# 基于双分支注意力网络的立体视频压缩

唐述，赵瑜，杨书丽，谢显中
（重庆邮电大学计算机科学与技术学院（示范性软件学院），重庆 400065）

**摘　要：**针对现有基于深度学习的立体视频压缩网络几乎只采用卷积操作来提取和融合特征，导致无法有效捕捉局部范围内的非重复纹理细节和忽略了全局特征等缺陷，严重影响了解码过程中图像重建质量的问题，提出了一种双分支注意力网络（DAN），通过开发和融合区域范围内的逐像素相似性和整幅图像的全局结构特征，实现更高质量的立体视频压缩编码。首先，提出了一种基于 Transformer 和通道注意力的局部和全局双分支编解码块（LGEDB），通过融合区域范围内每个像素点的自注意力和每个通道的全局注意力，实现对局部非重复纹理细节和全局结构信息的准确捕捉。其次，提出了一种基于可逆神经网络（INN）和门控机制的双分支高频信息融合模块（DHFFM），通过对运动补偿特征和视差特征中高频信息的准确提取以及逐像素点特征的筛选，实现对运动补偿特征和视差特征的高效融合。实验表明，DAN 在相同或更低比特率下能够实现更高质量重建，且模型参数量更少。

**关键词：**深度学习；立体视频压缩编码；双分支注意力；可逆神经网络；门控机制
**中图分类号：**TN919.8
**文献标志码：**A
**DOI：**10.11959/j.issn.1000–436x.2025139

# Dual-branch attention network-based stereoscopic video compression

TANG Shu, ZHAO Yu, YANG Shuli, XIE Xianzhong
School of Computer Science and Technology (National Exemplary Software School), Chongqing University of Posts and Telecommunications, Chongqing 400065, China

**Abstract:** Aiming at the problem that the existing deep learning-based stereoscopic video compression networks only used convolutional operations to extract and fuse features, which limited their ability to effectively capture non-repetitive texture details within local areas and cannot capture global features, thus affecting the quality of image reconstruction during decoding seriously, a dual-branch attention network (DAN) that achieved higher quality stereoscopic video compression coding by exploiting and fusing pixel-wise similarity within local regions and the global structural features of the entire image was proposed. Firstly, a local and global encoder-decoder block (LGEDB) based on Transformer and channel attention was proposed, which accurately captured non-repetitive texture details in local regions and global structural information by integrating pixel-level self-attention within each local area and global attention across channels. Secondly, a dual-branch high frequency fusion module (DHFFM) based on invertible neural network (INN) and gating mechanisms was proposed, which achieved efficient fusion of motion-compensated features and disparity features by accurate high-frequency information extraction and pixel-wise features filtering. Experimental results demonstrate that DAN achieves higher-quality reconstruction at the same or lower bitrate while maintaining fewer model parameters.

**Keywords:** deep learning, stereoscopic video compression coding, dual-branch attention, INN, gate mechanism

---

**收稿日期：**2025–04–25；**修回日期：**2025–08–06
**通信作者：**唐述，tangshu@cqupt.edu.cn
**基金项目：**国家自然科学基金资助项目（No.61601070）；重庆市自然科学基金资助项目（No.CSTB2023NSCQ-MSX0680）；重庆市教育委员会科学技术研究重大基金资助项目（No.KJZD-M202300101）；重庆邮电大学博士研究生创新人才基金资助项目（No.BYJS202217）
**Foundation Items:** The National Natural Science Foundation of China (No.61601070), The Chongqing Natural Science Foundation (No.CSTB2023NSCQ-MSX0680), The Major Project of Scientific and Technology Research of Chongqing Education Commission (No.KJZD-M202300101), The Innovation Talent Project for Doctoral Students of Chongqing University of Posts and Telecommunications (No.BYJS202217)

## 0 引言

立体双视图视频是一种通过提供2个不同视角的视频（左视角和右视角）来模拟人眼的视差，从而在观看者的感知中产生三维立体效果的视频技术。随着配备立体摄像头的自动驾驶汽车（AV, autonomous vehicle）的普及以及虚拟现实（VR, virtual reality）设备的发展和应用，立体视频数据（双视图）的重要性显著提升。在AV和VR设备的应用过程中，编解码器不仅需要保持低时延，而且更需要关注解码后重建图像的质量。因此，如何高效压缩立体视频数据成为人们研究的热点。

传统的立体视频压缩方法是将每个视图单独放入单视图编解码器（如H.264/AVC[1]或HEVC[2]）中，但是忽略了压缩后2个视图之间的相似性造成的视差冗余。为了解决这一缺陷，人们便在单视图编解码器上进行扩展，如H.264/AVC[3]的扩展H.264/MVC采用视差补偿的方式来消除压缩后的视差冗余，或者HEVC的扩展MV-HEVC[4]是采用编码树单元（CTU, coding tree unit）[5]等技术来消除视差冗余，它们都是通过组合帧间和视图间的预测方式来提高压缩性能。

近年来，因为深度神经网络（DNN, deep neural network）强大的学习能力，基于DNN的立体视频压缩编码方法得到了极大的发展，成为人们研究的热点。Chen等[6]首次提出了一种基于DNN的立体视频压缩（LSVC, learning-based stereo video compression）框架，通过在特征空间中使用运动压缩和视差压缩等方式来减少时间和双目的冗余，从而实现立体视频的高效压缩编码。Hou等[7]提出了一种基于DNN的低时延立体视频压缩（LLSS, low-latency neural stereo streaming）方法，通过引入双向特征位移模块以及运动信息和视差信息的共享，实现了左、右视图的并行处理，显著降低了时延并提升了率失真（RD, rate distortion）性能。虽然与传统的立体视频（双视图）压缩编码方法相比，现有基于DNN的立体视频压缩方法能够获得更优的率失真性能，但是他们几乎只采用了卷积操作来进行特征的提取和融合。虽然卷积操作能够有效提取局部信息，但是对于非重复且复杂的纹理信息的捕捉能力较差，并且卷积操作无法捕捉长距离的全局信息。因此，现有基于DNN的立体视频压缩方法的编码图像质量仍有较大提升空间。

根据上述分析，针对现有方法的缺陷，本文提出了一种双分支注意力网络（DAN, dual-branch attention network）来实现更高质量的立体视频压缩，DAN能够用相同或者更低的每像素点比特（BPP, bit per pixel）实现更高质量的重建图像，并且模型的参数量更少。本文在Cityspace[8]、KITTI 2012[9]和KITTI 2015[10]这3个立体数据集上分别进行了测试，DAN的性能明显优于近年来极具代表性的传统多视图视频压缩方法和基于DNN的立体视频压缩方法。本文的贡献主要体现在以下3个方面。

1) 提出了一种局部和全局双分支编解码块（LGEDB, local and global encoder-decoder block），LGEDB由窗口自注意力机制和通道注意力机制组成，通过融合区域范围内每个像素点的自注意力和每个通道的全局注意力，实现对局部非重复纹理细节和全局结构信息的准确捕捉。

2) 提出了一种新颖的双分支高频信息融合模块（DHFFM, dual-branch high frequency fusion module），DHFFM是由可逆神经网络（INN, invertible neural network）和门控机制组成。INN能够有效提取出运动补偿特征和视差补偿特征中的高频信息，门控机制可以有效进行逐像素点的特征筛选，从而实现运动补偿特征和视差补偿特征的高效融合。

3) 大量的实验结果表明，与近年来极具代表性的方法相比，本文方法能够用更低的BPP实现更高质量的图像重建，并且模型的参数量更少。

## 1 相关工作

### 1.1 单视图图像和单视图视频压缩方法

传统的单视图图像和单视图视频压缩方法大多采用人为设计的规则和先验知识来减少空间和时空冗余。在图像压缩方面，典型代表如JPEG[11]使用分块离散余弦变换和霍夫曼编码来消除空间冗余；JPEG2000[12]则采用全局小波变换获得更高质量的压缩效果。在视频压缩方面，H.264/AVC[1]通过帧间预测消除时间冗余、帧内预测消除空间冗余，而HEVC[2]通过进一步引入编码树单元，实现了更加灵活的块划分，并增加了更多的预测模式，从而提升了压缩质量，更加高效地消除了时空冗余。虽然这些传统方法在数据压缩效率上取得了一定的成

就，但它们在应对复杂非线性特征和高度动态场景时却受到了极大的限制，缺乏灵活性和自适应能力。

近年来，因为DNN强大的特征提取和建模能力，基于DNN的图像和视频压缩编码方法已成为人们研究的热点[13-32]。在基于DNN的图片研究初期，Toderici等[13]提出了基于循环神经网络（RNN, recurrent neural network）的图像压缩方法，通过结合RNN编码器与解码器、二值化器以及神经网络熵编码器，实现了可变压缩率的图片压缩。Muckley等[16]通过引入非二值判别器来增强图像的保真性和感知质量，在保持失真指标（如峰值信噪比（PSNR, peak signal-to-noise rate）和多尺度结构相似性（MS-SSIM, multi scale structural similarity index measure）[17]）不变的基础上提升了主观质量。由于Transformer在计算机视觉领域中的优异表现，能有效地捕捉全局信息，越来越多研究尝试将其引入图像压缩任务中[18-20]。例如，Nan等[18]结合Transformer与大核卷积，提出了一种双通道注意力机制，能够很好地处理长距离特征依赖和信息丢失问题。同时，生成对抗网络（GAN, generative adversarial network）也被广泛运用于图像压缩任务中[21-23]，以提高感知质量。Agustsson等[22]进一步提出了一种结合多尺度判别器的GAN压缩模型，在低比特率的情况下仍能恢复细节丰富的图像，有效提升了主观视觉效果。

除了基于DNN的图像压缩方法之外，基于DNN的视频压缩方法也获得了较大成功。DVC[24]是首个端到端DNN视频压缩框架，其使用光流估计完成运动补偿，并通过自编码网络对运动和残差信息进行压缩，显著减少了时空冗余，其性能优于传统的HEVC。DVC为基于DNN的视频压缩方法奠定了基础，后续的框架都遵循DVC混合编码框架，根据运动估计的运动信息产生补偿帧，然后通过残差编解码器得到重构帧。Habibian等[25]在该混合编码框架上，结合3D自编码器和自回归模型来捕获时间和空间的相关性，从而提高率失真性能。Agustsson等[26]提出了一种基于尺度空间光流（SSF, scale-space flow）的视频压缩方法，该方法引入了尺度空间变换，可以在运动预测不准确的区域自适应地模糊源内容。

