铁路计算机应用
RAILWAY COMPUTER APPLICATION

基于光流引导Transformer模型的重载铁路监控压缩视频质量增强方法

王文斌，宋宗莹，柴雪松，凌烈鹏，李健超

Compressed video quality enhancement method for heavy-haul railway surveillance based on optical flow guided Transformer model

WANG Wenbin, SONG Zongying, CHAI Xuesong, LING Liepeng, and LI Jianchao

引用本文:
王文斌, 宋宗莹, 柴雪松, 等. 基于光流引导Transformer模型的重载铁路监控压缩视频质量增强方法[J]. 铁路计算机应用, 2025, 34(1): 27–33.
WANG Wenbin, SONG Zongying, CHAI Xuesong, et al. Compressed video quality enhancement method for heavy-haul railway surveillance based on optical flow guided Transformer model[J]. Railway Computer Application, 2025, 34(1): 27-33.

在线阅读 View online: http://tljsjyy.xml-journal.net/2025/I1/27

您可能感兴趣的其他文章
Articles you may be interested in

新一代信息技术驱动下的智能重载铁路总体架构研究
Overall framework of intelligent heavy haul railway driven by new generation of information technology
铁路计算机应用. 2020, 29(6): 25–29

基于数据中台的重载铁路智能化应用研究
Intelligent application of heavy haul railway based on data middle platform
铁路计算机应用. 2023, 32(12): 67–72

重载铁路移动闭塞系统架构研究
Research on framework of moving block system for heavy haul railway
铁路计算机应用. 2021, 30(1): 67–71

重载铁路供电智能运维移动应用的设计与实现
Intelligent operation and maintenance mobile application for heavy haul railway power supply
铁路计算机应用. 2024, 33(4): 59–64

重载铁路港口车站智能管控平台设计
Intelligent control platform for heavy haul railway port stations
铁路计算机应用. 2023, 32(4): 79–83

重载铁路人车地全联锁防护系统的设计与实现
Full interlocking protection system for human–vehicle–ground in heavy–haul railway
铁路计算机应用. 2024, 33(6): 79–83

关注微信公众号，获得更多资讯信息

文章编号：1005-8451（2025）01-0027-07

基于光流引导 Transformer 模型的重载铁路监控压缩视频质量增强方法

王文斌[1]，宋宗莹[1]，柴雪松[2,3]，凌烈鹏[2,3]，李健超[2,4]
（1. 中国神华能源股份有限公司，北京 100080；
2. 中国铁道科学研究院集团有限公司 铁道建筑研究所，北京 100081；
3. 中铁科学技术开发有限公司，北京 100081；
4. 高速铁路轨道系统全国重点实验室，北京 100081）

摘 要：重载铁路视频监控系统的不断扩增，使得铁路视频数据急剧增长，对数据存储和传输等能力的要求更高。为此，提出了一种基于光流引导 Transformer 模型的重载铁路监控压缩视频质量增强方法。通过光流补全网络提取帧间运动信息，指导 Transformer 模型关注视频序列中的重要特征；结合多头自注意力机制和时间空间特征融合策略，有效提取视频帧的时空特征；通过在 Transformer 模型结构中融入光流引导的特征增强模块，进一步提升视频质量增强的准确性和效率。基于实际采集的重载铁路监控视频数据集的实验结果表明，该方法显著优于现有的视频质量增强方法，具有实用价值。

关键词：重载铁路；视频增强；光流；Transformer 模型；多头自注意力机制
中图分类号：U239.4 : TP391 **文献标识码：**A **DOI：**10.3969/j.issn.1005-8451.2025.01.04

Compressed video quality enhancement method for heavy-haul railway surveillance based on optical flow guided Transformer model

WANG Wenbin[1], SONG Zongying[1], CHAI Xuesong[2,3], LING Liepeng[2,3], LI Jianchao[2,4]
(1. China Shenhua Energy Company Limited, Beijing 100080, China; 2. Railway Engineering Research Institute, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China; 3. Zhongtie Science & Technology Development Co. Ltd., Beijing 100081, China; 4. State Key Laboratory of High-speed Railway Track System, Beijing 100081, China)

