direction for advancing neural network video coding technologies toward widespread adoption in mobile and edge computing environments to enable enhanced video quality and efficient bandwidth utilization with low power and resource demands.

Key words: loop filtering; feature compression; mobile deployment; AOMedia Video 1 (AV1); structural reparameterization; neural network model optimization; low memory footprint; low latency

0 引 言

近30年,视频编码技术在MPEG(moving picture expert group)、H. 26x、AVS(audio video coding standard)和AVx(AOMedia video)等标准的推动下实现了跨越式发展。其技术框架普遍采用基于块结构的混合预测与变换编码机制,通过空间/时间相关性建模、多粒度预测、动态块分割、自适应熵编码及率失真优化,显著提升了压缩效率。然而,有损压缩必然导致视觉损失,如边界模糊、块失真及振铃伪影等问题。因此,视频编码器往往通过设计一些基于规则的滤波器,如去块效应滤波器(deblocking filter, DBF)(Norkin等,2012)、样本自适应偏移(sample adaptive offset, SAO)(Fu等,2012)、自适应环路滤波器(adaptive loop-filter, ALF)(Tsai等,2013)、约束方向性增强滤波器(constrained directional enhancement filter, CDEF)(Midtskogen和Valin,2018)和环路恢复滤波器(loop restoration, LR)(Mukherjee等,2017)等,尽可能地还原原始像素,提升重建视频质量。

近年来,基于深度神经网络的方法备受瞩目(贾川民 等,2021),其凭借强大的非线性表达能力,相较于传统基于规则的滤波方法,取得了显著的性能提升。当前工作一般需要训练拥有数十万甚至数百万参数的模型(Ma等,2021;Huang等,2020;Wang等,2021)拟合压缩噪声,在提升性能的同时,也带来了高计算复杂度。显然,这些模型难以在资源受限的移动平台上部署。例如,一种基于卷积神经网络(convolutional neural network, CNN)的环路滤波方法——EDCNN(enhanced deep convolutional neural network)(Pan等,2020),所占用的模型大小为18.2 MB,在推理时大约需要5 GB的运行内存,这对于现有的移动设备来说是不切实际的。

为适应实际工业场景的部署需求,研究人员通常采用剪枝(Liu等,2017)、聚类(Han等,2015)、低秩近似(Jaderberg等,2014)以及神经网络搜索(Zoph和Le,2017)等技术手段,通过模型压缩和架构优化来构建轻量化的CNN模型。这些措施能够在保证准确性的同时,减少每秒浮点运算次数(floating point operations per second,FLOPs)和参数量。例如,深度可分离卷积(Howard等,2017)能够在等效替换标准卷积的同时,降低计算量和参数量。然而,剪枝虽然能降低模型大小,但更细粒度的剪枝(如Filter或Kernel级时),可能因为硬件架构的限制,并不会加速网络推理;聚类、低秩近似等技术会导致更高的计算量。并且,这些指标与模型的延迟效率缺乏强相关性(Dehghani等,2021)。例如,FLOPs这样的指标并没有考虑内存访问成本和并行度,仅依赖于FLOPs指标无法实现对推理速度的判断(Ma等,2021)。

针对模型内存占用问题,U-Net架构通过编码器—解码器结构减少特征处理量,从而降低内存需求和计算量。然而,U-Net的长跳跃连接在编码器阶段需要完整保存特征图,限制了其在资源受限设备上的应用;去掉跳跃连接则会导致模型性能大幅下降。因此,如何在速度、内存和质量之间实现平衡成为移动端实时环路滤波的关键挑战。

为解决上述难题,本文提出了一种面向移动端的实时神经环路滤波方法,权衡内存占用、推理速度和滤波质量。模型内存—速度—质量三维对比图如图1所示。图1基于1 080 P、ClassA2数据集的Intra编码结果,横轴表示推理速度(帧率(frames per second, FPS)),纵轴为相对于AV1环路滤波方法的BD-Rate指标,圆形标记的大小与各方法峰值内存成正比。

本文工作的主要贡献是:1)提出了一种基于U-Net架构结合结构重参数化方法的轻量级环路滤

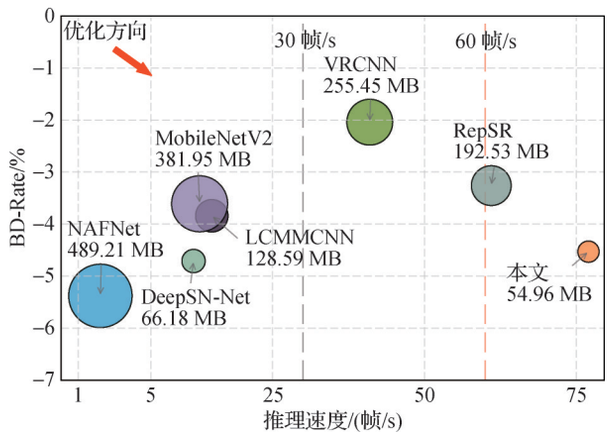

图 1　环路滤波网络的编码效率—复杂度关系

Fig. 1　Rate-complexity trade-off of loop filtering networks

波方法，实现 AV1（AOMedia Video 1）环路滤波功能，完成了内存—速度—质量的协同优化平衡；2）设计了一种通用的特征压缩融合模块，通过生成更紧凑的特征表示，有效缓解了 U-Net 架构中长跳跃连接造成的内存瓶颈问题；3）本文设计的模型相比 AV1 环路滤波基准方法，在帧内和帧间编码模式下获得了 −5.09% 和 −4.32% 的 BD-Rate（Bjøntegaard-Delta rate）增益。所提出的特征压缩融合模块替代了 U-Net 中长跳跃连接，在峰值内存阶段将特征图内存占用减少了 93.75%。将所设计的模型部署到骁龙 8Gen1 设备上，使用 NPU（neural processing unit）加速，针对 1 080 P 的高清视频，推理速度可达 77 帧/s，模型的峰值内存消耗仅 55 MB。

1　相关工作

1.1　基于神经网络的图像滤波方法

深度学习方法在加性高斯噪声去除方面取得了显著进展，但在应对图像中任意分布的复杂噪声时，表现仍不尽理想。随着神经网络在图像和视频压缩领域的广泛应用，学者们开始尝试将去噪网络用于压缩噪声抑制任务，Dai 等人（2016）提出的 VRCNN（variable-filter-size residue learning CNN）最早用于 H.265/HEVC（high efficiency video coding）环路滤波的去噪网络之一，采用双分支结构拓展网络宽度，以聚合更多信息。