在编码策略优化方面，Li等[27]通过引入潜在先验和双空间先验，在单一模型中实现了自适应量化和动态比特分配。为进一步挖掘上下文信息，Li等[28]提出了一种多样化上下文的视频压缩方法，通过在时间和空间维度上增加上下文信息，从而增强了表达能力，提升了压缩性能。为了提高运动信息的准确性，Rippel等[29]使用环内流预测模块，将之前传输的信息作为参考，从而生成更准确的运动估计信息。

为了满足实际部署要求，近年来也出现了许多更高效快速的视频压缩方法。Hu等[30]提出了一种双阶段的视频压缩方法，首先利用低分辨率特征快速完成初步的运动补偿，再进一步精细化运动信息。Le等[31]提出了一种针对移动设备的视频压缩方法，通过并行熵编码算法，实现了在移动设备上实时解码高清视频。周立新[32]通过编码器将高维视频数据压缩为低维潜在变量，再通过解码器实现高质量的视频重建。

### 1.2　立体双视图图像和立体双视图视频压缩方法

与传统的单视图图像和视频压缩方法相似，传统的立体双视图图像和视频压缩方法同样采用人为设计的规则和先验知识，并结合视差补偿的方式来实现立体图像和视频的压缩编码。例如，H.264/MVC[3]通过引入视差预测策略，消除左、右视图之间的视差冗余，实现立体图像压缩。在最新的标准中，MV-HEVC[4]通过采用CTU进行更加灵活的块划分和更多的预测模式，实现更加精准的视差预测，从而优化立体图像压缩。然而，传统方法的BPP较高，且编码后的图像具有较为明显的瑕疵，难以满足现代多媒体应用对高质量和高压缩率的需求。

近年来，DNN同样被快速发展到立体图像和视频的压缩领域。Liu等[33]提出了一种深度立体双视图图像压缩方法，利用立体图像之间的重叠来减少压缩码率。该方法使用参数化跳跃函数将视差特征进行变换并集成到第二张图像的编码和解码流程中。Deng等[34]进一步提出了一种基于单应性变换的DNN方法，用于立体双视图图像压缩，该方法采用回归模型来估计单应性矩阵，并将左图像进行空间变换，从而减少左、右图像之间的残差信息。Lei等[35]则设计了一种基于双向编码的深度双视图图像压缩方法，该方法通过双向上下文变换模块和双向条件熵模型，利用视图间上下文实现非线性变

换，并通过视图间对应关系改进熵编码的效率。此外，Wödlinger等[36]在潜在空间中使用全局平移和减法，仅对右图像的潜在表示残差进行编码，从而提高立体图像的压缩性能。

对于立体视频的压缩可以通过将立体视频拆分为2个单独的视频帧序列，再分别使用视频压缩技术进行压缩。Hu等[37]通过在特征空间中完成运动估计和补偿等操作，取得了比传统像素空间方法更好的性能。具体来说，FVC[37]采用可变形卷积进行运动补偿，并使用多帧特征融合模块来捕捉时间上下文信息。Li等[38]提出了一种基于条件编码的视频压缩方法，利用上下文特征作为条件，在特征域中执行运动估计和补偿，通过丰富的上下文信息提升了编解码效率。杜秀丽等[39]提出了一种结合3D可变形卷积与Transformer的视频压缩感知方法，通过3D可变形卷积捕获视频时空特征，同时利用Transformer建立关键帧的长距离依赖关系，从而提升压缩性能。同时，Yu等[40]提出的通道自回归熵模型（CWAEM, channel-wise autoregressive entropy model）能够为立体视频压缩方法提升性能。Chen等[41]提出了基于3D Transformer的时空交叉协方差注意力机制，将时空特征融合为联合表示并采用跨通道协方差注意力，从而降低计算开销。为了更好地平衡率失真，Mentzer等[42]提出了一种基于Transformer的视频压缩（VCT, video compression Transformer）方法，通过引入上下文长度为2的时序建模及Transformer模块，取得了更好的率失真性能。Ameur等[43]则另辟蹊径，利用INN能够有效恢复原始高分辨率图像细节能力，先将图像进行分辨率缩放后，再进行压缩，接着通过INN进行恢复，在保证重建图像质量的同时有效减少了存储空间占用。然而，这些单视图立体视频压缩方法不能有效去除视图间的冗余。

为克服上述问题，近年来研究者开始探索面向立体双视图的视频压缩框架。LSVC[6]是首个基于DNN的立体双视图视频压缩框架，其通过运动估计、视差估计和补偿机制，配合融合模块来进行立体视频压缩。同时，LSVC[6]还采用了参考缓冲器来缓存最后编码的帧间特征和相邻视图特征，并利用这些缓存信息进一步消除时间和视图冗余。由于LSVC编码耗时较长，难以满足实际需要，因此，Hou等[7]提出了一种低时延立体视频压缩方法，其引入了双向特征位移模块，在网络的编码器和解码器之间捕获和共享左右视图的互信息。LLSS[7]采用左右视图并行编码方式替代左右视图轮流编码，从而实现了低时延高效率的立体双视图视频压缩。由以上分析可知，虽然现有的基于DNN的立体视频压缩方法能够获得较好的率失真性能，但是他们几乎只采用了卷积操作来进行特征的提取和融合。虽然卷积操作能够有效提取局部信息，但是对于非重复且复杂的纹理细节信息的捕捉能力较差，并且卷积操作无法捕捉长距离的全局信息。因此，现有的基于DNN的立体视频压缩方法的编码图像的质量仍有较大提升空间。

## 2　双分支注意力网络

在本节中，将对本文提出的DAN进行详细的论述：首先介绍DAN的总体框架，然后详细介绍本文提出的局部和全局双分支编解码块（LGEDB）和双分支高频信息融合模块（DHFFM）。

### 2.1　总体框架

立体双视图视频由左（L）和右（R）相机同时拍摄$T$帧，即$\{X_t^{\mathrm{L}},X_t^{\mathrm{R}}\}_{t\in\{1,\cdots,T\}}$，其中，上标代表视图，下标$t$代表时间步长。本文采用与LSVC[6]相同的流程来压缩每一帧，如图1所示，按照先压缩右视图再压缩左视图的方式依次轮流进行压缩。

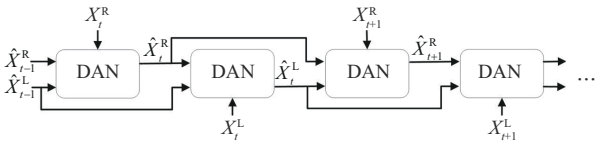



图1　立体双视图视频每帧压缩流程

本文的左右视图采用相同的压缩编码网络，如图2所示，本文以右视图为例来详细论述本文提出的DAN总体框架，其中，虚线框部分为由LGEDB组成的编解码模块和DHFFM，虚线连接代表3个编解码模块的参数共享。首先，将当前编码帧$X_t^{\mathrm{R}}$、时间相邻的重构帧$\hat{X}_{t-1}^{\mathrm{R}}$和视图相邻的重构帧$\hat{X}_{t-1}^{\mathrm{L}}$输入特征提取模块，得到对应的特征图$F_t^{\mathrm{R}}$、$F_{t-1}^{\mathrm{R}}$和$F_{t-1}^{\mathrm{L}}$。再将当前编码帧的特征$F_t^{\mathrm{R}}$和时间相邻的重构帧的特征$F_{t-1}^{\mathrm{R}}$送入运动估计模块，得到运动信息$M_t^{\mathrm{R}}$，然后将运动信息$M_t^{\mathrm{R}}$送入基于LGEDB的编解码模块得到运动重构信息$\hat{M}_t^{\mathrm{R}}$，再对运动重构信息$\hat{M}_t^{\mathrm{R}}$和$F_t^{\mathrm{R}}$进行运动补偿得到运动补偿特征$F^{\mathrm{M}}$。

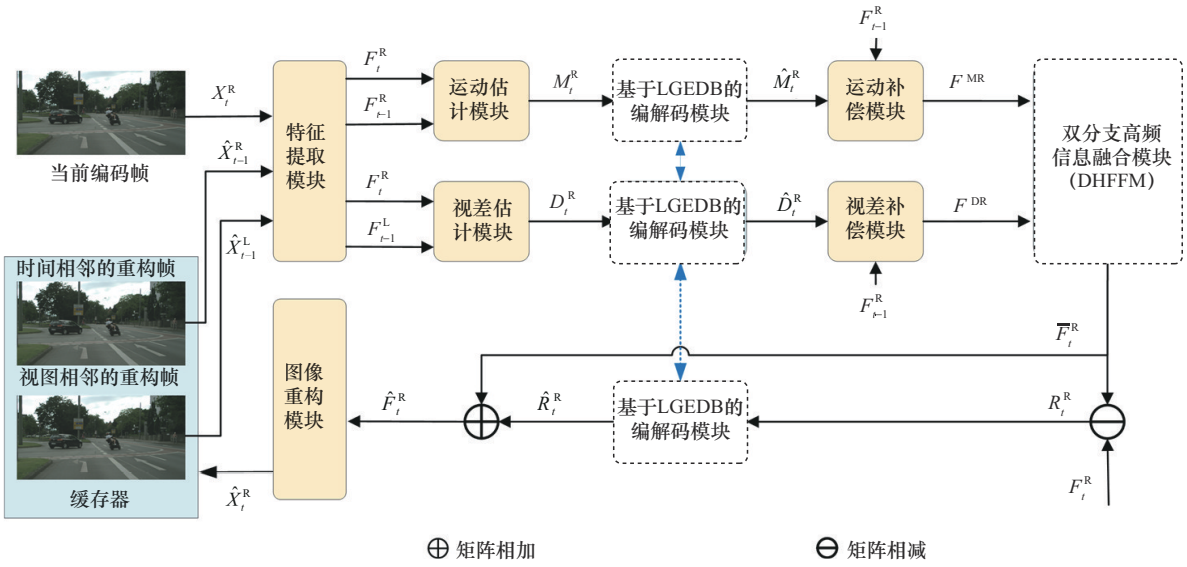



图2　DAN总体框架(以右视图为例)