Abstract: The continuous expansion of heavy-haul railway video surveillance systems has led to a sharp increase in railway video data, with higher requirements for data storage and transmission capabilities. To this end, this paper proposed a compressed video quality enhancement method for heavy-haul railway surveillance based on optical flow guided Transformer model: Extracting inter frame motion information through optical flow completion network to guide the Transformer model to focus on important features in the video sequence; Combining multi head self-attention mechanism and spatiotemporal feature fusion strategy, effectively extracting spatiotemporal features of video frames; By incorporating optical flow guided feature enhancement modules into the Transformer model structure, further improving the accuracy and efficiency of video quality enhancement. The experimental results based on the actual collected heavy-duty railway surveillance video dataset show that this method is significantly superior to existing video quality enhancement methods and has practical value.

Keywords: heavy haul railway; video enhancement; optical flow; Transformer model; multi-head self-attention mechanism

我国部分重载铁路为典型山区铁路，发生自然灾害和地质灾害的风险较大。近年来，我国铁路技术创新和安全保障能力不断提高，相关单位针对重载铁路布设了视频监控系统，视频采集点主要覆盖隧道、作业区域、边坡、平交道口、供电设施、行车室、通信机房等重点区域，具有点多、覆盖区域广、数据量大的特点。视频数据的急剧增长对数据

收稿日期：2024-07-02
基金项目：中国神华科技创新基金项目（SHGF-22-4）
作者简介：王文斌，正高级工程师；宋宗莹，正高级工程师。

存储和传输等能力提出了更高的要求。为应对上述问题，视频压缩技术应运而生，该技术可有效减少数据的存储空间和传输时间，但常以牺牲视频质量为代价，导致画面模糊和细节丢失，可能会影响识别、检测、跟踪等视觉任务的性能。因此，亟需研究如何提升经过压缩后的视频的质量。

传统的视频质量增强方法主要通过同质性或光流来探索视频的几何关系，从而实现视频内容在时间和空间上的有效增强[1-6]。例如，Huang 等人[7] 利用自然视频的固有特性，设计了一套能量方程，通过迭代优化运动场和帧质量，实现了连贯的视频画质提升。由于 Transformer 模型（简称：Transformer）[8] 在长时间空间建模方面的显著能力，已被广泛用于视频质量增强领域[9-11]， Zeng 等人[9]、Liu 等人[10] 和 Li 等人[11] 通过调整 Transformer，在较大的时间感受野内检索相似特征，实现了高质量的视频质量增强。在大多数基于 Transformer 的视频质量增强研究中，多个视频帧被单独编码为词元（token），通过各种 Transformer 块来估计与失真区域相关的词元与有效区域之间的特征相关性，再利用这些相关性对特征进行聚合，以恢复损坏区域的特征，其在性能方面大幅度超越传统视频质量增强方法。然而，采用 Transformer 进行视频质量增强仍存在运动信息的精确捕捉和查询退化等挑战。

综上，本文提出基于光流引导 Transformer 的重载铁路压缩视频质量增强（OFGT-CVQE，Optical Flow Guided Transformer for Compressed Video Quality Enhancement）方法，提升重载铁路监控系统中有损压缩后的视频序列的质量。

1 视频增强方法

本文提出的 OFGT-CVQE 方法，假设视频长度为 T；视频质量增强的输入为经有损压缩后的视频序列 $X=\{X_1,\cdots,X_T\}$，其对应的掩码序列为 $M=\{M_1,\cdots,M_T\}$，在每个掩码 M_t 中（$t\in[1,T]$），“1”标识失真区域，“0”标识有效区域。其目标是重构视频序列中的失真区域，并确保重构结果 $\widehat{Y}=\{\widehat{Y}_1,\cdots,\widehat{Y}_T\}$ 和压缩前无失真视频序列 $Y=\{Y_1,\cdots,Y_T\}$ 在时空上保持一致。

