基于多尺度时域交叉注意力的视频压缩感知重构网络

周超，陈灿，张登银

（南京邮电大学物联网学院，南京 210003）

摘要：视频压缩感知能够在编码端以低于奈奎斯特采样定律要求的采样率获取测量值，从而显著降低采集与编码的计算复杂度和功耗，在资源受限的视频采集与传输场景中具有重要应用价值。然而，欠采样导致的测量信息不足使重构问题呈现典型的病态特性，难以仅依赖单帧信息恢复出高质量视频。为了解决此问题，本文提出一种基于多尺度时域交叉注意力的视频压缩感知重构网络，通过充分挖掘视频信号的时空冗余实现高质量重构。针对不同场景下运动幅度与作用范围存在的差异性，本文在注意力融合阶段引入并行多尺度卷积表征，以兼顾不同尺度下的时空信息整合；为控制长序列带来的计算开销，网络采用线性注意力降低计算复杂度并提升推理效率；为了抑制递归重构过程中的不可靠信息并缓解误差随时间累积扩散，本文采用一种门控卷积前馈模块，将局部卷积建模与自适应门控机制结合，对跨帧传播的特征进行选择性增强与抑制。大量实验结果表明，与现有方法相比，本文所提出方法在多种采样率下都取得了更优异的重构性能。

关键词：压缩感知；视频重构；注意力机制；特征对齐；时域相关性；

中图分类号：TP391.41 文献标志码：A

引 言

压缩感知通过将采样与编码合并为一步，在低于奈奎斯特采样定律要求的采样率下完成信号观测，从而显著降低编码端的计算复杂度[1,2]，非常适用于无线传感网[3,4]、磁共振成像[5]等资源受限型应用。然而，欠采样观测带来的信息不足使解码端的信号重构呈现典型的病态特性，如何从有限的测量值中恢复出高质量原始信号是该领域的研究重点。

在视频场景下，由于视频序列在时间维度具有天然的连续性，相邻帧之间往往包含大量可被利用的时域冗余信息。因此，视频压缩感知重构的关键在于如何有效建模并利用帧间时域相关性，即如何从相邻帧中提取与当前帧内容高度相关的有效信息并进行有效融合，以补充欠采样观测中缺失的信息，从而提升重构质量[6]。

在传统视频压缩感知重构算法中，时域相关性的利用主要依靠人工设计的先验模型，例如多假设预测模型[7-13]、组稀疏[14]、残差稀疏[15]或者 3D 稀疏[16,17]等。然而，此类方法通常需要通过迭代优化求解，计算复杂度较高，难以满足实时性要求高的应用场景。与此同时，人工先验的表达能力往往受限于固定的统计假设，难以充分覆盖真实视频中复杂的时空变化。当先验与数据分布不匹配时，重构性能与稳定性容易受到限制。

基金项目：江苏省基础研究计划青年基金项目（BK20250682）；国家自然科学基金面上项目（62471241）。

收稿日期：2026-02-25；修订日期：2026-07-03

近年来，深度学习方法凭借较强的特征表示学习能力，逐渐成为视频压缩感知重构的主流研究方向。与传统迭代优化方法相比，此类方法能够从数据中自动挖掘更为有效的重构先验，并在保证重构质量的同时降低推理阶段的计算开销。目前，深度神经网络中对时域相关性的建模方式大致可以分为三类。第一类是将时域相关的视频帧直接级联，依靠神经网络自身的学习能力去挖掘帧间相关性来帮助重构。例如，VCSNet 构建了一种多级特征补偿框架，将关键帧中高采样率特征与非关键帧低采样率特征级联，以此提升非关键帧的重构质量[18]；TSRN 则利用时序位移操作，将相邻帧的部分特征图级联到当前帧，完成时域信息的高效交换[19]。此类方法虽然计算较为高效，但是时域依赖的建立依靠感受野有限的卷积操作，因此面对大尺度运动往往效果并不理想。第二类方法则利用运动估计/运动补偿来完成时域关系建模。例如，DUMHAN[20]在最大后验推导的深度展开框架下，采用了分阶段光流估计实现显式运动对齐。并且，通过多假设机制融合多组预测以强化帧内/帧间相关性建模，从而提升整体的重构性能。STM-Net 为提升非关键帧重构效果，引入遮挡感知的时域对齐与多帧融合机制，并采用改进的 PWCNet 估计光流实现显式跨帧对齐，以增强帧间信息补偿的有效性[21]。ABPN 则设计了一种双向传播机制，在每一个传播方向上利用可学习的光流估计进行运动补偿，并利用注意力机制帮助对齐后的特征进行融合[22]。上述方法对时域信息的利用高度依赖于光流估计的结果。但是，在视频压缩感知重构任务中，光流估计只能在粗糙的初始重构上进行，因此估计结果可能存在较大误差，从而影响重构性能。第三类方法则不再局限于利用显式的运动估计/运动补偿方法，而是隐式的建立时域依赖帮助完成时域相关性建模。文献[23]设计了一种静态与动态域先验增强的两阶段重构网络，并在对齐模块中结合可变形卷积[24]与注意力机制，构建一种塔式级联结构来对动态域先验进行建模。CollabNet[25]则首先根据像素域的光流估计结果进行粗对齐，然后再在特征域中通过可学习偏移的可变形卷积进行进一步细化对齐与融合。上述两种方法都借助可变形卷积来对时域相关性进行建模，但是，可变形卷积的对齐效果高度依赖于对偏移量的准确预测，因此同样容易受低质量初始重构的影响。FNLAN 则利用了非局部网络来实现长距离时域依赖的建立[26]，虽然计算复杂度较高，但是能够有效捕捉大尺度运动。非局部网络[27]的核心思想与 Transformer[28]类似，同样基于自注意力机制，但是缺少多头注意力、位置编码等关键机制，因此，在一定程度上限制了其在时域相关性建模上的表现。

为更有效地建模长距离时域依赖，本文提出一种基于多尺度时域交叉注意力的视频压缩感知重构网络。不同于传统 Transformer 仅在同一特征内部进行信息交互，该网络通过当前帧与相邻帧之间的交叉注意力，引入跨帧互补信息以提升重构质量。此外，为进一步适配视频压缩感知重构的特点，本文针对三个问题做出改进：（1）由于不同场景下运动幅度与作用范围差异较大，单一感受野往往难以同时兼顾不同尺度的帧间时域信息。为此，本文在交叉注意力的值分支引入并行多尺度卷积表征，为时域信息融合提供不同感受野下的候选内容，以此增强被融合内容的多尺度表达能力。（2）标准 Transformer 注意力的计算与存储开销随特征序列长度呈二次增长($O(N^2)$)，在处理高分辨率特征时对硬件资源要求较高。为降低复杂度，本文采用线性注意力[29]替代标准注意力以实现高效的时域信息交互。同时，在前馈网络中引入深度卷积以增强局部建模能力，缓解线性注意力不够聚焦的问题。(3) 为提高时域信息传播效率，视频重构算法往往采用 RNN 进行递归更新。但与此同时，当前帧的重构误差也可能随传播逐步累积而影响后续重构。为此，本文引入门控卷积前馈模块，通过自适应门控对传播特征进行有效筛选，抑制不可靠信息的传播并缓解误差扩散。