同理，将当前编码帧的特征 $F_t^{\mathrm{R}}$ 和视图相邻的重构帧的特征 $F_{t-1}^{\mathrm{L}}$ 送入视差估计模块，得到视差信息 $D_t^{\mathrm{R}}$，将视差信息 $D_t^{\mathrm{R}}$ 送入基于LGEDB的编解码模块得到视差重构信息 $\hat{D}_t^{\mathrm{R}}$，再对视差重构信息 $\hat{D}_t^{\mathrm{R}}$ 和 $F_t^{\mathrm{R}}$ 进行视差补偿便得到视差补偿特征 $F^{\mathrm{D}}$。接下来，将 $F^{\mathrm{M}}$ 和 $F^{\mathrm{D}}$ 送入DHFFM进行特征融合得到预测特征 $\bar{F}_t^{\mathrm{R}}$，并将 $\bar{F}_t^{\mathrm{R}}$ 和 $F_t^{\mathrm{R}}$ 相减得到残差信息 $R_t^{\mathrm{R}}$，即 $R_t^{\mathrm{R}} = F_t^{\mathrm{R}} - \bar{F}_t^{\mathrm{R}}$。再将残差信息 $R_t^{\mathrm{R}}$ 送入基于LGEDB的编解码模块得到残差重构信息 $\hat{R}_t^{\mathrm{R}}$。最后将残差重构信息 $\hat{R}_t^{\mathrm{R}}$ 与预测特征 $\bar{F}_t^{\mathrm{R}}$ 相加得到重构特征 $\hat{F}_t^{\mathrm{R}}$，即 $F_t^{\mathrm{R}} = \hat{R}_t^{\mathrm{R}} + \bar{F}_t^{\mathrm{R}}$，并将重构特征进行图像重构便得到重建图像 $\hat{X}_t^{\mathrm{R}}$，并将重建图像放入缓存器中。其中，特征提取模块、运动估计模块、运动补偿模块、视差估计模块、视差补偿模块以及图像重构模块均采用与文献[6]相同的设置，而基于LGEDB的编解码模块和DHFFM是本文的主要贡献和创新。因此，接下来，本文将详细论述基于LGEDB的编解码模块和DHFFM。

## 2.2　局部和全局双分支编解码块

现有的绝大多数立体双视图视频压缩编码模型仅使用纯卷积操作来构建编解码器，然而纯卷积操作不仅无法对非重复纹理的复杂区域进行有效的特征提取，同时还无法捕捉长距离的全局信息，导致压缩性能不佳。本文提出的LGEDB能够兼顾非重复复杂区域纹理特征和全局信息的有效提取。

如图5所示，本文提出的LGEDB是由局部窗口注意力分支（LWAB, local window attention branch）和全局通道注意力分支（GCAB, global channel attention branch）组成，分别通过LWAB和GCAB来有效提取局部的非重复复杂区域纹理特征和全局信息。

对于LWAB而言，本文使用局部的窗口注意力块获取局部特征，如图3所示。给定输入特征 $F \in R^{H \times W \times C}$，先将输入特征划分为 $p \times p$ 大小的不重叠窗口，即 $X^k \in R^{p^2 \times C}$，$k = 1,2,\cdots,L$（$L = \dfrac{H \times W}{p^2}$，即划分的块数）。再通过一个线形层将 $X^k$ 从通道方

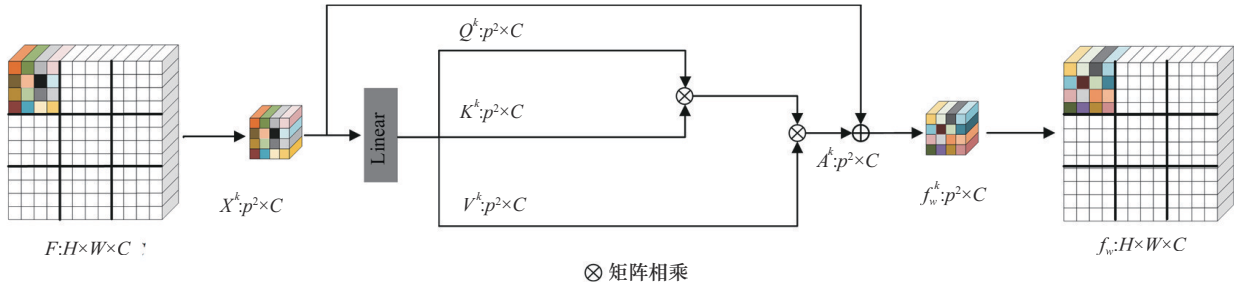



图3　局部的窗口注意力块

向等分为多个头，生成查询、键和值矩阵：$Q_s^k, K_s^k, V_s^k \in R^{p^2 \times d}$，$d = \frac{C}{S}$为每个头分到的通道数，$s = 1,2,\cdots,S$，$S$为头的数量，生成式分别为

$$Q_s^k = X^k P_Q^{k,s} \tag{1}$$

$$K_s^k = X^k P_K^{k,s} \tag{2}$$

$$V_s^k = X^k P_V^{k,s} \tag{3}$$

其中，$P_Q^{k,s}, P_K^{k,s}, P_V^{k,s} \in R^{C \times d}$分别代表第$k$个块中第$s$个头的线性投影矩阵。同时，不同块之间的$P_Q^{k,s}, P_K^{k,s}, P_V^{k,s}$的参数共享。接下来，计算第$k$个块中第$s$个头的特征相似性$A_s^k \in R^{p^2 \times d}$，表示为

$$A_s^k = \mathrm{Softmax}\left(\frac{Q_s^k\left(K_s^k\right)^{\mathrm{T}}}{\sqrt{d}} + E\right)V_s^k \tag{4}$$

其中，$E$是相对位置编码。因为有$S$个头，所以并行执行$S$次式(1)~式(4)的自注意力计算，然后将每个头计算得到的注意力特征$A_s^k$从通道维度上进行拼接得到特征$A^k \in R^{p \times p \times C}$，再将特征$A^k$与$X^k$相加得到第$k$个块的自注意力特征$f_w^k \in R^{P^2 \times C}$，表示为

$$f_w^k = A^k + X^k \tag{5}$$

最后将每个块的自注意力特征拼接回原来的位置得到整幅图像的LWAB输出特征$f_w \in R^{H \times W \times C}$，表示为

$$f_w = \mathrm{concat}\left(f_w^1, f_w^2, \cdots, f_w^L\right) \tag{6}$$

其中，concat(·)表示位置拼接。

由式(1)~式(6)和图3可知，LWAB是通过计算每个窗口内每个像素点与该窗口内其他像素点之间的相似性来构建该局部窗口的特征。因此，如果该局部窗口内存在重复的纹理，那么其相似性较高，经过LWAB后的响应会较大；反之，如果该局部窗口内存在非重复的纹理，那么其经过LWAB后的响应会较小。因此，LWAB是能够准确表达局部区域内的非重复纹理特征的。

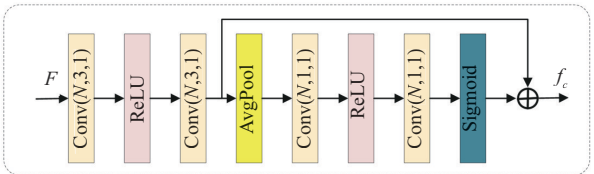
对于GCAB而言，本文采用了全局通道注意力块的思想来实现全局信息的有效提取。如图4(a)所示，给定输入特征$F \in R^{H \times W \times C}$，其计算过程为

$$f' = \mathrm{Conv3}\left(\mathrm{ReLU}\left(\mathrm{Conv3}(F)\right)\right) \tag{7}$$

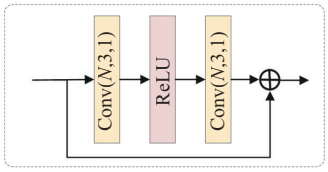
$$f'' = \mathrm{ReLU}\left(\mathrm{Conv1}\left(\mathrm{AvgPool}\left(f'\right)\right)\right) \tag{8}$$

$$f_c = \mathrm{Sigmoid}\left(\mathrm{Conv1}\left(f''\right)\right) + f' \tag{9}$$

其中，Conv3(·)代表3×3的标准卷积操作，Conv1(·)代表1×1的标准卷积操作，ReLU(·)代表ReLU激活函数，Sigmoid(·)代表Sigmoid激活函数。




(b) 全局通道注意力块




(b) 残差块

图4　全局通道注意力块和残差块

在分别经过了LWAB和GCAB之后，本文采用拼接、卷积和残差块（RB, residual block）对提取的局部特征和全局信息进行融合。

$$f = \mathrm{RB}\left(\mathrm{RB}\left(\mathrm{RB}\left(\mathrm{Conv3}\left(\mathrm{cat}\left(f_c, f_w\right)\right)\right)\right)\right) \tag{10}$$

其中，cat(·)表示通道拼接，RB(·)表示与文献[44]相同的残差块，如图4(b)所示。LGEDB的最终输出为

$$F_{\mathrm{LG\text{-}out}} = f + F \tag{11}$$

在实现了对非重复复杂区域纹理特征和全局信息的有效提取之后，本文将提出的LGEDB作为基础块，并结合通道自回归熵模型[45]来构建编解码模块，如图5所示，其中，Conv（$N$,$k$,$s$）为卷积操作，$N$为通道数，$k$为卷积核大小，$s$为步长，TConv（$N$,$k$,$p$）为转置卷积操作，$N$为通道数，$k$为卷积核大小，$p$为填充大小，RB为残差块。构建的编解码模块将分别作用于运动信息$M_t^{\mathrm{R}}$、视差信息$D_t^{\mathrm{R}}$和残差信息$R_t^{\mathrm{R}}$来生成对应的重构信息。以运动信息$M_t^{\mathrm{R}}$为例，输入$M_t^{\mathrm{R}}$，首先由LGEDB组成的编码器提取潜在表示$y$。然后$y$分别经过量化、CWAEM、算数编码（AE, arithmetic encoding）、算数解码（AD, arithmetic decoding）和由LGEDB组成的解码器得到运动重构信息$\hat{M}_t^{\mathrm{R}}$。该过程可表示为