1.1 方法流程

OFGT-CVQE 方法流程如图 1 所示，包含光流补全与特征提取模块、光流引导 Transformer 模块及质量增强模块。光流补全与特征提取模块先提取帧间运动信息，利用此信息指导 Transformer 关注视频序列中的重要特征，从而提高对动态内容的敏感度；光流引导 Transformer 模块结合多头自注意力机制和时间空间特征融合策略，有效提取视频帧的时空融合特征，获得用于视频质量增强的语义特征表示；质量增强模块利用融合特征映射、结合全卷积增强网络来计算残差映射，残差映射和压缩失真视频叠

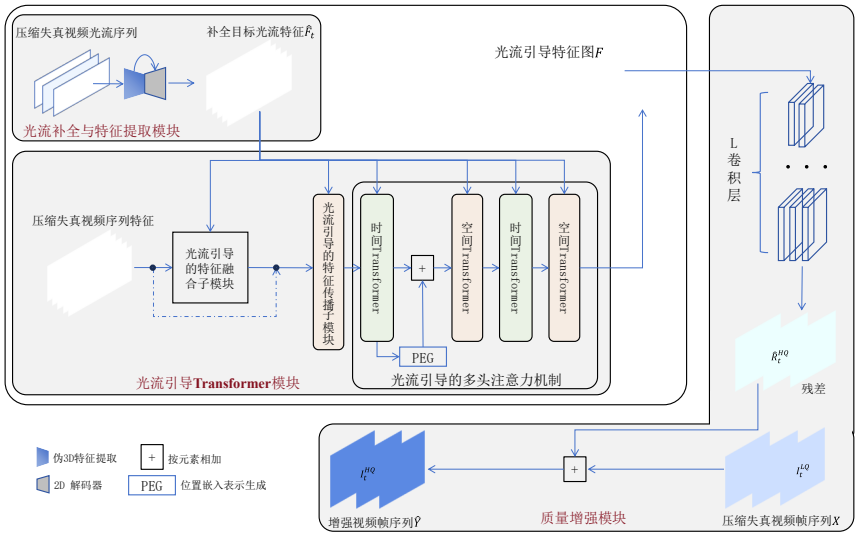

图1 OFGT-CVQE 方法流程示意

加生成增强视频。

1.2 光流补全与特征提取模块

由于视频一般具有较长序列，OFGT-CVQE 方法采用滑动窗口策略增强视频帧，在每次向前传递中，以第t帧为中心，对其局部帧$X_l=\{X_{t-s},\cdots,X_{t-1},X_t,X_{t+1},\cdots,X_{t+s}\}$和全局帧$X_g=\{X_r,X_{2r},\cdots\}$进行采样。其中，$s$为局部帧的采样步幅；$r$为全局帧的采样间隔；全局帧序列$X_g$用于扩大时间感受域。

针对有损压缩视频序列直接根据相邻帧估计光流并提取特征，会因存在失真区域导致光流信息不准确的问题，本方法在提取第 t 帧的光流信息时，先使用拉普拉斯填充光流法获得以第 t 帧为中心的初始的光流估计序列$\tilde{F}=\left\{\tilde{F}_{t-ni},\cdots,\tilde{F}_t,\cdots,\tilde{F}_{t+ni}\right\}$，$i$是连续光流之间的时间间隔，光流序列的长度为$2n+1$；利用伪 3D 卷积提取$\tilde{F}$的特征作为第 t 帧光流的特征表示，即利用第 t 帧临近帧的光流信息对第 t 帧的光流信息进行补全。设第m个伪 3D 卷积块的输入为$\tilde{f}^m$、输出为$\tilde{f}^{m+1}$。第 t 帧光流的特征聚合过程为

$$\tilde{f}^{m+1}=TC\left(\overline{SC}\left(\tilde{f}^m\right)\right)+\tilde{f}^m \tag{1}$$