1 网络结构

本文所提出的基于多尺度时域交叉注意力的视频压缩感知重构网络整体结构如图 1 所示。

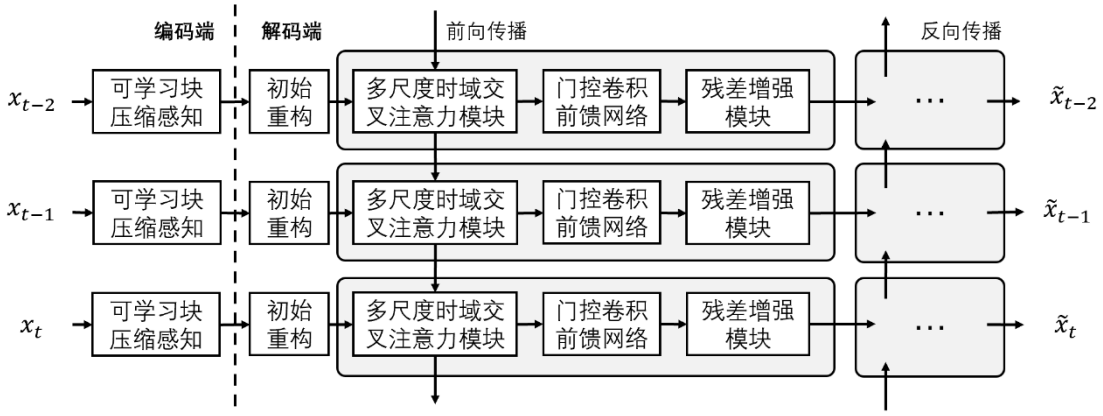

图 1 基于多尺度时域交叉注意力的视频压缩感知重构网络示意图

Fig.1 Overall architecture of the multi-scale temporal cross-attention network for video compressive sensing reconstruction

在编码端，本文通过一个可学习的块压缩感知模块实现对原始视频信号的观测。传统视频压缩感知所采用的块压缩感知[30]可由下述公式表示：

$$y_{t,i} = \Phi x_{t,i}, \tag{1}$$

其中，$x_{t,i}$为视频帧t中第i个图像块，$y_{t,i}$为其测量值，Φ为所有图像块共享的一个固定参数的随机采样矩阵。在可学习块压缩感知模块中，采样矩阵Φ由一个步长和卷积核大小固定为B（B为图像块大小）的卷积层进行参数化替代，并与后续解码端的重构网络一起进行端到端训练，从数据中自适应的学习到一个更合适的采样算子。

在解码端，初始重构模块首先对测量值$y_t \in R^{\frac{H}{B}\times\frac{W}{B}\times([SR\cdot B^2])}$进行初始重构（$H\times W$为原始帧尺寸，$SR$为采样率，$[SR\cdot B^2]$为每一个图像块的测量通道数，若$SR\cdot B^2$不为整数时向上取整）。此模块通过一个 1×1 卷积将通道维度从$SR\cdot B^2$扩张到B^2，然后再通过重塑操作将维度恢复成原始图像尺寸$R^{H\times W\times 1}$。由于初始重构结果仍较为粗糙，本文采用两个方向的时域传播对其进行逐步增强以获得最终重构结果。在每一个传播方向上，首先通过本文提出的多尺度时域交叉注意力模块，引入并利用相邻帧的时域相关信息以增强当前特征；随后接入门控卷积前馈网络，在对注意力输出进行逐位置非线性变换的基础上，对传播特征进行自适应筛选与更新，抑制不可靠信息的传递并缓解误差扩散；最后再级联由多个残差块组成的残差增强模块进行细节精炼与高频补偿。多尺度时域交叉注意力与门控卷积前馈网络将在后续章节中进一步说明。

1.1 多尺度时域交叉注意力

1.1.1 时域交叉注意力机制

传统 Transformer 中的自注意力机制可以由下式表示：

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V, \tag{2}$$

其中，Q代表查询矩阵，K和V分别代表键矩阵和值矩阵，$\sqrt{d_k}$为缩放因子。Q,K,V均来自于同一特征，主要

用于对特征内部的全局依赖关系进行建模与信息交互。而在视频压缩感知重构任务中，主要目标为对当前帧和相邻帧的时域关系进行建模，从相邻帧中查询对当前帧有益的高相关信息，并用于帮助当前帧重构。因此，本文对自注意力机制进行修改，提出更适合视频压缩感知的时域交叉注意力机制：

$$\mathrm{Attention}(Q_t, K_{(t,t-1)}, V_{(t,t-1)}) = \mathrm{softmax}(\frac{Q_t K_{(t,t-1)}^T}{\sqrt{d_k}}) V_{(t,t-1)}, \tag{3}$$

其中，Q_t作为查询矩阵仅来自于当前帧t，$K_{(t,t-1)}$和$V_{(t,t-1)}$则来自于当前帧t和相邻帧$t-1$组成的集合（在前向传播中）。将公式（3）按位置展开，有

$$\alpha_{t,i,j} = \frac{\exp(\frac{q_{t,i}^T k_{(t,t-1),j}}{\sqrt{d_k}})}{\sum_{m=1}^{N} \exp(\frac{q_{t,i}^T k_{(t,t-1),m}}{\sqrt{d_k}})}, \tag{4}$$

$$o_{t,i} = \sum_{j=1}^{N} \alpha_{t,i,j} v_{(t,t-1),j}, \tag{5}$$

其中，公式（4）的$\alpha_{t,i,j}$表示当前帧t的第i个位置上对当前帧和上一帧的集合中第j个位置分配的权重，N为将当前帧与相邻帧特征展平并拼接后的总位置数。公式（5）的$o_{t,i}$为当前帧t第i个位置的加权输出。

此机制以当前帧位置i的查询向量$q_{t,i}$为条件，在当前帧和上一帧的所有候选位置$\{k_{(t,t-1),j}\}_{j=1}^{N}$上计算相关性$q_{t,i}^T k_{(t,t-1),j}$，从而得到归一化权重$\{\alpha_{t,i,j}\}$。随后，对候选库中的值向量$v_{(t,t-1),j}$进行加权聚合，得到当前帧位置$i$的跨帧增强表示$o_{t,i}$。该过程本质上等价于在由$\{t, t-1\}$两帧构成的候选库中，自适应选择与当前位置内容最相关的特征进行融合，从而将相邻帧中对当前帧重构有用的补充信息引入当前特征表示。值得注意的是，本文以前向传播为例进行说明，此时相邻帧取$t-1$；后向传播时相邻帧对称地取$t+1$。

1.1.2 多尺度上下文内容表征

在时域交叉注意力中，查询/键分支主要用于对时空相关性进行度量，然后据此从候选库$\{k_{(t,t-1),j}\}_{j=1}^{N}$中检索与当前内容匹配的特征。而值分支则承载被检索和参与融合的内容表征，直接影响融合表征的质量。由于不同场景下运动幅度与作用范围差异较大，对值分支捕捉不同尺度时空信息的能力提出了更高要求。若值分支仅采用单一感受野进行内容建模，容易使候选库表征偏向某一尺度，从而限制其对复杂运动模式的适应能力。

因此，本文在时域交叉注意力的值分支引入并行多尺度卷积，提取候选位置的多尺度表征。具体的，首先将$V_{(t,t-1)}$按通道划分成四组$\{V_{(t,t-1)}^{(g)}\}_{g=1}^{4}$，每组通道数一致，随后每组通过一个对应尺寸的卷积层提取不同感受野的上下文信息，最后再在通道维度进行拼接融合，并经过一个1×1融合卷积$Conv_s$得到增强过后的多尺度值表征$V'_{(t,t-1)}$。以四分支结构为例，此过程可写为：

$$U_{(t,t-1)}^{(g)} = \mathrm{Conv}_{k_g \times k_g}(V_{(t,t-1)}^{(g)}), \quad k_g \in \{1,3,5,7\}, \tag{6}$$

$$V'_{(t,t-1)} = \mathrm{Conv}_s(\mathrm{Concat}(U_{(t,t-1)}^{(1)}, U_{(t,t-1)}^{(2)}, U_{(t,t-1)}^{(3)}, U_{(t,t-1)}^{(4)}))。 \tag{7}$$

值得注意的是，本文仅在值分支引入多尺度建模，而未扩展到查询/键分支，主要基于两点考虑：（1）查询/键分支用于相似性度量，过度的多尺度变换可能会破坏特征空间的一致性，使相关性计算不够稳定；（2）值分支承载待融合的时空上下文，在此注入多尺度信息能更直接地增强特征的表达能力。