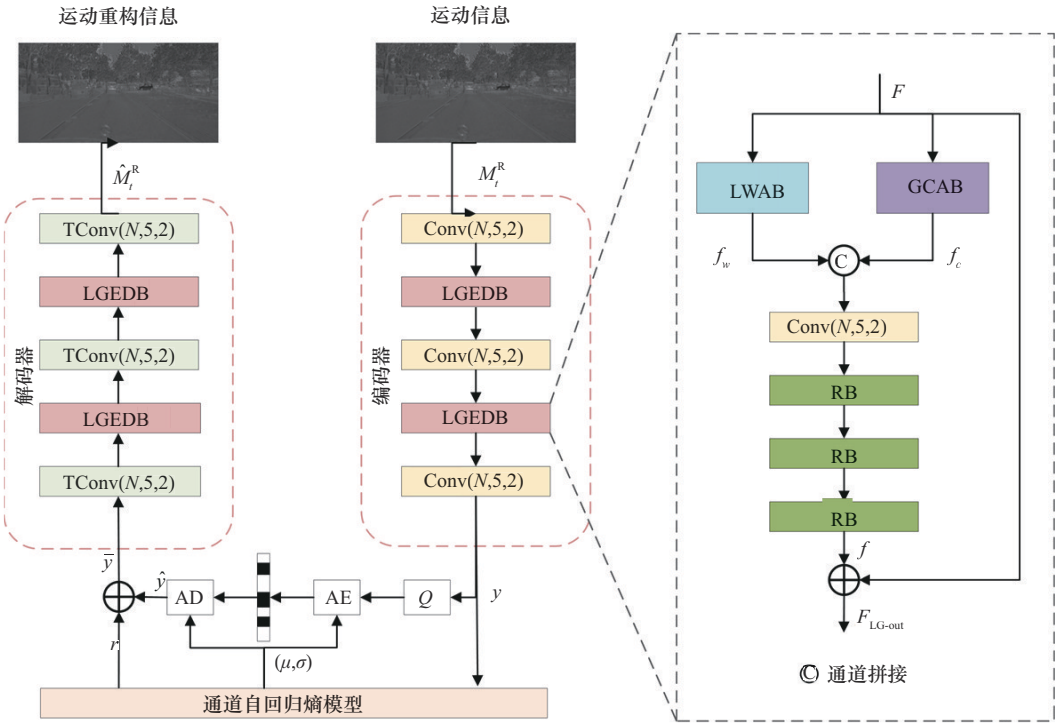



图5　编解码模块

$$y = E(M_t^{\mathrm{R}}) \tag{12}$$

$$\left(r,(\mu,\sigma)\right) = \mathrm{CWAEM}(y) \tag{13}$$

$$\hat{y} = \mathrm{AD}\left(\mathrm{AE}\left(Q(y),\mu,\sigma\right)\right) \tag{14}$$

$$\hat{M}_t^{\mathrm{R}} = D(\hat{y} + r) \tag{15}$$

其中，$E(\cdot)$和$D(\cdot)$分别表示由LGEDB组成的编码器和解码器，$Q(\cdot)$表示量化过程，CWAEM$(\cdot)$表示通道自回归熵模型，$\mu$、$\sigma$和$r$分别表示均值、方差和潜在空间残差，AE$(\cdot)$表示算数编码过程，AD$(\cdot)$表示算术解码过程。

在运动信息$M_t^{\mathrm{R}}$和视差信息$D_t^{\mathrm{R}}$分别经过基于LGEDB的编解码模块后，得到的运动重构信息$\hat{M}_t^{\mathrm{R}}$和视差重构信息$\hat{D}_t^{\mathrm{R}}$再经过运动补偿模块和视差补偿模块得到运动补偿特征$F^{\mathrm{M}}$和视差补偿特征$F^{\mathrm{D}}$，随后使用DHFFM融合$F^{\mathrm{M}}$和$F^{\mathrm{D}}$。

### 2.3　双分支高频信息融合模块

除了在编解码器中对局部非重复纹理信息和全局信息进行有效提取之外，对运动补偿信息和视差补偿信息的有效融合是影响立体双视图视频压缩编码质量的另一个关键因素。基于此，本文提出了一种新颖的DHFFM，以实现运动补偿特征和视差补偿特征中高频信息的有效提取和特征的准确筛选，从而实现运动补偿特征和视差补偿特征的高效融合。

Gomez等[46]已经证明INN的逆向变换特性能够确保高频信息在正向传递和逆向还原时不丢失。Zhou等[47]提出的INN块能够有效提取出高频信息。此外，Chen等[48]提出的NAFBlock（nonlinear activation free block）中的门控机制能够实现准确的特征筛选。因此，如图6所示（虚线表示区域中2个NAFBlock的参数共享，以及2个INN模块的参数共享），本文提出的DHFFM首先采用INN[47]和NAFBlock[48]分别对运动补偿信息$F^{\mathrm{M}}$和视差补偿信息$F^{\mathrm{D}}$进行高频信息的提取和特征的筛选。然后，将$F^{\mathrm{M}}$中提取的高频信息和筛选的特征与$F^{\mathrm{D}}$中提取的高频信息和筛选的特征分别进行逐元素相加。最后，再将相加后的高频信息与筛选特征经过拼接、卷积和残差块来实现运动补偿特征和视差补偿特征的高效融合，得到预测特征$\bar{F}_t^{\mathrm{R}}$。该过程可表示为

$$\varphi^{\mathrm{M}} = \mathrm{NAFBlock}\left(F^{\mathrm{M}}\right) \tag{16}$$

$$\varphi^{\mathrm{D}} = \mathrm{NAFBlock}\left(F^{\mathrm{D}}\right) \tag{17}$$

$$\varphi = \mathrm{NAFBlock}\left(\varphi^{\mathrm{M}} + \varphi^{\mathrm{D}}\right) \tag{18}$$

$$\theta = \mathrm{INN}\left(\mathrm{INN}\left(F^{\mathrm{M}}\right) + \mathrm{INN}\left(F^{\mathrm{D}}\right)\right) \tag{19}$$

$$F' = \mathrm{Conv3}\left(\mathrm{Conv1}\left(\mathrm{cat}\left(\varphi,\theta\right)\right)\right) \tag{20}$$

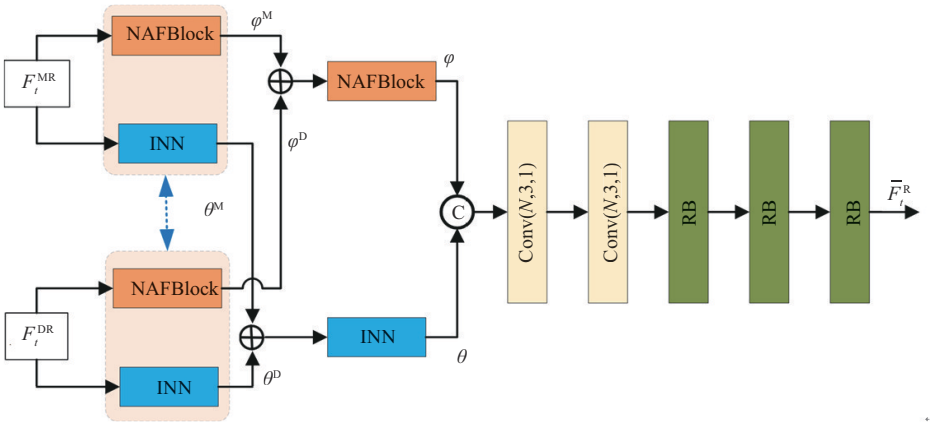



图 6 双分支高频信息融合模块

$$\bar{F}_t^{\mathrm{R}} = \mathrm{RB}\left(\mathrm{RB}\left(\mathrm{RB}\left(F'\right)\right)\right) \tag{21}$$

其中，NAFBlock( · )代表文献[48]中的NAFBlock，INN( · )代表文献[47]中的可逆神经网络。

由以上分析可知，由于INN和NAFBlock的引入，本文提出的DHFFM能够有效提取出运动补偿特征和视差补偿特征中的高频信息，以及逐像素点的特征筛选，从而实现运动补偿特征和视差补偿特征的高效融合，保证了双视图立体视频压缩编码性能的进一步提升。

在经过DHFFM得到预测特征$\bar{F}_t^{\mathrm{R}}$之后，将$\bar{F}_t^{\mathrm{R}}$与原始特征$F_t^{\mathrm{R}}$相减得到残差信息$R_t^{\mathrm{R}}$，即$R_t^{\mathrm{R}} = F_t^{\mathrm{R}} - \bar{F}_t^{\mathrm{R}}$。再将残差信息$R_t^{\mathrm{R}}$送入基于LGEDB的编解码模块得到残差重构信息$\hat{R}_t^{\mathrm{R}}$。随后，将残差重构信息$\hat{R}_t^{\mathrm{R}}$与预测特征$\bar{F}_t^{\mathrm{R}}$相加，得到重构特征$\hat{F}_t^{\mathrm{R}}$，即$\hat{F}_t^{\mathrm{R}} = \hat{R}_t^{\mathrm{R}} + \bar{F}_t^{\mathrm{R}}$，并通过图像重构模块得到重建图像$\hat{X}_t^{\mathrm{R}}$，该重建图像将被存入缓存器中，供下一帧的压缩编码使用。

### 2.4 损失函数

本文提出的DAN的损失函数为

$$L = \sum_{v\in\{\mathrm{L,R}\}}\sum_{t}\lambda D\left(X_t^v, \hat{X}_t^v\right) + R\left(y_{M,t}^v\right) + R\left(y_{D,t}^v\right) + R\left(y_{R,t}^v\right) \tag{22}$$

其中，$D(\cdot)$为失真损失，表示重构帧与原始帧之间的失真，本文使用均方误差（MSE, mean square error）损失和MS-SSIM[18]损失来计算失真损失；$R(\cdot)$表示压缩过程中的比特率，上标$v$表示使用左视图或者右视图，下标$t$表示时间步长，$t\in\{1,2,3,4\}$；$X_t^v$表示在视图$v$和时间$t$时的原始图像，$\hat{X}_t^v$表示原始图像对应的重构图像，$y_{M,t}^v$表示对应视图和时间的量化后运动特征，$y_{D,t}^v$表示对应视图和时间的量化后视差特征，$y_{R,t}^v$表示对应视图和时间的量化后残差特征；$\lambda$是控制率失真超参数，本文使用4个不同的$\lambda$值，即$\lambda$ = 512,1 024, 2 048,4 096。

## 3 实验结果和分析

在本节中，进行了大量的实验来验证本文提出的DAN的有效性和优越性，同时还进行了消融实验，以此来评估本文提出的贡献点的有效性。

### 3.1 实验设置

#### 3.1.1 数据集和客观评价指标

在本文所有实验中，均是首先使用单视图视频数据集Vimeo-90K进行预训练，然后再使用立体视频数据集Cityscape[8]的训练集进行训练。本文使用立体双视图视频Cityscape[8]的测试集、KITTI 2012[9]和KITTI 2015[10]进行测试和性能评估。

立体视频数据集Cityscape[8]包含训练集和测试集，分别有2 975和1 525个立体视频序列对，每个立体视频序列对包含30帧，每张图片的分辨率为2 048像素 × 1 024像素。KITTI 2012[9]和KITTI 2015[10]分别包含195和200个立体视频序列对，每个立体视频序列对包含21帧，每张图片的分辨率为1 241像素 × 376像素。本文遵循文献[6]的图片预处理过程，对Cityscape测试集和KITTI分别进行预处理，预处理后所有帧的分辨率分别对应为1 920像素 × 768像素 和 1 152像素 × 256像素。本文方法以及所有比较方法均采用此数据划分和裁剪策略。