式（1）中，TC表示一维时间卷积；$\overline{SC}$表示二维空间卷积。在拉普拉斯填充光流法的编码器中，除最后一个伪 3D 块和跳跃连接外，时间分辨率保持不变；在伪 3D 块和跳跃连接内，通过减小时间分辨率来获得目标帧光流的聚合特征。最终，通过二维解码器输出临近帧信息补全后的目标光流特征$\widehat{F}_t$。

1.3 光流引导 Transformer 模块

光流引导 Transformer 模块提取压缩失真视频序列的上下文语义信息，输出用于视频质量增强的特征图F。已有的研究表明直接利用 Transformer 建模TI会出现查询退化问题。不同于现有基于 Transformer 的方法仅对原始 RGB 序列进行语义特征提取，本文提出的光流引导 Transformer 模块的输入由 2 个部分组成：（1）前述局部帧和全局帧的组合$X_{in}=X_l\cup X_g$；（2）光流补全与特征提取模块得到的第t帧的光流特征$\widehat{F}_t$。可充分利用光流特征中蕴含的运动细节信息引导 Transformer 更关注有利于质量增强的运动细节。对于X_{in}，将其每一帧先通过预训练的卷积神经网络（在 ImageNet 数据集上预训练的 ResNet-50）转换为特征序列TI，以供 Transformer 建模处理。

光流引导 Transformer 模块包含 3 个部分：光流引导的特征融合（OFGFI，Optical Flow Guided Feature Integration）子模块、光流引导的特征传播（OFGFP，Optical Flow Guided Feature Propagation）子模块和光流引导的多头注意力机制（OFG-MHSA，Optical Flow Guided Multi-head Self-attention）子模块。

1.3.1 OFGFI 子模块

该子模块旨在将光流特征$\widehat{F}_t$与原始 RGB 帧特征TI进行融合。先将光流特征$\widehat{F}_t$编码为特征序列TF，然后根据TF和TI之间的加权交互来控制TF对TI的影响，公式为

$$\widehat{TF}_t=TF_t\oplus\mathrm{MLP}(C(TI_t,TF_t)) \tag{2}$$

式（2）中，C为拼接操作；$\oplus$为按位相加；MLP 为由多层感知机组成的全连接层；TI_t、TF_t和$\widehat{TF}_t$分别表示第t帧 RGB 模态的特征、原始光流特征和加权之后的光流特征。最后，将$\widehat{TF}_t$和TI_t连接起来形成光流增强的视频流第t帧特征表示为

$$\tilde{T}I_t=C\left(\widehat{TF}_t,TI_t\right) \tag{3}$$

视频序列所有帧经光流增强后的特征表示，记为特征空间序列$\tilde{T}I$。

1.3.2 OFGFP 子模块

该子模块基于光流信息来帮助跨帧传递，从而改善视频序列特征中的时间连续性和空间一致性。设从t到$t+1$帧的补全光流特征记为$\widehat{F}_{t\to t+1}$，从t到t-1帧的补全光流特征记为$\widehat{F}_{t\to t-1}$，光流引导的特征传播可表示为

$$\widehat{FI}_t^f=\mathrm{OFGFP}\left(TI_t,TI_{t-1},\widehat{F}_{t\to t-1}\right) \tag{4}$$

$$\widehat{FI}_t^b=\mathrm{OFGFP}(TI_t,TI_{t+1},\widehat{F}_{t\to t+1}) \tag{5}$$

$$\widehat{FI}t=\widehat{FI}_t^f\oplus\widehat{FI}_t^b \tag{6}$$

式（4）~式（6）中，$\widehat{FI}_t^f$和$\widehat{FI}_t^b$分别表示前向和后向传播特征；$\widehat{FI}_t$为特征传播融合后的特征；OFGFP 结构如图 2 所示，在特征按通道连接后，采用可变形卷积并经过特征相加后对传播后的特征进行聚合。