1.1.3 线性注意力计算

标准自注意力模型由于需要计算相关性矩阵QK^{T}，其计算与存储开销随特征序列长度N呈二次增长

$O(N^2)$。在时域交叉注意力中，候选库$\{k_{(t,t-1),j}\}_{j=1}^{N}$由当前帧与相邻帧共同构成，序列长度显著增加，并且，视频压缩感知重构任务通常需要处理较高分辨率的信号，进一步加重了计算负担。为降低复杂度并提升推理效率，本文采用线性注意力[29]来替代标准注意力，并通过增强网络的局部感知能力来补偿一定的潜在性能损失。

线性注意力的核心思想是通过引入非负核映射$\phi(\cdot)$替代公式（3）中 softmax 形式的相似性计算，避免直接计算$Q_t K_{(t,t-1)}^T$，从而将计算与存储开销从$O(N^2)$降低到与N近似线性相关。本文采用常用映射函数

$$\phi(x) = \mathrm{ELU}(x)+1, \tag{8}$$

其中，$\mathrm{ELU}(\cdot)$为指数线性单元。然后，将公式（3）重写为线性注意力形式：

$$\mathrm{Attention}(Q_t, K_{(t,t-1)}, V'_{(t,t-1)}) = \frac{\phi(Q_t)\left(\phi(K_{(t,t-1)})^T V'_{(t,t-1)}\right)}{\phi(Q_t)\left(\phi(K_{(t,t-1)})^T \mathbf{1}\right)+\varepsilon}, \tag{9}$$

其中，$\mathbf{1}$为全 1 的向量，ε为一个小常数，用于数值稳定。上述重写使得与候选库相关的项（$\phi\left(K_{(t,t-1)}\right)^T V'_{(t,t-1)}$和$\phi\left(K_{(t,t-1)}\right)^T \mathbf{1}$）可以先在序列维度进行汇总，随后各个位置的查询向量仅需与汇总结果进行乘法和点积即可得到输出[29]。整个时域交叉注意力模块如图 2 所示。

值得注意的是，虽然线性注意力有效降低了计算复杂度和内存开销，但是相比于 softmax，简单的核映射$\phi(\cdot)$通常会使得注意力不够集中，进而导致对局部细节的建模偏弱[31][32]。为了解决此问题，本文采用文献[32]的建议，在前馈网络中增加深度卷积以增强局部建模能力，并将其融入到下一小节介绍的门控机制中，从而在保证效率的同时尽量减少重构质量的损失。

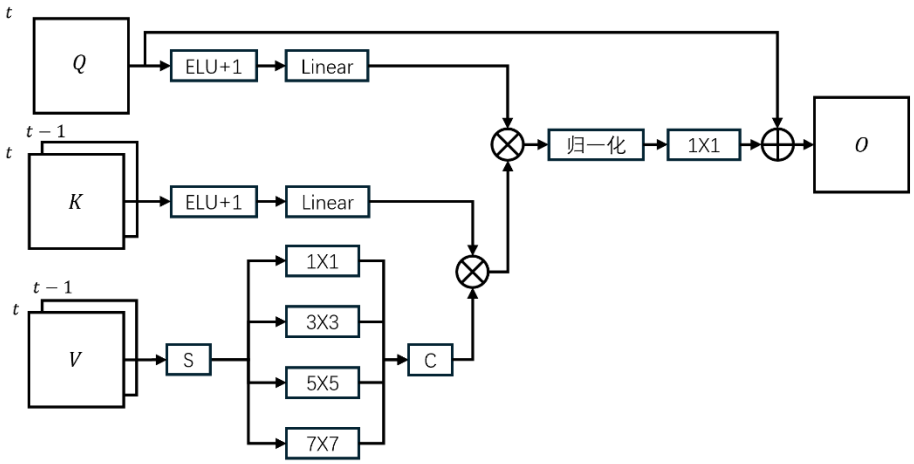

图 2 时域交叉注意力结构示意图

Fig.2 Diagram of the temporal cross-attention structure

1.2 门控卷积前馈网络

本文采用了双向传播的递归重构框架（如图 1 所示）。当前帧更新后的特征会作为时域先验传递至下一帧，帮助后续帧的重构。此机制虽然能实现高效的时域信息传递，但是也面临误差累计传播的风险。若某一帧的重构存在较大误差，那么该误差不仅会影响下一帧的重构，还可能随递归过程传播并逐帧扩散，从而损害全序列重构的精度与稳定性。为了解决此问题，本文对在时域交叉注意力之后的前馈网络进行修

改，增加了门控线性单元[33]来对传播特征进行自适应的筛选与更新，尽量保留有用信息并抑制不可靠更新。

具体结构如图 3 所示。给定输入特征X，在经过层归一化后，首先通过两路线性映射进行通道维度的扩展，生成特征分支和门控分支。特征分支用于提供待更新的内容表征，门控分支则用于生成逐位置的调制权重。每一个分支再单独经过一个深度卷积，以补充邻域信息并缓解线性注意力分布较平滑时带来的细节刻画不足[32]。然后，门控分支单独通过 GELU[34]进行非线性激活，并与特征分支进行逐像素相乘，实现对特征响应的自适应调制。最后，经过一个线性映射将通道维度恢复到原始大小，得到输出特征。整个过程可以表示为：

$$\tilde{X} = \mathrm{LN}(X), \tag{10}$$

$$X_f = \mathrm{DWConv}_f(W_f\tilde{X}), \tag{11}$$

$$X_g = \mathrm{DWConv}_g(W_g\tilde{X}), \tag{12}$$

$$Y = W_{out}(X_f \odot GELU(X_g)), \tag{13}$$

其中，$\mathrm{LN}(\cdot)$代表层归一化，W_f和W_g、DWConv_f和DWConv_g分别代表特征分支及门控分支的线性映射、深度卷积，⊙代表逐像素相乘，W_{out}为输出映射。

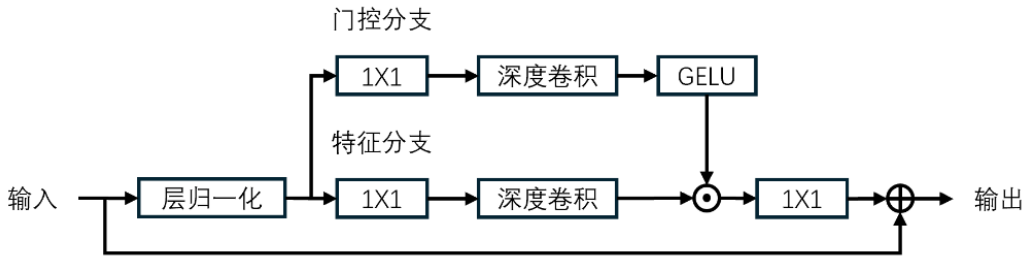

图 3 门控卷积分支结构示意图

Fig.3 Diagram of the gated convolutional branch.

2 实验结果分析

2.1 实验设置

本文采用视频重构任务中常用的 REDS 数据集作为训练集，对本文所提出的 MTCAN 网络进行训练。测试集则采用 REDS4 及 Vid4。在训练过程中，输入视频被随机裁剪成 256×256 大小，批次大小为 2。优化器采用 Adam[35]，其中β_1和β_2分别设置为 0.9 和 0.99。学习率设定为 2×10^{-4}，并最后降低至 1×10^{-5}。损失函数采用 Charbonnier 损失[36]。视频序列长度在训练时被设定为 8，测试时设定为 16。除非特殊说明，所有实验均基于 PyTorch 框架，在一张 NVIDIA RTX 4090 显卡上实现。

本文所提网络的部分关键参数设置如下：可学习块压缩感知中的块大小设置为 32，即采样卷积的核大小和步长都为 32。对于每个图像块，测量通道数为$\lceil SR \times 32^2 \rceil$，（$\lceil\cdot\rceil$为向上取整符号）。时域交叉注意力中多尺度并行卷积的卷积核大小为$\{1,3,5,7\}$；常数ε被设置为 1×10^{-6}；注意力头的数量为 8，每个头的维度为 32。在时域交叉注意力之前会首先通过一个基于块的预对齐模块[37]。残差增强模块的残差块个数为 5。