本文使用BPP来测量率失真，使用PSNR和MS-SSIM指标评估重建帧的保真度。本文使用

Bjøntegaard-Delta Rate(BD-BR)来表示相同质量下，2种方法的比特率节省情况（BD-BR的值越小越好）。

#### 3.1.2　训练流程和设置

本文遵循LSVC[6]的训练策略，分3个阶段训练模型。首先，在第一个阶段，使用Vimeo-90 K数据集预训练特征提取模块、运动估计模块、基于LGEDB的编解码模块、运动补偿模块和图像重构模块。以$5 \times 10^{-5}$的学习率进行200万次迭代。然后，使用第一阶段的预训练权重初始化DAN，使用Cityscape训练集和$1 \times 10^{-5}$的学习率对基于LGEDB的编解码模块和DHFFM进行40万次迭代，并冻结其他模块的参数。最后，使用Cityscape训练集和$5 \times 10^{-6}$的学习率先对整个DAN进行30万次迭代，再调整学习率为$1 \times 10^{-5}$进行30万次迭代，接着调整学习率为$5 \times 10^{-6}$进行30万次迭代。

此外，本文采用随机旋转90°、180°、270°和水平翻转来对训练数据集进行数据增强，并且所有方法均采用相同的数据增强策略。采用Adam[49]优化器。在预训练阶段，批量大小为8，输入的Vimeo-90K的大小被设置为256 × 256。后续阶段，批量大小为4，输入的Cityscape训练集大小为512 × 256。在测试过程中，对第一帧的压缩使用文献[50]的方法进行压缩，并且按照文献[50]的训练方式在Cityscape训练集上进行微调。本文的所有实验是在NVIDIA 4090 GPU和Pytorch[51]深度学习框架上进行训练和测试的。

### 3.2　消融实验

如前所述，本文提出的贡献点为LGEDB和DHFFM。因此，在本节中，将在Cityscape数据集上进行消融实验，以此来验证本文提出的LGEDB和DHFFM的有效性。消融实验中所有模型的设置和训练细节都一致。为了公平，本文首先创建了一个基线模型Baseline，即从DAN中移除LWAB、GCAB、NAFBlock和INN，其余成分保持不变：1) 将LGEDB编解码模块中的LWAB和GCAB移除，输入的特征$F$直接经过一个卷积层+3个残差块；2) 将DHFFM中的NAFBlock和INN移除，即运动补偿特征$F^{\mathrm{M}}$和视差补偿特征$F^{\mathrm{D}}$直接通道拼接，将拼接后的特征经过2个卷积层和3个残差块。

#### 3.2.1　LGEDB的有效性消融实验

本文提出的LGEDB很好地实现了非重复复杂区域纹理和全局信息的有效提取，能够有效降低比特率。因此，为了能够准确评估LGEDB的有效性，本文构建了一个新的网络模型Baseline+LGEDB：只将Baseline中的编解码模块替换为由本文提出的LGEDB组成的编码模块，其余成分保持不变。由此可见，Baseline和Baseline+LGEDB的区别就仅仅在于是否采用了LGEDB，因此，这两者间的性能对比是能够准确反映LGEDB的有效性。DAN的有效性消融实验如表1所示，其中$\lambda$为512，与Baseline相比，Baseline+LGEDB的PSNR提高了0.09 dB，且BPP降低了0.003。表1很好地证明了本文提出的LGEDB的有效性。由图5可知，LGEDB由GCAB和LWAB 2个分支组成，为了进一步验证GCAB和LWAB的有效性，本文又创建了2个模型：DAN-No-GCAB和DAN-No-LWAB。其中DAN-No-GCAB表示从DAN中仅移除GCAB，其余成分保持不变；而DAN-No-LWAB则表示从DAN中仅移除LWAB，其余成分保持不变。由表1可知，DAN-No-GCAB的PSNR为37.63 dB，BPP为0.035（PSNR和BPP相比DAN分别降低了0.09 dB和0.001）；DAN-No-LWAB的PSNR为37.68 dB，BPP为0.036（PSNR和BPP相比DAN分别降低了0.04 dB和0.002）。显而易见，在本文DAN中，移除任一分支都会导致性能下降，这表明GCAB和LWAB均对模型性能有显著贡献。这一结果验证了LGEDB的双分支设计在平衡图像质量和压缩效率方面的必要性。

**表1　　DAN的有效性消融实验**

| 模型 | 参数量/MB | FLOPS/GB | PSNR/dB (↑) | BPP (↓) |
|---|---|---|---|---|
| Baseline | 12.632 | 2 840 | 37.50 | 0.039 |
| Baseline+LGEDB | 14.268 | 3 169 | 37.59 | 0.036 |
| Baseline+DHFFM | 12.729 | 2 909 | 37.57 | 0.038 |
| DAN-No-GCAB | 13.972 | 3 120 | 37.63 | 0.035 |
| DAN-No-LWAB | 14.101 | 3 169 | 37.68 | 0.036 |
| DAN | 14.365 | 3 238 | 37.72 | 0.034 |

#### 3.2.2　DHFFM的有效性消融实验

为了准确评估本文提出的DHFFM的有效性，本文在Baseline的基础上又创建了一个新的模型Baseline+DHFFM：只将Baseline的融合模块替换为本文提出的DHFFM，其余成分保持不变。比较结果如表1所示，在模型的参数量和计算复杂度略高的情况下，Baseline+DHFFM的PSNR较Baseline

提升了 0.07 dB，BPP 降低了 0.001，很好地证明了本文提出的 DHFFM 的有效性。

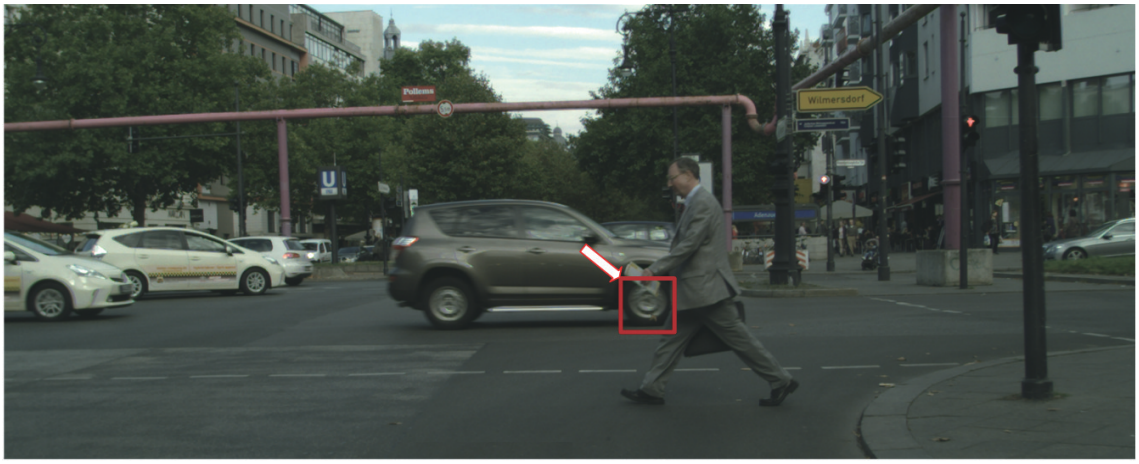
### 3.2.3　各贡献点有效性的可视化比较

为了能够更直观地证明本文提出的贡献点的有效性，将消融实验中的模型 Baseline、Baseline+LGEDB、Baseline+DHFFM、DAN-No-GCAB、DAN-No-LWAB 和 DAN 的重构结果进行了主观视觉效果的可视化比较，结果如图 7 所示。从图 7 中可以很明显看到，随着 LGEDB 和 DHFFM 的逐步加入，重构图像的主观视觉效果呈现出明显的上升趋势，如车胎位置的树叶轮廓逐步清晰（如图 7 中的放大区域所示）。此外，DAN-No-GCAB 和 DAN-No-LWAB 相较于 DAN，重构图像的主观视觉效果较弱，如车胎位置的树叶颜色较为暗淡（如图 7 中的放大区域所示），这一结果证明了 GCAB 和 LWAB 的有效性。最终，本文提出的 DAN 获得了最佳的可视化重构结果，最接近原始图像。图 7 从主观视觉效果方面证明了本文提出的贡献点的有效性。

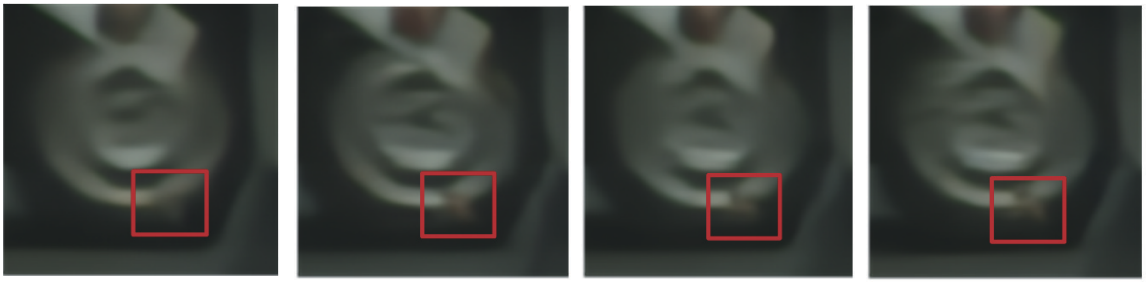
## 3.3　与前沿方法的比较实验

### 3.3.1　定量（客观评价指标）比较

为了验证本文方法的优越性，本节将本文方法与近年来最先进的 7 种方法进行了比较：LSVC[6]、LLSS[7]、FVC[37]、DCVC[38]、VCT[42]、H.265[52]和 MV-HEVC[53]。其中，LSVC[6]、LLSS[7]、FVC[37]、