随着模型结构日益复杂，深度不断增加，但去噪性能提升却逐渐趋于瓶颈。一些研究开始探索提升性能与降低复杂度的平衡路径。例如，有研究提出了一种简单而有效的两阶段图像去噪算法（丁岳皓等，2023），在降低模型复杂度的同时提升了对噪声的适应能力。Wang 等人（2022a）则提出 LCMMCNN（low complexity multi-model CNN）方法，使用多个小模型替代通用模型，通过多模型迭代训练机制，针对不同失真程度的视频内容构建多个小模型，每个模型由特定数量的残差块组成。尽管这些模型在各自的编码器中展现出显著性能，其核心手段仍是通过缩减模型的深度或宽度来降低复杂度，尚未从网络架构层面进行系统性优化。此外，多分支设计不仅可能因模块间同步等待降低并行效率，还会因缓存中间特征图导致内存占用增加，这对实时编解码场景构成了严峻挑战。

1.2　模型的轻量化

为满足实际应用需求，学者们开始探索神经网络模型的轻量化设计方法，以降低模型的 FLOPs。MobileNetV1（Howard 等，2017）是典型的方法之一，它提出了深度可分离卷积，将标准卷积分解为深度卷积和逐点卷积，显著降低了 FLOPs 和参数量，同时保持模型的原有性能。这一创新对轻量级网络架构设计有开创性意义，启发了后续诸多模型的结构范式演进。Chen 等人（2022）提出 NAFNet（nonlinear activation free network），进一步融合了 U-Net 的下采样结构与深度可分离卷积，在块间通过 U-Net 压缩特征分辨率降低计算量，在块内采用轻量卷积减少单模块复杂度，最终以更低的 FLOPs 实现了更优的滤波性能。利铭康等人（2025）探索了 Transformer 与 CNN 的结合方式，成功设计出一类 FLOPs 大幅减少而性能几乎不损的网络架构。此外，Deng 等人（2025）提出 DeepSN-Net（deep semi-smooth newton network），在传统半光滑牛顿方法的基础上进行了改进，用于求解图像逆问题，并据此设计了一种可解释的多任务通用盲图像复原网络。其轻量化版本在输入阶段对图像进行 4 倍下采样，并采用轻量卷积模块进行网络构建，从而显著降低了 FLOPs。

然而，上述网络优化策略过度聚焦于 FLOPs 指标，忽视了内存访问成本、并行度等实际部署中的关键约束。这种优化偏差在包含大量逐元素算子的网络中尤为明显，例如 LayerNorm 中的仿射变换、逐点卷积中的通道线性变换、SimpleGate 中用于替代非线性激活函数的逐元素乘法（Chen 等，2022）以及残差结构中的逐元素加法等算子，虽然在理论上具有

较低的计算复杂度，但由于缺乏数据局部性，无法通过缓存复用数据，导致内存访问成本高，计算强度（衡量内存带宽对计算速度的制约程度）较低，这类问题往往难以优化。因此，实际运行速度仍然受限。

1.3 硬件友好的模型架构

为了进一步加速硬件推理、降低内存占用，学者们研究了硬件友好的模型架构设计。为降低以残差结构为代表的多分支拓扑导致的高内存消耗，研究人员提出将跳跃连接或其他线性操作合并到单个线性分支中。例如，Ding等人（2021）引入了结构重参数化技术，通过将线性多分支拓扑转换为等效的单路径结构，实现了模型的高效推理，这种方式显著减少块内复杂度（单个神经网络模块内部的计算复杂度和资源消耗）。

基于该范式，ECBSR（edge-oriented convolution block for real-time super resolution）（Zhang等，2021）、RepSR（re-parameterization super resolution）（Wang等，2022b）等网络分别针对超分辨率任务特性，对多分支重参数化模块进行适应性改进，构建了适用于移动端实时推理的网络架构。不过，为了保持图像细节，这些方法都直接在原分辨率图像上进行处理，仍面临较大的计算资源压力。

相较而言，U-Net具有多层下采样，降低了块间复杂度（模块之间的连接方式、数据流动和依赖关系引入的额外开销），并构建了多尺度特征表达。这种优势使其成为资源受限场景的理想选择。因此，本文基于U-Net架构探索基于神经网络的轻量级环路滤波方法。然而，现有结构重参数化技术虽然解决了局部残差连接的合并问题，却无法消除U-Net架构存在的长跳跃连接。这是因为U-Net中的长跳跃连接主要用于连接编解码器，其主干包含了多个非线性变换，无法通过简单的线性等效转换实现分支合并，直接从U-Net中删除此类跳跃连接可能会导致严重的准确性损失。因此，在保持模型精度的前提下有效降低U-Net长跳跃连接的内存开销，是本研究面临的关键挑战之一。

2 本文工作

2.1 基于率失真优化的环路滤波决策模型

AV1编码器主要采用了以下环路滤波技术：去块效应滤波（DBF）、约束方向增强滤波（CDEF）、跨分量样本补偿（cross component sample offset，CCSO）（Du等，2021）以及环路恢复滤波（LR）。

图2展示了应用本文提出的环路滤波方案后的AV1编码器环路滤波架构。通过帧级的模式选择，编码器可以在基于CNN的滤波器和环路恢复滤波器之间进行切换，动态选择最优的滤波方案。

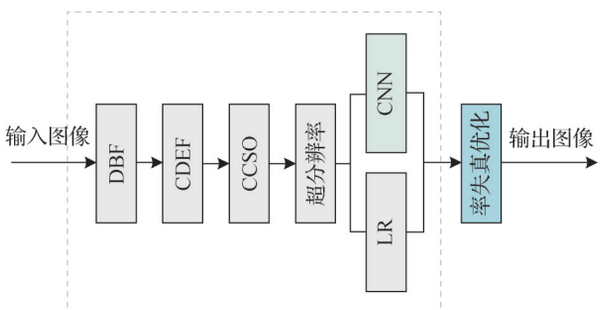

图2 基于率失真优化的AV1环路滤波流程

Fig. 2 RD-optimized loop filtering flow in AV1

由于增加了新的环路滤波方法，需要分别计算LR的率失真代价J_{LR}以及CNN滤波方法的率失真代价J_{CNN}，编码器将自动选择率失真代价较小的方法，作为当前帧的最佳滤波工具。率失真代价计算方法为

$$J = f_{\mathrm{MSE}}(\boldsymbol{x}, \boldsymbol{y}) + \lambda \cdot R_f \tag{1}$$