2.2 实验结果分析

2.2.1 重构精度对比

（1）REDS4 数据集对比结果

为了验证本文所提方法 MTCAN 的有效性，本文首先在测试集 REDS4 上将其与目前已有的视频压缩感知算法进行了对比，包括非深度学习方法 MH[9]、RRS[13]、以及深度学习方法 VCSNet[18]、ABPN[22]、FNLAN[26]、ImrNet[40]、STM-Net[21]、DUMHAN[20]、TSRN[19]、CollabNet[25]、SDPETs-Net[23]。具体测试结果如表 1 所示。其中，文献[9,13,18,19,22,25,26]的测试结果基于其官方或复现代码测试得出，其他对比算法的性能指标均直接引用自原文献。值得注意的是，本文方法 MTCAN 及 ABPN[22]、FNLAN[26]、TSRN[19]采用的是均匀的采样方法，即所有视频帧采样率相同；而其他对比方法采用 GOP 的采样策略，即每个 GOP 中包含 1 个高采样率关键帧和 7 个低采样率非关键帧。为保证平均每帧采样率一致，本文采用等效采样率进行比较：表中 SR=0.1 对应关键帧采样率 0.5、非关键帧采样率 0.1 的 GOP 采样方式，或等效平均采样率 0.15 的均匀采样方式（$0.1 \times 7 + 0.5 = 0.15 \times 8$）；表中 SR=0.01 对应关键帧采样率 0.5、非关键帧采样率 0.01 的 GOP 采样方式，或等效平均采样率 0.07 的均匀采样方式（$0.01 \times 7 + 0.5 \approx 0.07 \times 8$）。

当 SR=0.1 时，本文算法在平均 PSNR 和 SSIM 指标上分别领先次优算法 SDPETs-Net 方法 0.93dB 和 0.02。此结果充分说明了本文算法在重构质量上的优势。当 SR=0.01 时，本文算法相较于其他算法的性能优势更加明显，在平均 PSNR 和 SSIM 指标上分别领先次优算法 CollabNet 方法 1.68dB 和 0.05。这表明在低采样率条件下，MTCAN 通过充分利用时域相关性，能够更有效地补偿欠采样带来的信息缺失，从而获得更稳健的重构性能。

表 1　不同视频压缩感知重构算法在 REDS4 测试集的性能对比（PSNR(dB)/SSIM）

Table 1 Performance comparison of different video compressive sensing reconstruction algorithms on the REDS4 test set (PSNR(dB)/SSIM)

SR	算法	视频序列				平均值
		000	011	015	020	
0.01	MH	22.48/0.65	18.79/0.47	23.89/0.72	17.47/0.44	20.66/0.57
	RRS	20.72/0.58	21.64/0.56	25.34/0.73	21.45/0.59	22.29/0.62
	VCSNet	24.28/0.64	23.95/0.58	28.10/0.78	23.82/0.59	25.03/0.65
	ABPN	27.53/0.81	29.51/0.84	33.77/0.92	29.55/0.88	30.09/0.86
	FNLAN	27.17/0.79	29.14/0.83	33.11/0.92	29.39/0.88	29.70/0.85
	TSRN	27.10/0.79	28.94/0.82	32.93/0.91	29.20/0.88	29.54/0.85
	ImrNet	25.71/0.67	25.93/0.66	30.01/0.81	25.15/0.66	26.70/0.70
	STM-Net	26.45/0.73	26.89/0.71	30.67/0.84	25.98/0.71	27.50/0.75
	DUMHAN	27.74/0.77	26.72/0.70	31.02/0.85	25.97/0.70	27.86/0.76
	CollabNet	26.86/0.79	30.45/0.86	34.18/0.93	28.87/0.87	30.09/0.86
	SDPETs-Net	**29.44**/0.85	27.77/0.74	32.66/0.89	26.99/0.75	29.22/0.81
	本文 MTCAN	28.92/**0.87**	**32.10/0.90**	**34.17/0.94**	**31.91/0.93**	**31.77/0.91**
0.1	MH	24.29/0.72	20.33/0.53	25.15/0.75	19.03/0.49	22.20/0.62
	RRS	26.95/0.80	28.92/0.80	34.33/0.92	27.90/0.80	29.53/0.83
	VCSNet	28.51/0.82	29.56/0.82	33.71/0.91	29.89/0.85	30.42/0.85
	ABPN	29.83/0.89	32.71/0.91	36.97/0.96	33.16/0.94	33.17/0.92
	FNLAN	29.75/0.89	32.87/0.91	37.27/0.96	33.16/0.94	33.26/0.92
	TSRN	29.68/0.88	32.52/0.91	36.94/0.96	33.02/0.94	33.04/0.92
	ImrNet	29.09/0.85	32.29/0.89	36.33/0.94	31.23/0.90	32.24/0.90
	STM-Net	30.69/0.90	32.82/0.90	37.06/0.95	31.65/0.91	33.06/0.92
	DUMHAN	31.80/0.91	33.52/0.90	38.00/0.95	32.17/0.91	33.87/0.92
	CollabNet	29.56/0.89	34.10/0.93	37.99/0.96	32.55/ 0.94	33.55/0.93
	SDPETs-Net	**32.82/0.94**	34.36/0.92	**39.07**/0.96	33.16/0.93	34.85/0.94
	本文 MTCAN	32.46/**0.94**	**36.25/0.95**	38.46/**0.97**	**35.95/0.97**	**35.78/0.96**

（2）Vid4 数据集对比结果

为了进一步验证 MTCAN 的泛化性能，本文在 Vid4 测试集上进行对比测试，结果如表 2 所示。其中，

MCBCS[38]、RRS[13]、CSVNet[39]、IMFR-Net[41]的测试结果引用自文献[41]。CollabNet[25]测试结果由官方代码复现得到。所有方法的平均采样率等效为 0.0625 每帧。从表 2 中可以看出，在此数据集上，本文算法依然保持领先的重构性能，相比于次优算法 IMFR-Net，MTCAN 在 PSNR 和 SSIM 指标上分别领先 0.78dB 和 0.01。序列“Calendar”中，由于整体运动平缓，运动估计相对较为简单，因此基于显式运动估计的 IMFR-Net 方法表现更优。而在“City”和“Walk”这样的纹理和运动较为复杂的场景下，MTCAN 有着更加显著的性能优势。

表 2　不同视频压缩感知重构算法在 Vid4 测试集的性能对比（PSNR(dB)/SSIM）

Table 2 Performance comparison of different video compressive sensing reconstruction algorithms on the Vid4 test set (PSNR(dB)/SSIM)

算法	Calendar	City	Foliage	Walk	平均值
MCBCS	17.99/0.51	22.37/0.51	19.98/0.45	21.83/0.70	20.54/0.54
RRS	16.26/0.47	22.11/0.61	19.01/0.45	22.70/0.77	20.02/0.58
CSVNet	18.63/0.50	23.66/0.58	20.85/0.49	21.83/0.67	21.24/0.56
CollabNet	21.56/0.81	28.39/0.84	25.77/0.82	26.61/0.83	25.58/0.83
IMFR-Net	**24.84/0.88**	28.42/0.88	26.92/**0.87**	26.79/0.89	26.74/0.88
本文 MTCAN	23.89/0.87	**29.53/0.91**	**27.06**/0.86	**29.59/0.91**	**27.52/0.89**

（3）不同采样率下对比结果

为进一步分析本文方法在不同采样率条件下的重构性能变化，本文绘制了 MTCAN 在 REDS4 测试集上的 RD 曲线，结果如图 4 所示。当采样率由 0.01 提升至 0.15 时，PSNR 和 SSIM 增长较为明显。而在采样率继续增大后，性能提升幅度逐渐减小，表明模型在较高采样率下已能够恢复较充分的图像结构和纹理信息，进一步增加采样率带来的边际收益有所降低。