(a) Cityscape: berlin_000002_000003_rightImg8bit原图



(b) Baseline消融图　(c) Baseline+LGEDB消融图　(d) Baseline+DHFFM消融图　(e) DAN-No-GCAB消融图

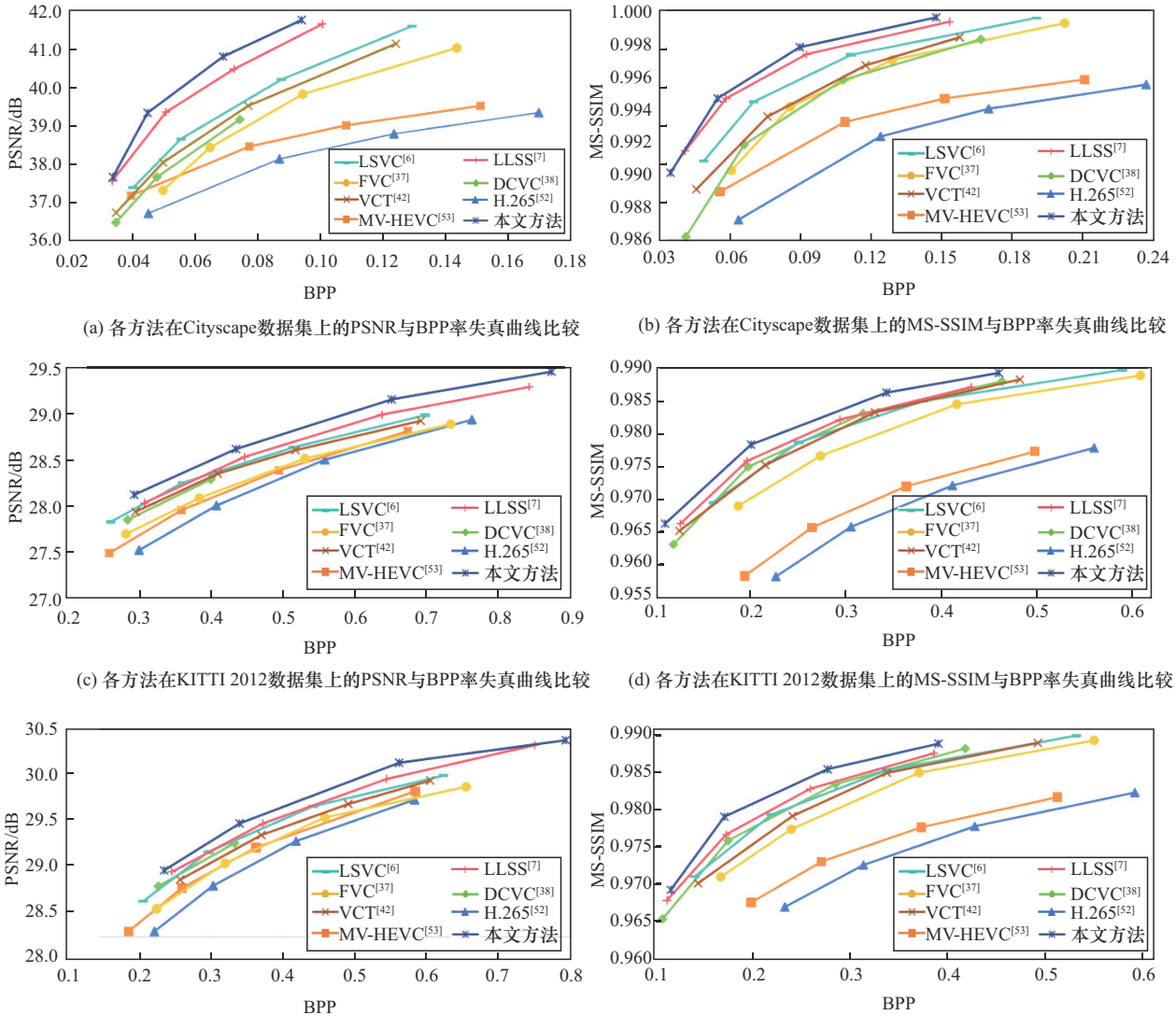
(f) DAN-No-LAWB消融图　(g) DAN主观视觉效果图　(h) GT图

图 7　各贡献点消融模型的主观视觉效果

DCVC[38]和 VCT[42]是基于 DNN 的方法，H.265[52]和 MV-HEVC[53]是传统方法。H.265[2]选择使用 HM-16.20[52]，配置选择“lowdelay_P_main”；MV-HEVC[4]选择使用 HTM-16.3[53]，配置为“baseCfg_2_view”。对于基于 DNN 的方法，使用本文提供的模型，并在 Cityscape 训练集上按照学习率为 $5 \times 10^{-6}$ 进行了 30 万次迭代微调。在测试过程中，测试集的分辨率等设置保持一致。

图 8 展示了 LSVC[6]、LLSS[7]、FVC[37]、DCVC[38]、VCT[42]、H.265[52]、MV-HEVC[53]和本文方法在 Cityscape[8]、KITTI 2012[9]和 KITTI 2015[10]数据集上的率失真曲线。如图8所示，本文方法的率失真曲线位于最上方，明显优于其他方法。

表2给出了LSVC[6]、LLSS[7]、FVC[37]、DCVC[38]、VCT[42]、H.265[52]和本文方法分别在 Cityscape[8]、KITTI 2012[9]和 KITTI 2015[10]数据集上的 BD-BR（BD-BR 的基线选择 MV-HEVC[51]）结果。由表 2 可知，首先，本文方法与基于 Transformer 的视频压缩模型 VCT[42]相比，在 Cityscape[8]、KITTI 2012[9]和 KITTI 2015[10]数据集上，BD-DR 分别提升了 27.8%、11.9% 和 14.4%。其次，本文方法在所有测试集上均优于近年来极具代表性的先进方法。与第二好的 LLSS[7]方法相比，本文方法在 Cityscape[8]、KITTI 2012[9]和 KITTI 2015[10]数据集上，BD-DR 分别提升了 3.6%、7.3% 和 7.5%。图 8 和表 2 在 Cityscape[8]、KITTI 2012[9]和 KITTI 2015[10]数据集上的数据，从客观评价指标方面很好地证明了本文方法的优越性和泛化性。




(a) 各方法在Cityscape数据集上的PSNR与BPP率失真曲线比较

(b) 各方法在Cityscape数据集上的MS-SSIM与BPP率失真曲线比较

(c) 各方法在KITTI 2012数据集上的PSNR与BPP率失真曲线比较

(d) 各方法在KITTI 2012数据集上的MS-SSIM与BPP率失真曲线比较

(e) 各方法在KITTI 2015数据集上的PSNR与BPP率失真曲线比较

(f) 各方法在KITTI 2015数据集上的MS-SSIM与BPP率失真曲线比较

图8　各方法分别在Cityscape[8]、KITTI 2012[9]和KITTI 2015[10]数据集上的PSNR和MS-SSIM与BPP率失真曲线比较

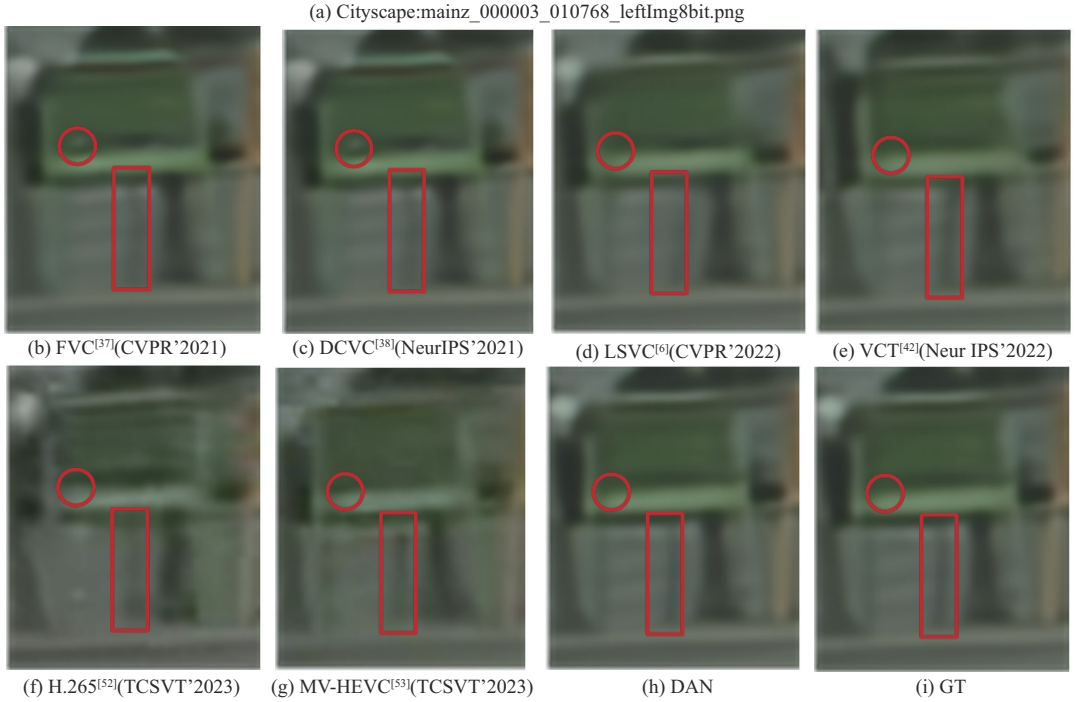
**表 2　各方法在数据集 Cityscape、KITTI 2012 和 KITTI 2015 上的 BD-BR 结果（以 MV-HEVC 为基线）**

| 模型 | Cityscape | KITTI 2012 | KITTI 2015 |
|---|---|---|---|
| FVC[37] (CVPR'2021) | –15.6 | –2.3 | 1.0 |
| DCVC[38] (NeurIPS'2021) | –15.2 | –13.7 | –12.3 |
| VCT[42] (NeurIPS'2022) | –26.4 | –13.6 | –8.9 |
| LSVC[6] (CVPR'2022) | –32.7 | –17.1 | –13.4 |
| H.265[52] (TCSVT'2023) | 33.3 | 7.9 | 12.7 |
| LLSS[7] (CVPR'2024) | –50.6 | –18.2 | –15.8 |
| 本文方法 | **–54.2** | **–25.5** | **–23.3** |

### 3.3.2　定性（主观视觉效果）比较

为了能够更加全面地证明本文方法的优越性，除了客观评价指标之外，本文还在主观视觉效果上将本文方法与 LSVC[6]、FVC[37]、DCVC[38]、VCT[42]、H.265[52] 和 MV-HEVC[53] 进行了比较，比较结果如图 9~图 12 所示。如图 10 的局部放大图所示，LSVC[6]、FVC[37]、DCVC[38]、VCT[42]、H.265[52] 和 MV-HEVC[53] 重构出的窗帘均存在不同程度的变形失真。如图 11 所示，在数据集 KITTI 2012 中，LSVC[6]、FVC[37]、DCVC[38]、VCT[42]、H.265[52] 和 MV-HEVC[53] 重构出的字母“T”的边缘出现了不同程度的变形失真，具有明显的拖尾和与字母“S”的粘连效应。如图 12 所示，对于数据集 KITTI 2015 而言，LSVC[6]、FVC[37]、DCVC[38]、VCT[42]、H.265[52] 和 MV-HEVC[53] 重构出的字牌内容均存在不同程度的模糊，无法重建