式中，R_f表示应用不同环路滤波工具时，编码相关参数所需的比特数。λ为拉格朗日乘数，代表了在既定程度的视频质量下，码率代价与视频质量之间的关系。使用均方误差（mean squared error，MSE）作为失真评价指标，其中，f_{MSE}为求均方误差，$\boldsymbol{x}$和$\boldsymbol{y}$分别表示滤波后图像和原图像。

2.2 低内存实时环路滤波模型

为了构建适用于硬件约束的低内存实时环路滤波模型，本研究结合架构特性设计了以下准则，并根据此准则构建了图3（a）所示的网络架构。

1）采用逐级下采样—上采样的U-Net架构，以降低块间复杂度并减少处理数据量，从而提升整体吞吐率。与常见的深度可分离卷积不同，本研究选用标准卷积作为基本算子。标准卷积可受益于Winograd（Lavin和Gray，2016）等优化算法进行加速。而深度可分离卷积通过拆分空间维度和通道维度的相关性，减少了卷积计算所需要的参数个数，并

在一些研究中被证实提升了卷积核参数的使用效率(Chollet,2017)。然而,其中的深度卷积由于每个通道独立计算,计算更加碎片化;逐点卷积由于核尺寸为 1 × 1,Winograd 算法对此类操作的优化空间有限,使其计算效率相比标准卷积更低。

2)在使用标准卷积的基础上,引入结构重参数化技术,将训练和推理阶段的模块结构解耦:训练时采用多分支、过参数化卷积结构以提升特征表达能力;推理时将其融合为单分支标准卷积,从而在不增加推理复杂度的情况下获得额外性能提升(Ding 等,2021;Wang 等,2022b)。同时,尽量减少块内多分支结构,以避免推理过程中因并行分支同步等待而导致计算资源利用率下降,并减少中间特征缓存以降低内存压力和访存延迟,从而提升整体推理效率。

3)为解决 U-Net 架构的长跳跃连接导致的高内存占用,提出特征压缩融合模块。在编码器阶段,将各层级特征图下采样至最小分辨率,并进行通道融合与压缩,生成紧凑特征表示;在解码器阶段再将其逐步重建回目标尺度,实现特征融合。该设计显著降低了推理过程中保存的中间特征图峰值内存消耗,在保证重建质量的同时减少内存消耗,其具体过程如图 3(b)示意。

4)减少逐元素操作,此类算子属于带宽受限算子,需要通过向量化数据访问的方式提升算子的性能,且无法有效利用局部性原理,就推理效率而言低于计算密集型的算子。因此,除了 ReLU(rectified linear unit)等必要的激活函数以外,本文模型规避了此类算子的使用。

根据以上 4 个准则,本文使用图 3(c)的过参数卷积块(RepConvBlock)作为基本结构,并在推理时将其重参数为标准卷积。

5)通道数设计遵循 2 的幂次准则(如 16/32/64),这是因为当前硬件架构的计算库大多对 2 的幂次准则的维度张量操作实现了内存对齐优化和指令级加速,可显著提升算子执行效率。

6)不使用倒置残差结构。该结构由 MobileNetV2(Sandler 等,2018)提出,其瓶颈层通过将通道数扩展至输入通道的 t 倍(扩展因子),这种通道扩张机制与残差结构相叠加,共同加剧了内存占用的增长。并且,当扩展因子 t 为非整数时,将违反准则 5。

图 3 中的 H、W、C 分别表示对应特征图的高宽和通道数;α_i 和 β 分别表示各阶段的压缩系数;BN 为 batchnorm。

2.3 特征压缩融合模块

本文的网络架构采用对称的编码器—解码器的 U-Net 结构,各包含 3 个阶段。各个阶段输出的特征图内存占用量为 M_E,具体为

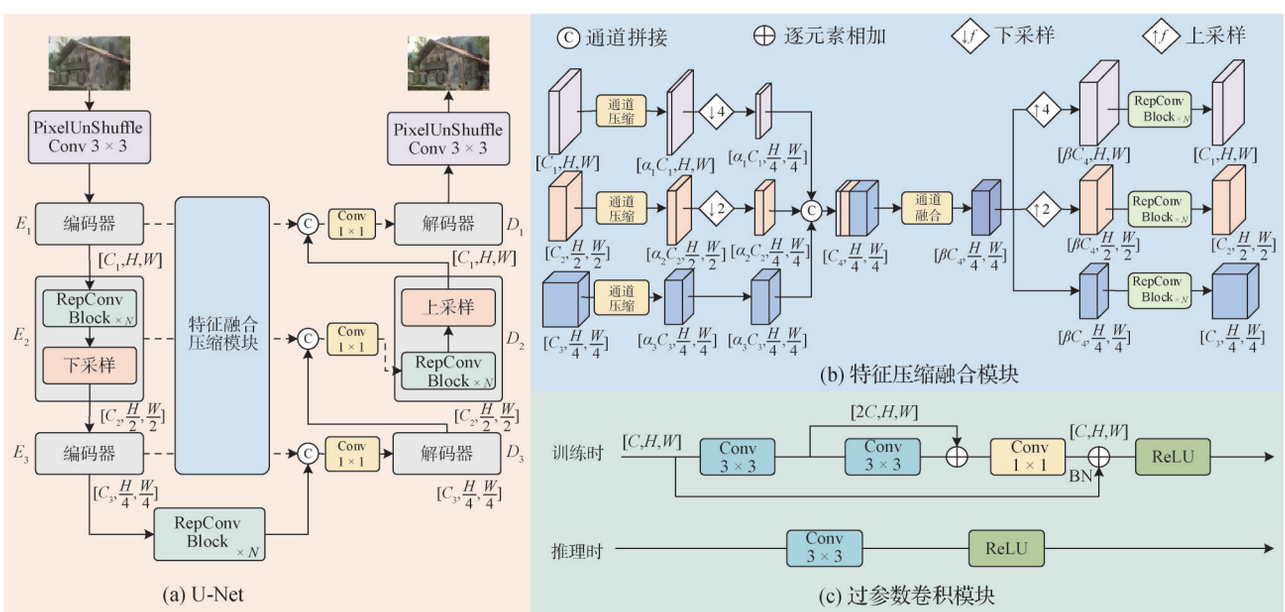

图 3　网络总体架构图

Fig. 3　Overall architecture of the network ((a) U-Net network; (b) feature fusion and compression module; (c) RepConvBlock)