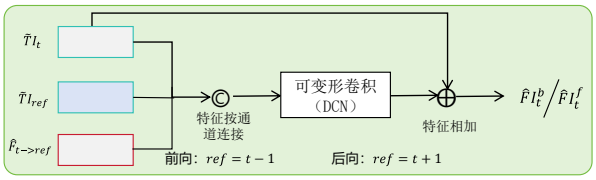

图2 OFGFP 子模块结构示意

1.3.3 OFG-MHSA 子模块

OFGFP 子模块的输出并未作为常规时空 Transformer 结构的输入进行语义上下文建模。OFGT-CVQE 方法在时间和空间维度上对时空 Transformer 进行了解耦，即 Transformer 由时间 Transformer 单元和空间 Transformer 单元组成。时间 Transformer 用于挖掘视频帧间的时序语义，空间 Transformer 用于挖掘视频帧内的空间语义。时间 Transformer 单元和空间 Transformer 单元与现有结构最大的不同在于，使用 OFG-MHSA 用于计算查询、键和值向量，即通过补全的光流特征$\widehat{F}_{t\to t-1}$和$\widehat{F}_{t\to t+1}$来建模第t帧光流特征TI_t与第$t-1$和$t+1$帧光流特征之间的相关性。光流引导的多头注意力机制中的查询、键和值向量编码流程如下。

（1）使用软合成（SC）将TI_{t-1}和TI_{t+1}的嵌入表示转换到特征空间并获得相应的特征，软合成是指使用 Softmax 计算注意力权重对输入特征进行加权平均，达到“软”组合的效果。公式为

$$\widehat{FI}_{t-1} = SC(TI_{t-1}) \tag{7}$$

$$\widehat{FI}_{t+1} = SC(TI_{t+1}) \tag{8}$$

（2）在得到$\widehat{FI}_{t-1}$和$\widehat{FI}_{t+1}$后，利用相应补全光流$\widehat{F}_{t\to t-1}$和$\widehat{F}_{t\to t+1}$来聚合沿时间维度扭曲位置的相关内容，并使用软分裂（SS）将特征映射到词元嵌入表示空间（token representation），软分裂是指对特征进行指数归一化，以进行注意力操作。

$$TI_{t-1\to t} = SS(\mathcal{W}(\widehat{FI}_{t-1}, \widehat{F}_{t\to t-1})) \tag{9}$$

$$TI_{t+1\to t} = SS(\mathcal{W}(\widehat{FI}_{t+1}, \widehat{F}_{t\to t+1})) \tag{10}$$

式（9）~式（10）中，$\mathcal{W}$表示沿时间维度反向扭曲操作；$TI_{t-1\to t}$和$TI_{t+1\to t}$分别表示从$t-1$到t和$t+1$到t的扭曲的嵌入表示。

（3）根据扭曲的嵌入表示经一层全连接网络对查询、键和值向量进行编码。

经过时间 Transformer 单元和空间 Transformer 单元后得到最终的光流引导特征图F。为使整个光流引导 Transformer 模块能够准确建模视频帧的前后位置关系，在特征经过第一个时间 Transformer 后与位置编码生成器（PEG，Position Encoding Generator）模块生成位置信息向量相加，用于补充输入特征的帧先后位置信息。

1.4 质量增强模块

质量增强模块的核心思想是充分利用特征图F中的互补信息，生成增强的目标帧I_t^{HQ}。为利用残差学习，本文通过一个非线性映射$\mathcal{F}_{\theta_{qe}}(\boldsymbol{F})$来预测增强残差，公式为

$$\widehat{\mathcal{R}}_t^{HQ} = \mathcal{F}_{\theta_{qe}}(\boldsymbol{F}) \tag{11}$$