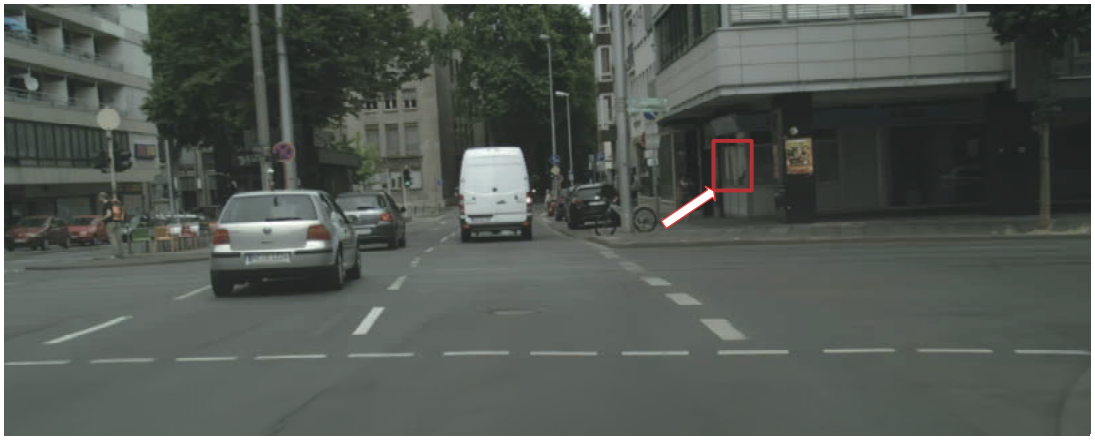
(a) Cityscape:mainz_000003_010768_leftImg8bit.png



(b) FVC[37](CVPR'2021)　(c) DCVC[38](NeurIPS'2021)　(d) LSVC[6](CVPR'2022)　(e) VCT[42](Neur IPS'2022)

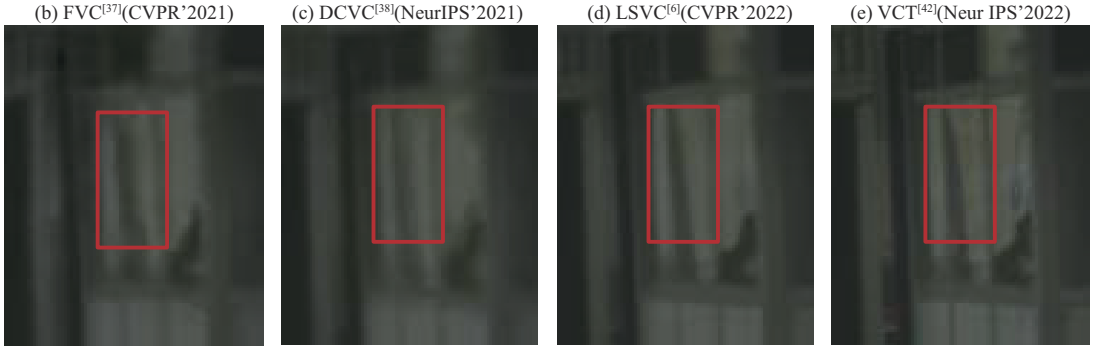
(f) H.265[52](TCSVT'2023)　(g) MV-HEVC[53](TCSVT'2023)　(h) DAN　(i) GT

图 9　各方法在 Cityspace 数据集上的主观视觉效果比较（椅子放大视图）



(a) Cityscape: mainz_000003_010768_leftImg8bit.png

(b) FVC[37](CVPR'2021)　(c) DCVC[38](NeurIPS'2021)　(d) LSVC[6](CVPR'2022)　(e) VCT[42](Neur IPS'2022)



(f) H.265[52](TCSVT'2023)　(g) MV-HEVC[53](TCSVT'2023)　(h) DAN　(i) GT

图 10　各方法在 Cityspace 数据集上的主观视觉效果比较（窗帘放大视图）

出锐化的边缘轮廓和细节。相比之下，本文方法不仅能有效降低比特率，而且能够重建出最准确的图像，尤其是针对复杂的边缘和纹理细节区域，是最接近原始图像的方法。图 9~图 12 从主观视觉效果方面很好地证明了 DAN 的优越性，即能够实现对局部复杂的非重复纹理细节和全局信息的准确捕捉，重构出更高质量的图像。同时，DAN 在 Cityscape[8]、KITTI 2012[9] 和 KITTI 2015[10]数据集上的主观视觉效果图也很好地证明了其泛化性。

### 3.3.3　参数量、计算复杂度和运行时间的比较

除了对方法性能的评估之外，本文还评估了 DAN 与 FVC[29]、DCVC[32]、LSVC[6]和 VCT[42]的参数量、计算复杂度和运行时间（输入立体双视图视频分辨率为 1 024 像素×512 像素）。实验设置为 Intel i7-7820X CPU 和单个 NVIDIA 2080Ti GPU。通过表 3 可知，本文方法参数量为 14.37 MB，为第二低，但是其计算复杂度和运行时间却较多。这很可能是因为本文提出的 LWAB 采用了窗口自注意力，而该自注意力相比卷积运算具有更多的乘法和加法的运算操作。同时，本文提出的 DHFFM 采用了较多的深度可分离卷积，从而导致内存访问次数的增加[54]，而内存的速度远低于 GPU 和 CPU 的计算速度，因此导致 DAN 的运行时间较多。本文提出的 DAN 是不会影响部署的时间，但是会影响实际工作中的运行时间，还无法达到实时性的效率要求。

基于此，如何能够在保证高性能和低参数量的同时有效降低网络的计算复杂度和运行时间成为未来研究工作的重点。目前，可行的轻量化设计和优化策略包括：1) 缩小 LWAB 中窗口自注意力的窗口大小；2) 用轻量级的 NAFBlock 代替 LWAB；3) 减少 DAN 中深度可分离卷积的数量，减少访存次数，以缩短 DAN 的运行时间。

**表 3　各方法的参数量、计算复杂度和运行时间**

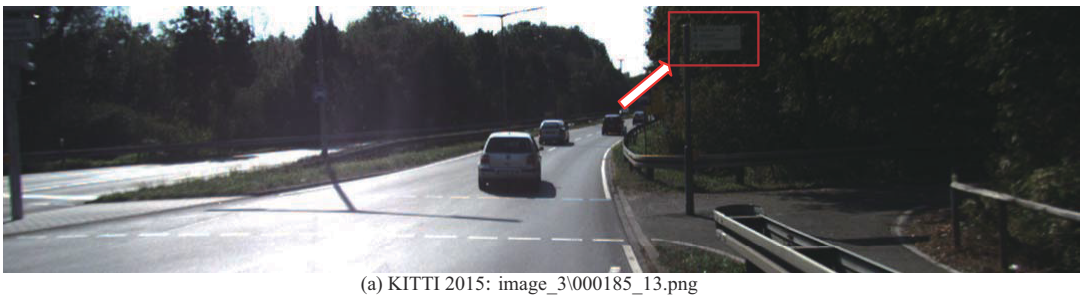
| 模型 | 参数量/MB | FLOPS/GB | 运行时间/ms |
|---|---|---|---|
| FVC[37] (CVPR'2021) | 26 | 462 | 201 |
| DCVC[38] (NeurIPS'2021) | 6.53 | 603 | 618 |
| VCT[42] (NeurIPS'2022) | 2 | 1 340 | 596 |
| LSVC[6] (CVPR'2022) | 27.76 | 2 423 | 443 |
| DAN | 14.37 | 3 238 | 539 |

(a) KITTI 2012: image_3\000143_20.png

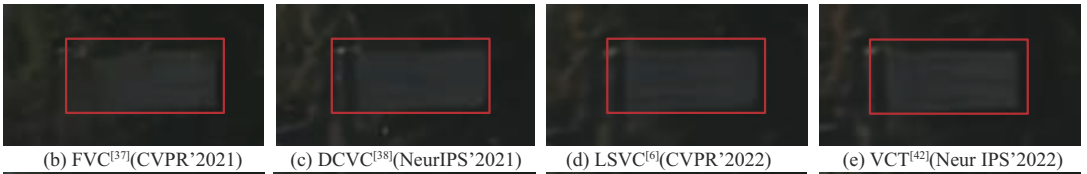
(b) FVC[37](CVPR'2021)　(c) DCVC[38](NeurIPS'2021)　(d) LSVC[6](CVPR'2022)　(e) VCT[42](Neur IPS'2022)

(f) H.265[52](TCSVT'2023)　(g) MV-HEVC[53](TCSVT'2023)　(h) DAN　(i) GT

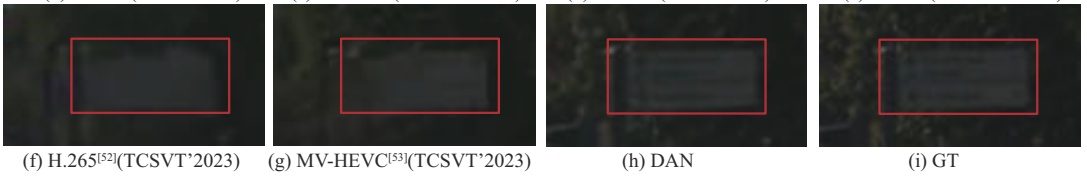
图 11　各方法在 KITTI 2012 数据集上的主观视觉效果比较



(a) KITTI 2015: image_3\000185_13.png



(b) FVC[37](CVPR'2021)　(c) DCVC[38](NeurIPS'2021)　(d) LSVC[6](CVPR'2022)　(e) VCT[42](Neur IPS'2022)



(f) H.265[52](TCSVT'2023)　(g) MV-HEVC[53](TCSVT'2023)　(h) DAN　(i) GT

图 12　各方法在 KITTI 2015 数据集上的主观视觉效果比较

## 4　结束语

本文提出了一种基于 Transformer 和通道注意力机制的 LGEDB，通过融合局部范围内像素点的自注意力和全局通道注意力，实现对局部非重复纹理细节和全局信息的精准捕捉。同时，还提出了一种基于可逆神经网络和门控机制的 DHFFM，该模块能够有效提取运动补偿特征和视差补偿特征中的高频信息，并通过逐像素特征的筛选，实现两类特征的高效融合。首先，相较于现有的基于卷积的立体编码方法，DAN 采用局部–全局协同注意力机制，解决了局部卷积算子导致的特征表征碎片化问题。其次，相比混合卷积–Transformer 框架，DAN 创新性地引入可逆神经网络进行运动–视差补偿特征融合，通过可逆性保证高频信息无损传递，配合门控机制实现的特征空间动态结构化建模，彻底突破了传统方法在跨模态融合时简单线性叠加的局限性。在 Cityscape、KITTI 2012 和 KITTI 2015 数据集上的大量实验结果证明，本文方法能够在保证较低比特率的同时，实现更高质量的图像重建，与当前最佳方法 LLSS 相比，本文方法的 BD-BR 分别提升了 3.6%、7.3% 和 7.5%。这些提升意味着在相同带宽条件下，DAN 重建的高保真不仅能显著提升自动驾驶系统的实时感知精度，还能有效消除 VR 等应用场景中的运动模糊和边缘伪影，从而极大增强用户的沉浸式体验。这些是目前其他方法未能同时达到的关键性能改进。然而，值得注意的是，DAN 在实现高质量图像重建的同时，仍存在计算复杂度较高、运行时间较长的问题。因此，如何在保持高重建质量和低比特率的基础上，进一步降低计算复杂度并加速网络的运行时间，将是未来研究工作的重点方向。

## 参考文献：


[1] WIEGAND T, SULLIVAN G J, BJONTEGAARD G, et al. Overview of the H.264/AVC video coding standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2003, 13(7): 560-576.