$$\begin{cases} M_{E_1} = H \times W \times C \\ M_{E_2} = \dfrac{H}{2} \times \dfrac{W}{2} \times 2C = \dfrac{M_{E_1}}{2} \\ M_{E_3} = \dfrac{H}{4} \times \dfrac{W}{4} \times 4C = \dfrac{M_{E_1}}{4} \end{cases} \tag{2}$$

式中，M_{E_i}表示在各个编码阶段的内存占用量，H，W，C分别表示对应特征图的高宽和通道数。

在相邻阶段之间，通过下采样操作，将特征图大小逐级减半、通道数加倍。这种方法能够将计算量和参数量减半。类似地，解码过程中，通过对称的上采样操作逐步恢复空间分辨率（特征图尺寸逐级加倍、通道数折半），实现像素级的特征重建。网络通过3层跳跃连接建立跨层级特征融合，将编码器各阶段的特征图分别直连至对应尺度的解码器阶段，具体如图3（a）虚线所示。各个编码阶段结束后，需要将对应特征图暂存于内存中，直到对应的解码阶段才可释放，将由此产生的特征图内存占用量记为M_{sc}，构成模型整体内存消耗的关键组成部分，其计算式为

$$M_{\mathrm{sc}} = \left\{ \sum_{i=1}^{n} M_{E_i} \middle| n = 1, 2, 3 \right\} \tag{3}$$

在编码器阶段，可求得$M_{\mathrm{sc}} = \{ M_{E_1}, \dfrac{3M_{E_1}}{2}, \dfrac{7M_{E_1}}{4} \}$。在$E_3$阶段之后，内存占用达到峰值，通过解码器阶段才能逐级释放，这对片上资源受限的设备提出了挑战。

为了应对上述挑战，本文设计了特征压缩融合块，其结构如图3（b）所示。首先，在编码器阶段，通过逐点卷积对各层级特征进行通道降维，每层设置一个压缩系数$\alpha \in (0, 1]$，逐步压缩通道维度以降低冗余。随后通过双线性插值将多尺度特征统一缩放到编码器末端的最小分辨率，完成空间对齐的特征图沿通道维度进行拼接实现跨层级的特征融合。最后，引入压缩系数$\beta \in (0, 1]$，对拼接特征实施二次压缩，生成更紧凑的特征表示。在相应的解码器阶段，将输入的紧凑特征通过双线性上采样为原始尺度后，采用一组过参数化卷积模块进行特征增强，最终输出融合多尺度上下文信息的增强特征。

通过上述方法，各编码器特征图内存占用为

$$\begin{cases} M_{E_1'} = \alpha_1 \times \dfrac{H}{4} \times \dfrac{W}{4} \times C = \dfrac{\alpha_1 M_{E_1}}{16} \\ M_{E_2'} = \alpha_2 \times \dfrac{H}{4} \times \dfrac{W}{4} \times 2C = \dfrac{\alpha_2 M_{E_1}}{8} \\ M_{E_3'} = \alpha_3 \times \dfrac{H}{4} \times \dfrac{W}{4} \times 4C = \dfrac{\alpha_3 M_{E_1}}{4} \end{cases} \tag{4}$$

式中，$M_{E_i'}$表示在应用特征压缩融合模块后编码器特征图内存占用，α_i表示对应压缩阶段的压缩系数，β表示融合阶段的压缩系数。

应用特征融合压缩模块后，可得到每个阶段的内存占用，具体为

$$M_{\mathrm{sc}}' = \left\{ \frac{\alpha_1 M_{E_1}}{16}, \frac{(\alpha_1 + 2\alpha_2) M_{E_1}}{16}, \frac{(\alpha_1 + 2\alpha_2 + 4\alpha_3) M_{E_1}}{16} \right\}$$

式中，集合中的第i个元素记为M_{sc_i}'，表示在应用特征压缩融合模块后，模型第i个阶段产生的跳跃连接特征图的内存占用。当$i = 3$时整个模型的内存占用达到峰值。

因此，对比每个阶段内存的压缩率为

$$\varphi = \left\{ \frac{\alpha_1}{16}, \frac{\alpha_1 + 2\alpha_2}{24}, \frac{\beta(\alpha_1 + 2\alpha_2 + 4\alpha_3)}{28} \right\}$$

该方法有效降低了U-Net中长跳跃连接导致的内存消耗，显著提升了模型的内存效率。同时，该架构在整合多尺度上下文信息的基础上，实现了与传统跳跃连接结构相当的重建性能，并在内存开销上具备更突出的优势，适用于移动端等资源受限场景。

3 实验结果与分析

3.1 数据集

使用DIV2K数据集构建训练集和验证集。首先，将DIV2K的所有图像转换为YUV格式，作为未压缩原始帧。然后，使用编码器在帧内模式下压缩这些帧，生成低质量重建帧。测试工具使用AOM（alliance for open media）标准工作组代码库中的AVM（AOM video model）稳定版本v3.0.0。

测试集在编码环境配置上与训练集保持完全一致，测试集选自Class A1—A5，共5类标准视频序列，涵盖多种分辨率和内容场景，包括4 K（8个序列）、1 080 P（22个序列）、720 P（8个序列）、360 P（6个序列）和270 P（4个序列）。每个序列选取连续30帧作为测试样本。该设置覆盖从高分辨率到低分辨率、从静态到动态、从简单到复杂的多种图像内容，有效验证了所提方法在不同分辨率和场景下的鲁棒性与泛化能力。