式（11）中，$\boldsymbol{F}$为光引导特征矩阵，映射$\mathcal{F}_{\theta_{qe}}(\boldsymbol{F})$是通过一个由步幅为 1 的多个卷积层组成的卷积神经网络（CNN，Convolutional Neural Networks）来实现，除了最后一层，所有层均有 3 个具有 ReLU 激活函数的卷积滤波器，最终层的卷积输出无失真画面和失真之后画面之间的残差$\widehat{\mathcal{R}}_{t_0}^{HQ}$。

则增强后的目标帧生成公式为

$$I_t^{HQ} = \widehat{\mathcal{R}}_t^{HQ} + I_t^{LQ} \tag{12}$$

式（12）中，$I_{t_0}^{LQ}$是t_0帧原始有损失真后的低质量画面。

2 损失函数

损失函数通过度量 OFGT-CVQE 方法生成视频帧和原始无失真视频帧间的差异，来指导方法的优化。该方法中包括 4 个损失函数，从不同维度量化增强帧与真实帧的差异，分别是空域重构损失L_{yc}和L_{yv}，频域幅值重构损失L_{amp}和对抗性损失L_{adv}。

2.1 空域重构损失函数

空域重构损失计算公式为

$$L_{yc} = \left\| \boldsymbol{M}_t \odot (Y_t - \widehat{Y}_t) \right\|_1 / \|\boldsymbol{M}_t\|_1 \tag{13}$$

$$L_{yv} = \left\| (1 - \boldsymbol{M}_t) \odot (Y_t - \widehat{Y}_t) \right\|_1 / \|(1 - \boldsymbol{M}_t)\|_1 \tag{14}$$

式（13）~式（14）中，$\boldsymbol{M}_t$为掩码矩阵；Y_t为真实的高质量视频帧；$\widehat{Y}_t$为增强得到的高质量视频帧；$\odot$表示按位置相乘。损失函数L_{yc}衡量的是压缩视频非失真区域经 OFGT-CVQE 方法增强前后的差异；损失函数L_{yv}衡量的是压缩视频失真区域经 OFGT-

CVQE 方法增强前后的差异。

2.2 频域幅值重构损失函数

为衡量增强帧与真实帧之间的频域差异，本文引入了频域幅值重构损失函数L_{amp}。在训练过程中，通过傅里叶变换将输入图像从空间域转换到频域。傅里叶变换后的频域信息包括实部和虚部，进而计算出振幅分量。振幅反映了频域信号的强度，能够有效反映图像在频域上的特征差异。频域幅值重构损失函数L_{amp}通过对两帧图像在频域振幅上的差异进行度量，使得网络能够更好地恢复图像的频域特征，进而提升增强质量。

2.3 对抗性损失函数

为进一步提升增强帧的真实性，本文采用对抗性损失函数L_{adv}。沿用文献 [12] 中的对抗性损失计算方法，通过引入一个判别器来评估增强图像的质量。判别器的目标是判断生成的图像是否为真实图像，而生成网络则通过最小化判别器的判定结果来优化自身参数，从而生成更逼真的增强视频帧。

综上，总损失函数L_y是以上各项损失的线性组合，公式为

$$L_y = \lambda_{y1}L_{yc} + \lambda_{y2}L_{yv} + \lambda_{y3}L_{amp} + \lambda_{y4}L_{adv} \tag{15}$$

借鉴已有研究，本文设置λ_{y1}和λ_{y2}为 1、λ_{y3}为 0.1、λ_{y4}为 0.01。

3 实验分析

3.1 数据集

本文采用某重载铁路沿线监控系统中采集的 130 段具有不同分辨率和内容的传输前未压缩视频，其中 106 个视频用于训练，其余 24 个视频用于验证，部分视频截图如图 3 所示。在实验测试阶段，视频序列包括不同分辨率，为模拟经传输压缩失真后的效果，所有视频均采用 H.265/HEVC 参考软件 HM16.5 进行压缩，配置为低延迟 P 模式（LDP，Low Delay P）。为评估在不同压缩级别下的性能，本文在 4 个不同的量化参数下进行视频压缩，分别为 22、27、32 和 37。