[2] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding (HEVC) standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12): 1649-1668.

[3] VETRO A, WIEGAND T, SULLIVAN G J. Overview of the stereo and multiview video coding extensions of the H. 264/MPEG-4 AVC standard[J]. Proceedings of the IEEE, 2011, 99(4): 626-642.

[4] TECH G, CHEN Y, MÜLLER K, et al. Overview of the multiview and 3D extensions of high efficiency video coding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 26(1): 35-49.

[5] KIM I K, MIN J, LEE T, et al. Block partitioning structure in the HEVC standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12): 1697-1706.

[6] CHEN Z H, LU G, HU Z H, et al. LSVC: a learning-based stereo video compression framework[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 6063-6072.

[7] HOU Q Q, FARHADZADEH F, SAID A, et al. Low-latency neural stereo streaming[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 7974-7984.

[8] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 3213-3223.

[9] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2012: 3354-3361.

[10] MENZE M, GEIGER A. Object scene flow for autonomous vehicles[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 3061-3070.

[11] WALLACE G K. The JPEG still picture compression standard[J]. Communications of the ACM, 1991, 34(4): 30-44.

[12] TAUBMAN D S, MARCELLIN M W. JPEG2000: standard for interactive imaging[J]. Proceedings of the IEEE, 2002, 90(8): 1336-1357.

[13] TODERICI G, VINCENT D, JOHNSTON N, et al. Full resolution image compression with recurrent neural networks[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 5435-5443.

[14] MINNEN D C, BALLÉ J, TODERICI G. Joint autoregressive and hierarchical priors for learned image compression[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, 31: 10794-10803.

[15] HE D L, YANG Z M, PENG W K, et al. ELIC: efficient learned image compression with unevenly grouped space-channel contextual adaptive coding[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 5708-5717.

[16] MUCKLEY M J, EL-NOUBY A, ULLRICH K, et al. Improving statistical fidelity for neural image compression with implicit local likelihood models[J]. arXiv Preprint, arXiv: 2301.11189, 2023.

[17] WANG Z, SIMONCELLI E P, BOVIK A C. Multiscale structural similarity for image quality assessment[C]//Proceedings of the Thrity-Seventh Asilomar Conference on Signals, Systems & Computers. Piscataway: IEEE Press, 2003: 1398-1402.

[18] NAN R L, SUN G L, ZHENG B W, et al. Hybrid transformer and con-

volution for image compressed sensing[J]. Electronics, 2024, 13(17): 3496.s

[19] 张新岩, 祝勇俊, 吴宏杰, 等. 基于并行Transformer和CNN的图像压缩感知重构网络[J]. 科技导报, 2025, 43(2): 108-116.
ZHANG X Y, ZHU Y J, WU H J, et al. A parallel Transformer-CNN network for image compression sensing reconstruction[J]. Science & Technology Review, 2025, 43(2): 108-116.

[20] HU H, SHI Y H, WANG J, et al. Feature enhanced spherical transformer for spherical image compression[J]. Displays, 2025, 88: 103002.

[21] MENTZER F, TODERICI G, TSCHANNEN M, et al. High-fidelity generative image compression[J]. arXiv Preprint, arXiv: 2006.09965, 2020.

[22] AGUSTSSON E, TSCHANNEN M, MENTZER F, et al. Generative adversarial networks for extreme learned image compression[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 221-231.

[23] AGUSTSSON E, MINNEN D, TODERICI G, et al. Multi-realism image compression with a conditional generator[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 22324-22333.

[24] LU G, OUYANG W L, XU D, et al. DVC: an end-to-end deep video compression framework[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 10998-11007.

[25] HABIBIAN A, ROZENDAAL T V, TOMCZAK J, et al. Video compression with rate-distortion autoencoders[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 7032-7041.

[26] AGUSTSSON E, MINNEN D, JOHNSTON N, et al. Scale-space flow for end-to-end optimized video compression[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 8500-8509.

[27] LI J H, LI B, LU Y. Hybrid spatial-temporal entropy modelling for neural video compression[C]//Proceedings of the 30th ACM International Conference on Multimedia. New York: ACM Press, 2022: 1503-1511.

[28] LI J H, LI B, LU Y. Neural video compression with diverse contexts[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 22616-22626.

[29] RIPPEL O, ANDERSON A G, TATWAWADI K, et al. ELF-VC: efficient learned flexible-rate video coding[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 14459-14468.

[30] HU Z H, LU G, GUO J Y, et al. Coarse-to-fine deep video coding with hyperprior-guided mode prediction[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 5911-5920.

[31] LE H, ZHANG L, SAID A, et al. MobileCodec: neural inter-frame video compression on mobile devices[C]//Proceedings of the 13th ACM Multimedia Systems Conference. New York: ACM Press, 2022: 324-330.

[32] 周立新. 基于深度学习的视频压缩编码技术研究[J]. 电视技术, 2024, 48(9): 13-16.
ZHOU L X. Research on video compression coding technology based on deep learning[J]. Video Engineering, 2024, 48(9): 13-16.

[33] LIU J, WANG S L, URTASUN R. DSIC: deep stereo image compression[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 3136-3145.

[34] DENG X, YANG W Z, YANG R, et al. Deep homography for efficient stereo image compression[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 1492-1501.

[35] LEI J J, LIU X R, PENG B, et al. Deep stereo image compression via bi-directional coding[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 19637-19646.

[36] WÖDLINGER M, KOTERA J, XU J, et al. SASIC: stereo image compression with latent shifts and stereo attention[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 651-660.

[37] HU Z H, LU G, XU D. FVC: a new framework towards deep video compression in feature space[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 1502-1511.

[38] LI J H, LI B, LU Y. Deep contextual video compression[J]. Advances in Neural Information Processing Systems, 2021, 34: 18114-18125.

[39] 杜秀丽, 朱金耀, 高星, 等. 一种3D可变形卷积结合Transformer的视频压缩感知方法[J]. 计算机科学, 2024: 1-11.
DU X L, ZHU J Y, GAO X, et al. A video compression sensing method based on 3D deformable convolution and Transformer[J]. Computer Science, 2024: 1-11.

[40] YU Y, HE X H, WU X H, et al. Learned video compression with channel-wise autoregressive entropy model[J]. Journal of Electronic Imaging, 2023, 32: 063013.

[41] CHEN Z H, RELIC L, AZEVEDO R, et al. Neural video compression with spatio-temporal cross-covariance transformers[C]//Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM Press, 2023: 8543-8551.

[42] MENTZER F, TODERICI G, MINNEN D, et al. VCT: a video compression transformer[J]. Advances in Neural Information Processing Systems, 2022, 35: 13091-13103.

[43] AMEUR Z, DEMARTY C H, MÉNARD D, et al. 3R-INN: how to be climate friendly while consuming/delivering videos? [C]//European Conference on Computer Vision. Berlin: Springer, 2024: 146-163.

[44] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.

[45] MINNEN D, SINGH S. Channel-wise autoregressive entropy models for learned image compression[C]//Proceedings of the 2020 IEEE In-

ternational Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2020: 3339-3343.

[46] GOMEZ A N, REN M Y, URTASUN R, et al. The reversible residual network: backpropagation without storing activations[J]. arXiv Preprint, arXiv: 1707.04585, 2017.

[47] ZHOU M, HUANG J, FANG Y C, et al. RETRACTED: pan-sharpening with customized transformer and invertible neural network[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 3553-3561.

[48] CHEN L Y, CHU X J, ZHANG X Y, et al. Simple baselines for image restoration[C]//European Conference on Computer Vision. Berlin: Springer, 2022: 17-33.

[49] KINGMA D P, BA J. Adam: a method for stochastic optimization[C]// Proceedings of the 3rd International Conference on Learning Representations. Vancouver: ICLR, 2015: 6-20.

[50] CHENG Z X, SUN H M, TAKEUCHI M, et al. Learned image compression with discretized Gaussian mixture likelihoods and attention modules[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 7936-7945.

[51] PASZKE A, GROSS S, MASSA F, et al. PyTorch: an imperative style, high-performance deep learning library[J]. arXiv Preprint, arXiv: 1912.01703, 2019.

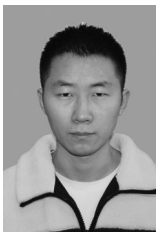
[52] HEVC. Hevc test model (HM)[R]. 2024.

[53] MV HEVC. Multiview high efficiency video coding (MV-HEVC)[R]. 2024.

[54] WU B C, WAN A, YUE X Y, et al. Shift: a zero FLOP, zero parameter alternative to spatial convolutions[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 9127-9135.

[作者简介]

**唐述**（1981– ），男，重庆人，博士，重庆邮电大学副教授、博士生导师，主要研究方向为深度学习、图像超分辨率重建、模糊图像复原、视频压缩等。



**赵瑜**（1997– ），男，四川广元人，重庆邮电大学硕士生，主要研究方向为深度学习、视频压缩。

**杨书丽**（1995– ），女，河南驻马店人，重庆邮电大学博士生，主要研究方向为深度学习、图像超分辨率重建。

**谢显中**（1966– ），男，四川巴中人，博士，重庆邮电大学教授、博士生导师，主要研究方向为信号与信息处理、计算机通信、通信与信息系统。