3.2 实验设置

训练时，图像样本被随机裁剪成128 × 128像素固定大小的块。采用Adam优化器对模型进行优

化，总共进行20万次迭代。初始学习率为1×10^{-3}，使用余弦退火策略动态调整学习率减少到1×10^{-4}，批次大小设置为64。

本文实验中的模型在Intel Core i7-8770k CPU @ 3.70 GHz、32 GB内存和NVIDIA GeForce 2080 Ti GPU的硬件环境下进行训练。为评估移动端推理性能，模型测试基于高通骁龙8 Gen1移动平台，采用SNPE（snapdragon neural processing engine）推理框架提供的后量化方法，将浮点模型量化为8比特全量化模型，确保边缘计算环境下的部署可行性。

在编码性能方面，本文采用BD-Rate作为主要评价指标，基于峰值信噪比（peak signal to noise ration，PSNR）与码率曲线计算平均码率差异，用于评估不同滤波方案在保证图像质量前提下的压缩效率变化，数值越小表示在相同图像质量下所需码率越低，编码性能越优。参考基准为AVM（v3.0.0）编码器默认配置下的标准滤波器组合（DBF + CDEF + CCSO + LR）。

在运行效率方面，本文引入了理论计算复杂度（FLOPs）、峰值内存占用、推理时间，以及计算密度（即每秒实际执行的浮点运算次数）。其中，计算密度反映了模型的硬件执行效率，是衡量模型硬件适配性的关键参考指标。

3.3 实验结果

3.3.1 编码性能增益对比分析

为评估各模型在不同编码场景下的压缩性能，本文分别在帧内预测模式和帧间预测模式下，对比分析了多个代表性模型的BD-Rate指标，测试序列涵盖从270 P至4 K的多种分辨率，确保了评估的全面性与泛化性，如表1和表2所示。

在帧内预测模式下，本文方法在大多分辨率上均实现了优于多数对比方法的BD-Rate表现，整体平均为−5.09%。其中，在ClassA1和ClassA3上分别取得−5.52%和−6.75%的压缩增益，显著优于VRCNN（−1.07%、−4.16%）、MobileNetV2（−2.24%、−3.40%）以及RepSR（−1.96%、−5.66%）。相比于现有DeepSN-Net（平均−5.23%）和LCMMCNN（平均−3.81%），所提方法在整体上达到相当甚至略优的性能，且表现更为稳定，接近当前主流网络NAFNet（−5.91%）。

在帧间预测模式下，各模型在BD-Rate上的差异略有收敛，但本文方法依然保持领先，整体平均为−4.32%，优于LCMMCNN（−3.69%）、MobileNetV2（−2.58%）与VRCNN（−2.33%），在ClassA1和ClassA3等典型序列上仍展现出明显优势。DeepSN-Net在此设置下的平均BD-Rate为−4.35%，略高于本文方法。

图4展现了各模型在帧内编码QP210时的可视化滤波结果。

综上，无论在帧内还是帧间预测模式下，本文提出的方法均表现出良好的BD-Rate增益，特别是在中高分辨率段及对比时展现出明显优势，验证了所提结构在实用场景下的有效性与推广潜力。

3.3.2 运行效率对比实验

在运行资源效率对比中，统一采用ClassA2序列进行测试，以保证评估过程中分辨率的一致性。其

表1 各模型在帧内预测模式下的编码性能

Table 1 Intra prediction mode coding performance of each model

/%

方法	ClassA1（4 K）	ClassA2（1 080 P）	ClassA3（720 P）	ClassA4（360 P）	ClassA5（270 P）	平均
RepSR(Wang等，2022b)	−1.96	−3.26	−5.66	−4.26	−3.91	−3.81
LCMMCNN(Wang等，2022a)	−2.56	−3.84	−3.40	−4.82	−4.46	−3.81
VRCNN（Dai等，2016）	−1.07	−2.04	−4.16	**−7.40**	−2.73	−3.48
NAFNet（Chen等，2022）	−4.84	**−5.39**	**−8.08**	−5.90	**−5.33**	**−5.91**
MobileNetV2（Sandler等，2018）	−2.24	−3.61	−3.40	−2.17	−1.88	−2.66
DeepSN-Net（Deng等，2025）	−5.37	−5.01	−7.23	−4.51	−4.01	−5.23
本文	**−5.52**	−4.53	−6.75	−4.56	−4.08	−5.09

注：加粗字体表示各列最优结果。

表2　各模型在帧间预测模式下的编码性能

Table 2　Inter prediction mode coding performance of each model

/%

方法	ClassA1（4 K）	ClassA2（1 080 P）	ClassA3（720 P）	ClassA4（360 P）	ClassA5（270 P）	平均
RepSR(Wang等,2022b)	−3.25	−3.06	−4.71	−2.85	−2.63	−3.30
LCMMCNN(Wang等,2022a)	−3.68	−3.39	−5.22	−3.23	−2.94	−3.69
VRCNN(Dai等,2016)	−2.47	−2.23	−3.29	−1.82	−1.84	−2.33
NAFNet(Chen等,2022)	−3.95	**−5.09**	−3.88	**−4.66**	**−4.91**	**−4.50**
MobileNetV2(Sandler等,2018)	−2.17	−2.56	−2.83	−3.01	−2.35	−2.58
DeepSN-Net(Deng等,2025)	−4.15	−3.61	**−6.23**	−4.41	−3.35	−4.35
本文	**−4.61**	−3.53	−5.77	−4.06	−3.65	−4.32