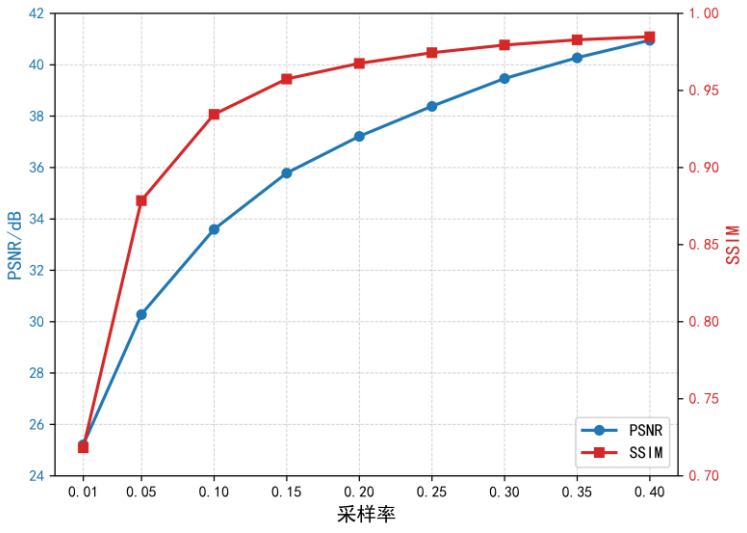

图 4　不同采样率下 MTCAN 重构性能对比

Fig.4 Comparison of MTCAN reconstruction performance under different sampling rates

2.2.2 视觉效果对比

为了更加直观的对比本文算法与其他视频压缩感知重构算法的重构性能差异，本小节通过图 5 和图 6 展示不同采样率下的各算法的视觉重构结果。

图 5 展示了当平均采样率为 0.07 时不同算法的重构结果。从中可以看出非深度学习方法 MH 和 RRS 重构效果较差，伴随着严重的块效应，图像内容中许多纹理结构都重构失败。深度学习方法中，VCSNet 由于结构简单，网络深度较浅，重构效果不佳。其余深度学习方法在整体画面质量上更高，但是本文所提 MTCAN 在细节上有着更优异的表现，例如图 5 中局部放大的车灯部分，MTCAN 的重构图中可以更清楚的看出车灯的结构。

图 6 则展示了平均采样率为 0.15 时，各个算法的重构表现。相比于低采样率，所有算法都有着明显的提升，但是非深度学习方法仍然存在部分图像块重构失败。对于重构质量较高的 ABPN、FNLAN、TSRN、CollabNet 和 MTCAN，在将局部细节放大之后，可以看到 MTCAN 的重构图中面部细节更加清晰、锐利，与原图最为接近。这表明所提方法在充分利用时域冗余信息后，能够更有效地恢复高频纹理与边缘结构等细节，从而获得更强的细节保持能力与结构还原能力。

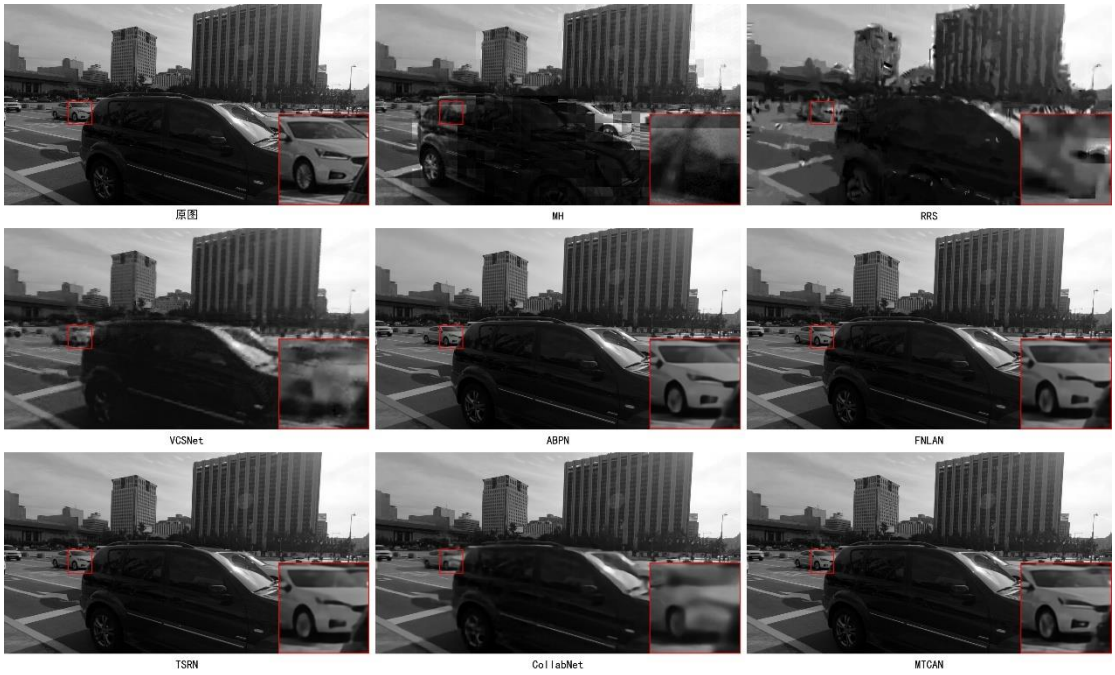

图 5　不同算法视觉重构效果对比 (SR=0.07)

Fig.5 Visual reconstruction comparison of different algorithms (SR=0.07)

图 6　不同算法视觉重构效果对比 (SR=0.15)

Fig.6 Visual reconstruction comparison of different algorithms (SR=0.15)

2.2.3 重构时间及网络参数量对比

表 3 展示了不同算法在不同分辨率下的平均每帧 FLOPs、参数量和单帧重构耗时。为了保证测试结果的可比性，除 CollabNet 外，其余算法均在同一台搭载 NVIDIA RTX 4090 显卡的服务器平台上完成推理测试，深度学习框架均为 PyTorch。由于 CollabNet 依赖的部分模块存在环境兼容性问题，本文将其单独部署在搭载 NVIDIA RTX 3090 显卡的服务器上进行测试。测试集为 REDS4，输入视频帧分辨率分别设置为 1280×720 和 512×512。FLOPs 和耗时均按平均每帧统计。采样率为平均 0.15 每帧。

对于 1280×720 分辨率，各方法的参数量整体处于同一数量级。在 FLOPs 方面，ABPN 的计算量最高，主要原因在于其采用多阶段双向传播结构，整体计算开销相对较大。FNLAN 和 TSRN 的 FLOPs 相对较低，其中 FNLAN 主要采用单向传播结构进行时域信息补偿，TSRN 则通过轻量化的时序平移操作实现帧间信息交互，因此计算量得到一定控制。CollabNet 的平均每帧 FLOPs 也较低，这是因为其采用关键帧与非关键帧结合的 GOP 重构方式，虽然关键帧重构计算负担较重，但占多数的非关键帧重构相对轻量。相比之下，本文所提 MTCAN 的平均每帧 FLOPs 仅为 1.26T，低于所有对比方法。其主要原因在于，MTCAN 采用了轻量化的线性注意力模块来替代传统注意力模块，同时，相比于其他方法，网络特征维度较低，残差增强模块的深度较浅。

从单帧重构耗时来看，各方法的整体趋势与 FLOPs 基本一致。MTCAN 的单帧耗时为 116 ms，同样低于其他对比方法。需要注意的是，ABPN、FNLAN、TSRN 和 MTCAN 均以 16 帧作为一个输入张量，并一次性输出 16 帧结果，因此推理效率较高。而 CollabNet 采用 GOP 重构模式，需要先重构关键帧，再利用关键帧结果依次重构非关键帧。这种串行重构方式使得此方法虽然平均 FLOPs 不高，但单帧耗时明显高于其他并行方法。此外，CollabNet 测试所用 GPU 性能也低于其他方法，进一步影响了其推理效率。