3.2 实验参数设置

本文实验基于 PyTorch 框架实现，使用 RAFT[13] 来估计光流。OFGT-CVQE 方法共采用 4 个时间和空间 Transformer 块；利用 Adam 优化器训练网络，训练迭代次数为 500 000 次，初始学习率设置为 1e^4，在迭代 120 000 次和 400 000 次后将学习率降低 10 倍。在训练阶段，对于每一个视频序列滑动窗口处理时，本文采样 5 个时间邻近的帧作为局部帧，并采样额外的 3 个帧作为全局帧。

图3　重载铁路场景中实际采集的视频数据

3.3 评价指标

本文采用峰值信噪比（PSNR，Peak Signal-to-Noise Ratio）和结构相似性指数（SSIM，Structural Similarity Index Measure）作为评价指标，用以衡量增强后图像帧的质量，二者均为数值越大，所评价的方法性能越好。

3.3.1 PSNR

给定一个大小为$h \times w$的原始图像X和一个带有噪声的压缩图像Y，它们之间的均方误差为

$$MSE(X,Y) = \frac{1}{hw}\sum_{i=0}^{h-1}\sum_{j=0}^{w-1}(X(i,j)-Y(i,j))^2 \tag{16}$$

基于此，PSNR公式为

$$PSNR(X,Y) = 10\log_{10}\frac{\max_I^2}{MSE(X,Y)} \tag{17}$$

式（17）中，$\max_I^2$是图像帧中可能的最大像素值。通常对于 uint8 的数据格式，最大像素值为 255，而对于浮点数据格式，最大像素值为 1。

3.3.2 SSIM

SSIM 的公式为

$$SSIM(X,Y)=\frac{(2\mu_X\mu_Y+c_1)(2\sigma_{XY}+c_2)}{(\mu_X^2+\mu_Y^2+c_1)(\sigma_X^2+\sigma_Y^2+c_2)} \quad (18)$$

式（18）中，μ_X和μ_Y分别为X和Y的像素点均值；σ_X^2和σ_Y^2表示 X 和 Y 的方差；σ_{XY}为X和Y协方差。为了避免除零，添加了常数c_1、c_2、c_3，$c_1=(k_1L)^2$，$c_2=(k_2L)^2$，$c_3=c_2/2$，一般取k_1=0.01，k_2=0.03。

3.4 实验结果对比分析

与目前广泛应用的基于 Transformer 的视频质量增强算法 FGT[14] 进行对比，证明 OFGT-CVQE 方法的有效性，实验结果如表 1 所示。在不同的量化参数设置下，由表 1 可知，OFGT-CVQE 方法的平均 PSNR 和 SSIM 均超越了 FGT 算法。在量化参数为 37 条件下，平均 PSNR 为 0.58 dB，SSIM 为 1.12%，相比于 FGT 显示出明显的优势。

表1 在 4 个不同量化参数下测试视频的实验结果

量化参数个数	方法	PSRN/dB	SSIM
37	FGT	0.46	0.88%
	OFGT-CVQE	0.58	1.12%
32	FGT	0.44	0.61%
	OFGT-CVQE	0.54	0.89%
27	FGT	0.41	0.37%
	OFGT-CVQE	0.51	0.63%
22	FGT	0.33	0.17%
	OFGT-CVQE	0.48	0.43%

3.5 实验可视化效果

OFGT-CVQE 方法的可视化效果如图 4 所示，压缩后的帧因各种压缩伪影而遭受严重扭曲，视频质量增强方法则在参考帧的帮助下取得了更好的效果。OFGT-CVQE 方法在对抗压缩伪影方面表现出更强的鲁棒性，能够有效地利用时空信息，更精确地恢复结构细节。