注：加粗字体表示各列最优结果。

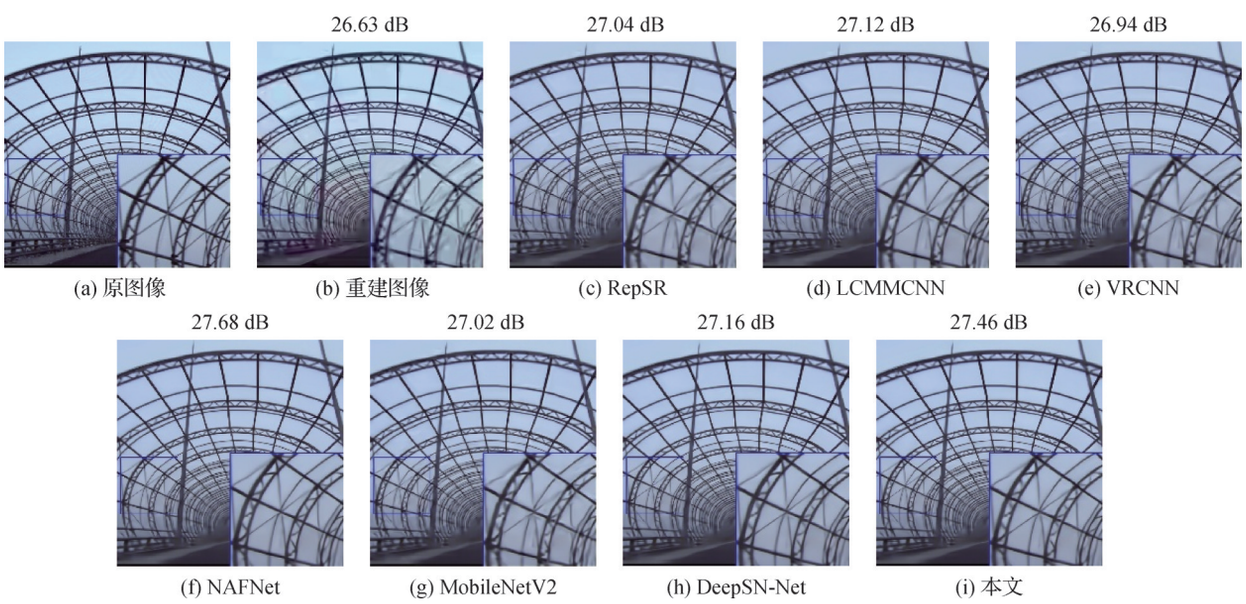

图4　各模型帧内编码QP210时的滤波结果

Fig. 4　Filter results of each model for intra code at QP210 ((a) original image; (b) reconstruction; (c) RepSR; (d) LCMMCNN; (e) VRCNN; (f) NAFNet; (g) MobileNetV2; (h) DeepSN-Net; (i) ours)

分辨率与移动端视频处理场景高度匹配，适合作为模型运行效率的标准测试基准。其中，LCMMCNN基于残差块堆叠构建；VRCNN采用浅层双分支宽通道架构；RepSR采用重参数化的卷积模型，上述模型均在原分辨率图像上进行处理。DeepSN-Net在输入阶段对图像进行4倍下采样，显著降低了FLOPs。NAFNet（U-Net结构）和MobileNetV2采用倒置残差结构并深度优化FLOPs，其中MobileNetV2同样保持原分辨率处理。

从表3可见，在峰值内存上，LCMMCNN、VRCNN和NAFNet等以残差结构为代表的多分支模型相比于单分支模型，表现出更高的内存消耗（准则3）。另外，由于NAFNet和MobileNet采用了倒置残差结构，进一步加剧了内存占用的增长（准则6）。

图5展示了各个模型在推理过程中每层的实时内存分配。其中，内存曲线的尖峰表示对应残差张量相加操作时的过程。相比之下，DeepSN-Net和本文提出的模型展现出更平缓的内存分配特性。尽管各模型的FLOPs存在细微差异，计算密度作为衡量实际计算效率的重要指标，能够更直观地反映模型的硬件执行性能。本文提出的模型在对比模型中具有最高的计算密度，展现出卓越的计算效率。

表 3　与当前滤波模型计算复杂度、内存占用对比

Table 3　Complexity and memory comparison with existing filtering models

模型	计算量/MFLOPs	峰值内存/MB	推理时间/ms	计算密度/MFLOPS
RepSR(Wang等,2022b)	231.70	192.53	16.22	14 285.08
LCMMCNN(Wang等,2022a)	211.40	128.59	66.49	3 179.48
VRCNN(Dai等,2016)	226.06	255.45	24.18	9 349.21
NAFNet(Chen等,2022)	236.35	489.21	453.22	521.49
MobileNetV2(Sandler等,2018)	222.68	381.94	72.66	3 064.74
DeepSN-Net(Deng等,2025)	230.13	264.72	87.12	2 641.53
本文	239.15	**54.96**	**12.98**	**18 424.37**

注:加粗字体表示各列最优结果。

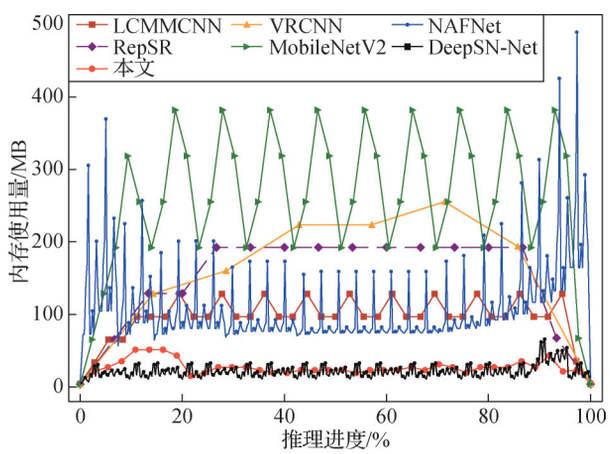

图 5　模型运行过程中的内存占用对比图

Fig. 5　Memory usage comparison during model running

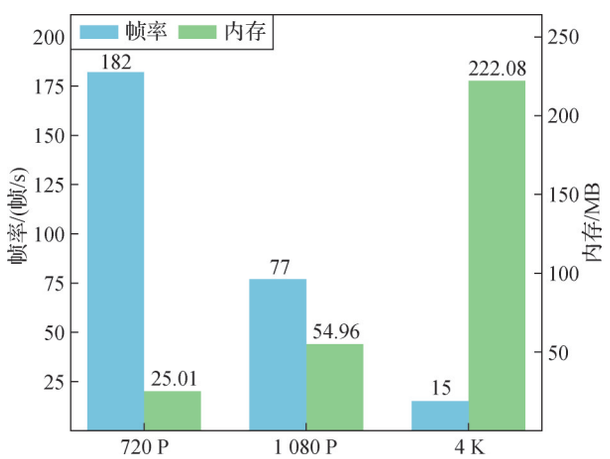

图 6　不同分辨率下的帧率与内存占用对比

Fig. 6　Speed and peak memory by resolution

本文测试了除 1 080 P 外主流的 720 P 和 4 K 分辨率下模型的推理帧率以及峰值内存,如图 6 所示。结果表明,在 720 P 下模型可实现 182 帧/s,峰值内存为 25. 01 MB;而在 4 K 下帧率仅 15 帧/s,峰值内存升至 222. 08 MB,难以满足实时处理需求,表明针对高分辨率场景仍需进一步优化模型。