对于 512×512 分辨率，所有方法的网络参数量保持不变，FLOPs 和单帧耗时则随输入分辨率降低而明显下降。由于 512×512 图像的像素数量约为 1280×720 图像的 28.4%，各方法的 FLOPs 整体随空间分辨率近似比例下降，单帧推理时间也呈现相似下降趋势。

表 3　不同算法在不同分辨率下参数量和重构时间对比

Table 3 Comparison of parameter count and reconstruction time of different algorithms at different resolutions

分辨率	算法	FLOPs (T)	参数量（M）	单帧耗时（ms）
1280×720	ABPN	10.32	8.50 M	568
	FNLAN	2.66	4.61 M	203
	TSRN	2.64	4.07 M	210
	CollabNet	2.57	3.60 M	1574
	MTCAN	1.26	5.06 M	116
512×512	ABPN	3.00	8.50 M	150
	FNLAN	0.77	4.61 M	52
	TSRN	0.77	4.07 M	55
	CollabNet	0.72	3.60 M	439
	MTCAN	0.37	5.06 M	31

2.3 消融实验

2.3.1 关键模块消融实验

为了验证不同模块对于网络重构性能的贡献，本文设置了一系列的消融实验进行验证。不同模块组合及其重构结果如表 4 所示。实验在 REDS4 测试集上进行，并统一将输入裁剪为 512×512。采样率为 0.15。本文所提出的完整的 MTCAN 算法作为性能参照，记为组合 3。

首先，为了验证引入时域信息对重构质量的贡献，组合 0 移除了时域交叉注意力机制，仅在帧内做自注意力。与组合 3 相比，其重构质量出现了较大的退化，PSNR 和 SSIM 分别下降了 2.46dB 和 0.0255。这说明时域冗余是视频压缩感知重构的关键信息来源。在测量值信息不足的情况下，若缺少相邻帧的补充，网络难以重构出高质量结果。其次，为了验证线性注意力对于重构性能的影响，组合 1 将线性注意力替换为标准 softmax 注意力，其重构质量相较于采用了线性注意力的组合 3 提升了 0.1dB 和 0.0018，但单帧重构时间增加了 53ms，约提升至原来的 2 倍以上。此现象符合预期，线性注意力作为标准注意力的近似，本质上是牺牲了一定的重构质量换取更高的重构效率，并且显存要求也降低，更加适合分辨率较高的视频重构场景。最后，为了验证门控卷积前馈网络的有效性，组合 2 用标准的前馈网络替代了门控结构，其重构质量相较于组合 3 在 PSNR 上下降了 0.11dB，说明门控机制能够对特征更新进行自适应调制，抑制噪声信息的传递，对重构质量起到正向作用。并且，从重构时间角度来看，几乎没有增加额外的负担。

表 4　关键模块消融实验对比结果

Table 4 Comparison results of ablation experiments on key modules

组合	模块			重构质量（PSNR/SSIM）	重构时间（ms）
	时域交叉注意力	线性注意力	门控卷积前馈网络		
0	×	√	√	33.96/0.9382	31
1	√	×	√	36.52/0.9655	84
2	√	√	×	36.31/0.9632	31
3	√	√	√	36.42/0.9637	31

2.3.2 多尺度结构消融实验及可视化分析

为进一步分析时域交叉注意力中多尺度特征提取结构对重构性能的影响，本文设计了不同卷积核组合形式的多尺度结构消融实验，具体结果如表 5 所示。实验所采用的测试集为 REDS4（分辨率为 1280×720）。采样率为 0.15。实验中，将值分支的特征沿通道划分为数量相同的 4 组。通过改变 4 组卷积核的尺寸组合，考察不同尺度配置对网络性能的影响。作为性能参照，本文所提算法 MTCAN 采用的卷积核尺寸组合为{1, 3, 5, 7}。

首先，当四组卷积核均采用 1×1（即{1, 1, 1, 1}）时，该设置相当于注意力机制的默认实现，其重构效果相较于本文采用的{1, 3, 5, 7}组合在 PSNR 指标上下降 0.17dB。当保持单一卷积核尺寸并逐步增大感受野时（{3, 3, 3, 3}、{5, 5, 5, 5}、{7, 7, 7, 7}），PSNR 虽呈现一定波动并在{7,7,7,7}时达到 35.71dB，但仍低于多尺度组合{1,3,5,7}。值得注意的是，从参数量角度看，{7, 7, 7, 7}的卷积参数已明显高于{1, 3, 5, 7}，但性能仍未占优，说明仅依赖单一尺度难以充分适配视频中可能出现的不同尺度的纹理结构与运动模式。进一步地，本文还测试了由两种尺度构成的组合结构（{1, 1, 3, 3}、{3, 3, 5, 5}、{5, 5, 7, 7}），其 PSNR 在 35.56dB 和 35.70dB 之间，整体仍低于{1, 3, 5, 7}组合，说明仅依靠两种尺度依然不够。此外，本文还对更大尺度范围的组合{3, 5, 7, 9}进行了对比，其 PSNR 达到 35.72dB，仍然略低于{1, 3, 5, 7}组合的性能表现，这说明一味的增加尺度并不能带来更好的重构效果。综合来看，{1, 3, 5, 7}组合在保持较小参数开销的同时，能够更均衡地兼顾局部细节与大尺度运动信息，从而实现最优的重构精度。

表 5　不同多尺度结构消融实验对比结果

Table 5 Comparison results of ablation experiments on different multi-scale structures

多尺度结构				PSNR	SSIM
1	3	5	7	35.78	0.9574
1	1	1	1	35.61	0.9566
3	3	3	3	35.66	0.9564
5	5	5	5	35.59	0.9555
7	7	7	7	35.71	0.9567
1	1	3	3	35.56	0.9560
3	3	5	5	35.65	0.9568
5	5	7	7	35.70	0.9565
3	5	7	9	35.72	0.9568

此外，为了进一步说明值分支中多尺度卷积结构对特征表达的影响，本文对时域交叉注意力模块中不同尺度值分支的响应结果进行了可视化分析，结果如图 7 所示。从图中可以看出，不同尺度卷积核所学习到的特征响应呈现出明显差异，并且这种差异不仅体现在静态纹理结构上，也与视频中不同范围的运动变化密切相关。1×1 分支主要关注局部运动细节和细粒度纹理变化，例如图中砖块边缘、墙面纹理附近均出现较为集中的响应，说明小尺度卷积更倾向于捕捉局部位置上的细微变化。随着卷积核尺寸增大，3×3 和

5×5 分支的响应范围进一步扩大，不仅能够覆盖局部边缘，还能够在邻域范围内形成更连续的结构响应，从而补充局部运动周围的上下文信息。相比之下，7×7 分支表现出更强的大范围结构响应。由于该序列中背景区域存在整体位移，画面左侧和上方边缘在 7×7 分支中形成了较为连续的条带状响应，说明了大尺度卷积对较大范围背景结构变化的感知能力。上述结果表明，不同尺度的值分支并非产生重复响应，而是对局部运动细节、邻域结构变化和大范围背景位移形成互补表征。

图 7 不同尺度值分支的特征响应可视化结果(SR=0.15，“011”序列中第 11 帧)

Fig.7 Visualization results of feature responses from value branches at different scales (SR=0.15, the 11th frame of sequence “011”)

2.3.3 不同序列长度下门控模块稳定性分析

为进一步验证门控卷积前馈网络的作用，本文在不同测试序列长度下，对有无门控模块的模型进行了更全面的对比实验，结果如表 6 所示。其中，“√”表示采用门控卷积前馈网络，“×”表示将其替换为普通前馈网络。测试集为 REDS4，分辨率为 512×512，采样率为 0.15。

从表 6 可以看出，在所有测试序列中，引入门控模块后模型的 PSNR 均得到提升。并且，随着测试序列长度由 8 帧增加至 64 帧，门控模块带来的平均 PSNR 增益整体呈上升趋势，由 0.0848dB 提升至 0.1361dB。这表明，当序列变长时，时域传播过程中误差累积的影响更加明显，而门控模块能够在较长时间范围内保持更稳定的特征更新。值得注意的是，序列 011 和 020 在长序列条件下的提升更加明显。这主要是因为 011 和 020 序列中运动变化更为复杂，时域相关性建模过程中更容易引入不可靠信息，因此更需要门控模块对传播特征进行自适应调制。