综合来看,虽然本文模型在重建质量上较最优模型 NAFNet 分别在帧内预测和帧间预测模式下略低 0. 82% 和 0. 18%,但在 1 080 P 分辨率下,其峰值内存仅为 NAFNet 的 11. 23%,推理速度提升近 34 倍,显著提升了在移动和边缘设备上的实时推理可行性,提供了更优的工程化解决方案。

3. 3. 3　消融实验

为了统一测试条件并评估模块在高清移动端场景下的内存占用与推理延迟,本节消融实验统一基于 ClassA2(1 080 P)数据集,在帧内预测模式下进行,以验证本文所提模块的有效性。

表 4 展示了不同跳跃连接方式与压缩策略下的性能对比。

从表 4 可以看出,传统的 AddSkip 与 ConcatSkip 均未采用压缩策略,BD-Rate 分别为 -4. 26% 和 -4. 35%,但特征图峰值内存(M_{sc3})高达 27. 69 MB。相比之下,本文提出的压缩融合结构在引入通道压缩和空间对齐的基础上,显著降低了内存开销。例如在压缩系数 $\alpha_1 = \alpha_2 = \alpha_3 = \beta = 0.5$ 时,峰值内存 M_{sc3} 降至 1. 73 MB,节省约 93. 75%,同时 BD-Rate 提升至 -4. 53%。进一步增加压缩系数至 0. 75 可带来轻微性能提升(-4. 57%),但相应内存和时间开销增加,性价比不高。而在极端压缩(压缩系数均为 0. 25)下,虽然 M_{sc3} 可压缩至 0. 43 MB,但 BD-Rate 明显下降至 -3. 31%。实验表明,压缩系数的减小可显著降低特征图峰值内存,但过度压缩可能带来信息损失,从而影响编码性能,验证了该模块在资源受限场景下的可调性。鉴于当前尚缺乏可统一度量编码

表4　不同跳跃连接与压缩策略下的性能对比

Table 4　Performance comparison under different skip connection and compression strategies

方法	特征压缩	BD-Rate/%	M_{sc1}/MB	M_{sc2}/MB	M_{sc3}/MB	时间/ms
w/o Skip	×	−0.81	–	–	–	**9.74**
AddSkip	×	−4.26	15.82	23.73	27.69	10.14
ConcatSkip	×	−4.35	15.82	23.73	27.69	10.27
本文($\alpha_1=\alpha_2=\alpha_3=\beta=0.75$)	√	**−4.57**	0.74	2.22	3.89	14.63
本文($\alpha_1=\alpha_2=\alpha_3=\beta=0.5$)	√	−4.53	0.50	1.48	1.73	12.98
本文($\alpha_1=\alpha_2=\alpha_3=\beta=0.25$)	√	−3.31	**0.25**	**0.74**	**0.43**	11.88

注：加粗字体表示各列最优结果，“–”表示无对应数据，“√”表示该方法包括对应的策略，“×”表示不包括该策略。

性能与资源消耗之间平衡的客观指标，为保证实验对比的公平性与部署策略的一致性，本文所有实验统一采用压缩系数 $\alpha_1=\alpha_2=\alpha_3=\beta=0.5$。

4　结　论

本文提出了一种基于结构重参数化技术和U-Net架构的高效环路滤波方法。该方法通过纯卷积模块和特征压缩融合模块，在显著提升编码性能的同时，有效降低了NPU部署复杂度和内存占用。相较于AV1环路滤波基准，模型在帧内和帧间编码模式下分别实现了−5.09%和−4.32%的BD-Rate增益，同时在骁龙8 Gen1处理器上，推理速度达77帧/s，峰值内存仅为55 MB。

然而，模型存在以下局限性：首先，训练仅针对YUV420格式，缺乏对HDR、10-bit视频以及YUV422、YUV444等其他色彩格式的支持，导致泛化能力有限；其次，本文方法主要优化了1 080 P分辨率，对于4 K等高分辨率场景仍无法实现实时推理；最后，模型中的压缩系数需要人工设定，缺乏针对不同场景的灵活调整机制。因此，未来的工作将着重于高分辨率视频的支持、增强模型的适应性以及探索自适应压缩系数的机制，以提高模型在不同应用场景中的灵活性和表现。

参考文献(References)

Chen L Y, Chu X J, Zhang X Y and Sun J. 2022. Simple baselines for image restoration//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 17-33 [DOI: 10.1007/978-3-031-20071-7_2]

Chollet F. 2017. Xception: deep learning with depthwise separable convolutions//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 1800-1807 [DOI: 10.1109/CVPR.2017.195]

Dai Y Y, Liu D and Wu F. 2016. A convolutional neural network approach for post-processing in HEVC intra coding//Proceedings of the 23rd International Conference on Multimedia Modeling. Reykjavik, Iceland: Springer: 28-39 [DOI: 10.1007/978-3-319-51811-4_3]

Dehghani M, Tay Y, Arnab A, Beyer L and Vaswani A. 2021. The efficiency misnomer//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: ICLR: #12894

Deng X, Zhang C X, Jiang L, Xia J Y and Xu M. 2025. DeepSN-Net: deep semi-smooth Newton driven network for blind image restoration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(4): 2632-2646 [DOI: 10.1109/TPAMI.2024.3525089]

Ding X H, Zhang X Y, Ma N N, Han J G, Ding G G and Sun J. 2021. RepVGG: making VGG-style ConvNets great again//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 13728-13737 [DOI: 10.1109/CVPR46437.2021.01352]

Ding Y H, Wu H, Kong F L, Xu D and Yuan G W. 2023. A dual of real image noise-related blind denoising technique. Journal of Image and Graphics, 28(7): 2026-2036 (丁岳皓, 吴昊, 孔凤玲, 徐丹, 袁国武. 2023. 面向真实图像噪声的两阶段盲去噪. 中国图象图形学报, 28(7): 2026-2036) [DOI: 10.11834/jig.211020]

Du Y X, Zhao X and Liu S. 2021. Cross-component sample offset for image and video coding//Proceedings of 2021 International Conference on Visual Communications and Image Processing (VCIP). Munich, Germany: IEEE: 1-5 [DOI: 10.1109/VCIP53242.2021.9675355]