表 6 不同序列长度下有无门控模块 PSNR 对比（dB）

Table 6 PSNR comparison with/without the gated module under different sequence lengths (dB)

视频序列长度		8	16	24	32	40	48	56	64
000	×	33.7028	34.4314	34.4889	34.3757	34.2817	34.2096	34.1622	34.1263
	√	33.7603	34.5237	34.5855	34.4736	34.3770	34.3048	34.2491	34.2169
	差值	**0.0575**	**0.0923**	**0.0966**	**0.0979**	**0.0953**	**0.0952**	**0.0869**	**0.0906**
011	×	37.0286	37.6001	37.9398	38.0387	38.0333	37.9418	37.9417	37.8615
	√	37.1369	37.7294	38.0566	38.1589	38.1680	38.098	38.0964	38.0274
	差值	**0.1083**	**0.1293**	**0.1168**	**0.1202**	**0.1347**	**0.1562**	**0.1547**	**0.1659**
015	×	37.9596	37.7584	37.6825	37.6555	37.6876	37.6997	37.6654	37.8357
	√	38.0170	37.8579	37.8014	37.7810	37.8015	37.8194	37.7843	37.9588
	差值	**0.0574**	**0.0995**	**0.1189**	**0.1255**	**0.1139**	**0.1197**	**0.1189**	**0.1231**
020	×	34.2637	35.4506	35.6592	35.4424	35.3319	35.3864	35.5073	35.6556
	√	34.3796	35.557	35.7721	35.5796	35.4809	35.5356	35.6659	35.8205
	差值	**0.1159**	**0.1064**	**0.1129**	**0.1372**	**0.1490**	**0.1492**	**0.1586**	**0.1649**
平均差值		**0.0848**	**0.1069**	**0.1113**	**0.1202**	**0.1232**	**0.1301**	**0.1298**	**0.1361**

3 结束语

本文提出了一种基于多尺度时域交叉注意力的视频压缩感知重构网络（MTCAN）。首先，为了充分利用视频序列的时域相关性，构建了时域交叉注意力机制，在当前帧与相邻帧之间建立依赖，实现了时域相关信息的动态检索与融合。其次，针对不同场景下运动幅度与作用范围存在的差异性，引入并行多尺度卷积表征结构，增强不同感受野下的局部特征表达能力。再次，针对标准注意力计算复杂度随序列长度呈二次增长的问题，采用线性注意力降低计算与显存开销，并结合门控卷积前馈模块抑制误差传播，缓解递归更新过程中不可靠信息的累积影响。实验结果表明，与现有方法相比，MTCAN 在保持参数量和重构时间相近的同时，能够取得最优的重构精度，并且在高频细节上优势明显。消融实验也充分证明了所提出的各个关键模块的有效性。

参考文献

[1] BARANIUK R G. Compressive sensing [J]. IEEE Signal Processing Magazine, 2007, 24(4): 118–121.

[2] DONOHO D L. Compressed sensing [J]. IEEE Transactions on Information Theory, 2006, 52(4): 1289–1306.

[3] PUDLEWSKI S, PRASANNA A, MELODIA T. Compressed-sensing-enabled video streaming for wireless multimedia sensor networks [J]. IEEE Transactions on Mobile Computing, 2012, 11(6): 1060–1072.

[4] 张健,刘鹏博,汤健.联合 WPT 和 MEC 的无线传感网时延优化算法[J].数据采集与处理,2025,40(1):163-175. ZHANG Jian, LIU Pengbo, TANG Jian. Delay Optimization Algorithm in Wireless Sensor Networks Combining WPT and MEC[J]. Journal of Data Acquisition and Processing, 2025, 40(1): 163-175.

[5] 薛磊,段继忠.基于多尺度特征融合预处理与深度稀疏网络的并行磁共振成像重建[J].数据采集与处理,2025,40(4):1082-1095.

XUE Lei, DUAN Jizhong. Parallel Magnetic Resonance Imaging Reconstruction Based on Multi-scale Feature Fusion Preprocessing and Deep Sparse Networks[J]. Journal of Data Acquisition and Processing, 2025, 40(4): 1082-1095.
[6] BARANIUK R G, GOLDSTEIN T, SANKARANARAYANAN A C, et al. Compressive video sensing: Algorithms, architectures, and applications [J]. IEEE Signal Processing Magazine, 2017, 34(1): 52–66.
[7] 郑学炜, 杨春玲, 禤韵怡. CVS 中基于视频运动特征的多假设-双稀疏重构算法[J]. 电子学报, 2020, 48(02): 249–257.
ZHENG Xuewei, YANG Chunling, XUAN Yunyi. Video motion features based multi-hypothesis-dual-sparsity reconstruction algorithm in compressed video sensing [J]. Acta Electronica Sinica, 2020, 48(2): 249–257.
[8] 欧伟枫, 杨春玲, 戴超. 一种视频压缩感知中两级多假设重构及实现方法[J]. 电子与信息学报, 2017, 39(07): 1688–1696.
OU Weifeng, YANG Chunling, DAI Chao. A two-stage multi-hypothesis reconstruction and implementation method in video compressive sensing [J]. Journal of Electronics & Information Technology, 2017, 39(7): 1688–1696.
[9] TRAMEL E W, FOWLER J E. Video compressed sensing with multihypothesis [C]//Proceedings of the Data Compression Conference. Snowbird: IEEE Press, 2011: 193–202.
[10] ZHENG S, CHEN J, ZHANG X P, et al. A new multihypothesis-based compressed video sensing reconstruction system [J]. IEEE Transactions on Multimedia, 2021, 23: 3577–3589.
[11] 常侃, 覃团发, 唐振华. 基于多重假设的视频压缩感知分层重建[J]. 数据采集与处理, 2013, 28(6): 730-738.
CHANG Kan, QIN Tuanfa, TANG Zhenhua. Multi-hypothesis based hierarchical reconstruction for compressed video sensing[J]. Journal of Data Acquisition and Processing, 2013, 28(6): 730-738.
[12] 汤瑞东, 杨春玲, 禤韵怡. 视频压缩感知多假设局部增强重构算法[J]. 自动化学报, 2022, 48(08): 1984–1993.
TANG Ruidong, YANG Chunling, XUAN Yunyi. Local enhancement reconstruction algorithm based on multihypothesis prediction in compressed video sensing [J]. Acta Automatica Sinica, 2022, 48(8): 1984–1993.
[13] CHEN C, ZHOU C, LIU P, et al. Iterative reweighted Tikhonov-regularized multihypothesis prediction scheme for distributed compressive video sensing [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(1): 1–10.
[14] 和志杰, 杨春玲, 汤瑞东. 视频压缩感知中基于结构相似的帧间组稀疏表示重构算法研究[J]. 电子学报, 2018, 46(03): 544–553.
HE Zhijie, YANG Chunling, TANG Ruidong. Research on reconstruction algorithm of inter-frame group sparse representation based on structural similarity in video compressive sensing [J]. Acta Electronica Sinica, 2018, 46(3): 544–553.
[15] ZHAO C, MA S, ZHANG J, et al. Video compressive sensing reconstruction via reweighted residual sparsity [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(6): 1182–1195.
[16] LLULL P, LIAO X, YUAN X, et al. Coded aperture compressive temporal imaging[J]. Optics Express, 2013, 21(9): 10526-10545.
[17] WAKIN M B, LASKA J N, DUARTE M F, et al. Compressive imaging for video representation and coding[C]//Proceedings of Picture Coding Symposium. Beijing, China: Picture Coding Symposium, 2006.
https://www.researchgate.net/publication/220043723_Compressive_Imaging_for_Video_Representation_and_Coding
[18] SHI W, LIU S, JIANG F, et al. Video compressed sensing using a convolutional neural network [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(2): 425–438.

[19] GU Z F, ZHOU C, LIN G F. A temporal shift reconstruction network for compressive video sensing [J]. IET Computer Vision, 2024, 18(4): 448-457.
[20] YANG X, YANG C. MAP-inspired deep unfolding network for distributed compressive video sensing [J]. IEEE Signal Processing Letters, 2023, 30: 309–313.
[21] WEI Z, YANG C, XUAN Y. Efficient video compressed sensing reconstruction via exploiting spatial-temporal correlation with measurement constraint [C]// Proceedings of the IEEE International Conference on Multimedia and Expo. Shenzhen: IEEE Press, 2021: 1–6.
[22] ZHOU C, CHEN C, ZHANG D. Deep video compressive sensing with attention-aware bidirectional propagation network [C]// Proceedings of the 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. Beijing: IEEE Press, 2022: 1–5.
[23] 杨春玲, 梁梓文. 静态与动态域先验增强的两阶段视频压缩感知重构网络[J]. 电子与信息学报, 2024, 46(11): 4247–4258.
YANG Chunling, LIANG Ziwen. Static and dynamic-domain prior enhancement two-stage video compressed sensing reconstruction network [J]. Journal of Electronics & Information Technology, 2024, 46(11): 4247–4258.
[24] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks [C]// Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE Press, 2017: 764–773.
[25] LIU J H, YANG C L. A dual-domain collaboration network for VCS reconstruction [C]//Proceedings of the IEEE International Conference on Image Processing. Abu Dhabi: IEEE Press, 2024: 1732-1738.
[26] ZHOU C, CHEN C, ZHANG D. A flow-guided non-local alignment network for video compressive sensing reconstruction [C]// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes: IEEE Press, 2023: 1–5.
[27] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE Press, 2018: 7794–7803.
[28] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017: 5998-6008.
[29] KATHAROPOULOS A, VYAS A, PAPPAS N, et al. Transformers are RNNs: Fast autoregressive transformers with linear attention[C]//Proceedings of the 37th International Conference on Machine Learning. Online: PMLR, 2020: 5156-5165.
[30] FOWLER J E, MUN S, TRAMEL E W. Block-based compressed sensing of images and video [J]. Foundations and Trends in Signal Processing, 2012, 4(4): 297–416.
[31] FAN Q, HUANG H, AI Y, et al. Rectifying magnitude neglect in linear attention[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE Press, 2025: 21505-21514.
[32] CAI H, LI J, HU M, et al. EfficientViT: Lightweight multi-scale attention for high-resolution dense prediction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE Press, 2023: 17302-17313.
[33] SHAZEER N. GLU Variants Improve Transformer[EB/OL]. https://arxiv.org/abs/2002.05202, 2020-02-12.
[34] HENDRYCKS D, GIMPEL K. Gaussian Error Linear Units (GELUs)[EB/OL]. https://arxiv.org/abs/1606.08415, 2016-06-27.
[35] KINGMA D P, BA J. Adam: A method for stochastic optimization [C]// Proceedings of International Conference on Learning Representations. San Diego: ICLR, 2015: 1-15.
[36] CHARBONNIER P, BLANC-FERAUD L, AUBERT G, et al. Two deterministic half-quadratic regularization algorithms for computed imaging [C]// Proceedings of the IEEE International Conference on Image Processing. Austin: IEEE Press, 1994: 168–172.

[37] SHI S, GU J, XIE L, et al. Rethinking alignment in video super-resolution transformers [C]// Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022: 36081–36093.
[38] MUN S, FOWLER J E. Residual reconstruction for block-based compressed sensing of video [C]// Proceedings of the Data Compression Conference. Snowbird: IEEE Press, 2011: 183–192.
[39] XU K, REN F. CSVideoNet: A real-time end-to-end learning framework for high-frame-rate video compressive sensing [C]// Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe: IEEE Press, 2018: 1680–1688.
[40] YANG X, YANG C. IMRNet: An iterative motion compensation and residual reconstruction network for video compressed sensing [C]// Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE Press, 2021: 2350–2354.
[41] ZHANG Y, ZHANG W X, ZHAO Z F, et al. IMFR-Net: Interval measurement and full recovery network for video compressive sensing [J]. Neurocomputing, 2025, 612: 130803.

作者简介：

周超（1994-），男，讲师，硕士生导师，研究方向：图像处理和计算机视觉。E-mail: zhouchao@njupt.edu.cn

陈灿（1993-），**通信作者**，男，讲师，硕士生导师，研究方向：图像处理和计算机视觉。E-mail: chencan@njupt.edu.cn

张登银（1964-），男，研究员，博士生导师，研究方向：信号与信息处理和信息安全。

A Multi-Scale Temporal Cross-Attention Network for Video Compressive Sensing Reconstruction

ZHOU Chao, CHEN Can, ZHANG Dengyin

(School of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Video compressive sensing (VCS) provides an efficient acquisition paradigm by sampling video signals at rates far below the Nyquist requirement, but severe undersampling also makes video reconstruction a highly ill-posed inverse problem. To tackle this ill-posed reconstruction problem, this paper proposes a multi-scale temporal cross-attention network for VCS reconstruction, which improves reconstruction quality by exploiting the spatiotemporal redundancy within video sequences. The proposed method adopts an end-to-end reconstruction framework composed of a learnable compressive sampling module and a temporal feature propagation network. At the sampling stage, a learnable block-based sensing module is used to replace the fixed random measurement matrix, so that the sampling operator can be jointly optimized with the reconstruction network. At the reconstruction stage, the measurements are first mapped to an initial estimate of the video frames. Based on this initial reconstruction, bidirectional temporal propagation is further employed to introduce information from neighboring frames and refine the current frame representation. Within each propagation direction, a temporal cross-attention mechanism is designed to model the correlation between the current frame and its adjacent frame. Different from conventional self-attention, the query is generated from the current frame, while the key and value are constructed from both the current and neighboring frames. This design enables the network to retrieve temporal information from adjacent frames and integrate it into the reconstruction of the current frame. Considering that motion patterns vary significantly across different video scenes, multi-scale content representation is introduced into the value branch of the cross-attention module. This design enhances the ability of the network to represent local details and broader motion-related structures without disturbing the correlation calculation in the query and key branches. To reduce the computational burden caused by long video sequences and high-resolution features, linear attention is adopted instead of standard attention. Furthermore, a gated convolutional feed-forward module is introduced after temporal feature fusion. By combining local convolutional modeling with adaptive gating, this module selectively enhances reliable propagated features and suppresses unreliable updates, which helps alleviate error accumulation during recursive reconstruction.Experimental results under multiple sampling rates demonstrate the effectiveness of the proposed method. Compared with existing reconstruction approaches, the proposed network achieves improvements of 2.55 dB in PSNR and 0.10 in SSIM at the sampling rate of 0.01, and 0.93 dB in PSNR and 0.02 in SSIM at the sampling rate of 0.10. These results indicate that the proposed framework can make better use of inter-frame information, especially under severe undersampling conditions. Overall, the proposed method provides an effective reconstruction solution for video compressive sensing and shows strong potential for low-power and resource-limited video acquisition applications.

Highlights:

1. A multi-scale temporal cross-attention network is proposed for video compressive sensing reconstruction.
2. Multi-scale convolutions are embedded into the value branch to enhance temporal feature fusion under different motion ranges.
3. Linear attention is adopted to reduce the computational cost of long-sequence video reconstruction.
4. A gated convolutional feed-forward module is introduced to suppress unreliable propagated features and alleviate error accumulation.

5. Experimental results show that the proposed method achieves superior reconstruction performance under multiple sampling rates.

Keywords: Compressive sensing, Video reconstruction, Attention mechanism, Feature alignment, Temporal correlation

Foundation items: Basic Research Program of Jiangsu (BK20250682) and National Natural Science Foundation of China (62471241)

*__Corresponding author, E-mail:__ chencan@njupt.edu.cn