Fu C M, Alshina E, Alshin A, Huang Y W, Chen C Y, Tsai C Y, et al. 2012. Sample adaptive offset in the HEVC standard. IEEE Transactions on Circuits and Systems for Video Technology, 22(12): 1755-1764 [DOI: 10.1109/TCSVT.2012.2221529]

Han S, Mao H Z and Dally W J. 2015. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding [EB/OL]. [2016-02-15].

https://arxiv.org/pdf/1510.00149.pdf

Huang Z J, Li Y C and Sun J. 2020. Multi-gradient convolutional neural network based in-loop filter for Vvc//Proceedings of 2020 IEEE International Conference on Multimedia and Expo (ICME). London, United Kingdom: IEEE: 1-6 [DOI: 10.1109/ICME46284.2020.9102826]

Howard A G, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. 2017. MobileNets: efficient convolutional neural networks for mobile vision applications [EB/OL]. [2026-01-11]. https://arxiv.org/pdf/1704.04861.pdf

Jaderberg M, Vedaldi A and Zisserman A. 2014. Speeding up convolutional neural networks with low rank expansions//Proceedings of 2014 British Machine Vision Conference. Nottingham, UK: BMVA Press

Jia C M, Ma H C, Yang W H, Ren W Q, Pan J S, Liu D, et al. 2021. Video processing and compression technologies. Journal of Image and Graphics, 26(6): 1179-1200 (贾川民, 马海川, 杨文瀚, 任文琦, 潘金山, 刘东, 等. 2021. 视频处理与压缩技术. 中国图象图形学报, 26(6): 1179-1200) [DOI: 10.11834/jig.200861]

Lavin A and Gray S. 2016. Fast algorithms for convolutional neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 4013-4021 [DOI: 10.1109/CVPR.2016.435]

Li M K, Liu W and Chen W D. 2025. Image denoising via Swin Transformer V2 and feature fusion U-Net. Journal of Image and Graphics, 30(11): 3524-3534 (利铭康, 柳薇, 陈卫东. 2025. Swin Transformer V2和特征融合的U-Net图像去噪方法. 中国图象图形学报, 30(11): 3524-3534) [DOI: 10.11834/jig.240659]

Liu Z, Li J G, Shen Z Q, Huang G, Yan S M and Zhang C S. 2017. Learning efficient convolutional networks through network slimming//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 2755-2763 [DOI: 10.1109/ICCV.2017.298]

Ma D, Zhang F and Bull D R. 2021. MFRNet: a new CNN architecture for post-processing and in-loop filtering. IEEE Journal of Selected Topics in Signal Processing, 15(2): 378-387 [DOI: 10.1109/JSTSP.2020.3043064]

Midtskogen S and Valin J M. 2018. The AV1 constrained directional enhancement filter (CDEF)//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE: 1193-1197 [DOI: 10.1109/ICASSP.2018.8462021]

Mukherjee D, Li S Y, Chen Y, Anis A, Parker S and Bankoski J. 2017. A switchable loop-restoration with side-information framework for the emerging AV1 video codec//Proceedings of 2017 IEEE International Conference on Image Processing (ICIP). Beijing, China: IEEE: 265-269 [DOI: 10.1109/ICIP.2017.8296284]

Norkin A, Bjontegaard G, Fuldseth A, Narroschke M, Ikeda M and Andersson K. 2012. HEVC deblocking filter. IEEE Transactions on Circuits and Systems for Video Technology 22(12): 1746-1754 [DOI: 10.1109/TCSVT.2012.2223053]

Pan Z Q, Yi X K, Zhang Y, Jeon B and Kwong S. 2020. Efficient in-loop filtering based on enhanced deep convolutional neural networks for HEVC. IEEE Transactions on Image Processing 29: 5352-5366 [DOI: 10.1109/TIP.2020.2982534]

Sandler M, Howard A, Zhu M L, Zhmoginov A and Chen L C. 2018. MobileNetV2: inverted residuals and linear bottlenecks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA: IEEE: 4510-4520 [DOI: 10.1109/CVPR.2018.00474]

Tsai C Y, Chen C Y, Yamakage T, Chong I S, Huang Y W, Fu C M, et al. 2013. Adaptive loop filtering for video coding. IEEE Journal of Selected Topics in Signal Processing, 7(6): 934-945 [DOI: 10.1109/JSTSP.2013.2271974]

Wang S, Fu Y B, Zhu C, Song L and Zhang W J. 2022a. Low-complexity multi-model CNN in-loop filter for AVS3//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, Singapore: IEEE: 1630-1634 [DOI: 10.1109/ICASSP43922.2022.9746146]

Wang X T, Dong C and Shan Y. 2022b. RepSR: training efficient VGG-style super-resolution networks with structural re-parameterization and batch normalization//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: Association for Computing Machinery: 2556-2564 [DOI: 10.1145/3503161.3547915]

Wang Z, Ma C Y, Liao R L and Ye Y. 2021. Multi-density convolutional neural network for in-loop filter in video coding//Proceedings of 2021 Data Compression Conference (DCC). Snowbird, USA: IEEE: 23-32 [DOI: 10.1109/DCC50243.2021.00010]

Zhang X D, Zeng H and Zhang L. 2021. Edge-oriented convolution block for real-time super resolution on mobile devices//Proceedings of the 29th ACM International Conference on Multimedia. [s.l.]: Association for Computing Machinery: 4034-4043 [DOI: 10.1145/3474085.3475291]

Zoph B and Le Q V. 2017. Neural architecture search with reinforcement learning//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: ICLR: #1578

作者简介

毛可权，男，硕士研究生，主要研究方向为人工智能和视频图像编码。E-mail：kqmao@stu.hznu.edu.cn

丁丹丹，通信作者，女，教授，主要研究方向为多媒体通信、智能视频编码、三维点云编码与重建。

E-mail：dandanding@hznu.edu.cn

姚杰，男，硕士研究生，主要研究方向为人工智能和视频图像编码。E-mail：yaoge.yj@alibaba-inc.com

余国生，男，硕士研究生，主要研究方向为人工智能和视频图像编码。E-mail：guosheng.ygs@alibaba-inc.com