3.6 消融实验

为证明光流特征对压缩视频的质量增强具有促进作用。本文将光流补全与特征提取模块替换为算法 FGVC 中的光流特征提取方案[15]，结果如表 2 所示，表 2 中，“FGVC→OFGT-CVQE”表示采用从 FGVC 中补全的光流来执行特征融合与传播。实验结果表明，OFGT-CVQE 方法的 PSRN 提升了 1.33%，SSIM 提升了 0.89%，增强后图像帧的质量有一定的提升，能更合理地补全光流引导特征融合与传播后的区域

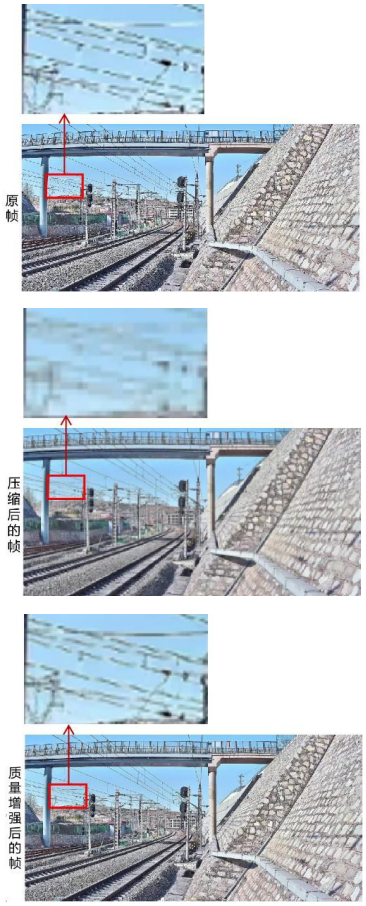

图4 实验可视化效果展示

表2 光流特征的有效性消融实验

方法	PSRN/dB	SSIM
FGVC→OFGT-CVQE	35.12	0.69%
OFGT-CVQE	36.45	1.58%

填充，证明了本文所采用的光流方法所形成的准确运动轨迹在视频增强中的重要性。

4 结束语

本文提出了一种基于光流引导 Transformer 模型的重载铁路监控压缩视频质量增强方法，该方法在多个层面上利用了光流引导。为了解决查询退化问题，设计了光流引导特征融合子模块和光流引导特征传播子模块，对特征的相关性进行时间建模并沿时间维度进行特征传播获得具有高可表示性的特征表示。此外，在时间和空间维度上对时空 Transformer 进行了解耦，设计光流引导的多头注意力机制，对 Transformer 中的查询、键和值矩阵的 RGB 帧信息和光流信息同时建模，有助于挖掘二者中管理。实验

结果验证了本文所提方法的有效性。

本文研究仍存在局限性，性能在很大程度上依赖于光流补全的质量，当视频中物体在相邻两帧有较大位移像素时，补全的光流可包含较大误差，从而影响光流引导的效果。下一步，应重点关注解决上述问题，并且将光流引导的想法应用到更多基于 Transformer 的视频处理方法中。

参考文献

[1] Bertalmio M, Bertozzi A L, Sapiro G. Navier-stokes, fluid dynamics, and image and video inpainting[C]//2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 8-14 December, 2001, Kauai, HI, USA. New York, USA: IEEE, 2001. I.

[2] Granados M, Tompkin J, Kim K, et al. How not to be seen —object removal from videos of crowded scenes[J]. Computer Graphics Forum, 2012, 31(2pt1): 219-228.

[3] Matsushita Y, Ofek E, Ge W N, et al. Full-frame video stabilization with motion inpainting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(7): 1150-1163.

[4] Ebdelli M, Le Meur O, Guillemot C. Video inpainting with short-term windows: application to object removal and error concealment[J]. IEEE Transactions on Image Processing, 2015, 24(10): 3034-3047.

[5] Granados M, Kim K I, Tompkin J, et al. Background inpainting for videos with dynamic objects and a free-moving camera[C]//12th European Conference on Computer Vision, 7-13 October, 2012, Florence, Italy. Berlin, Heidelberg: Springer, 2012: 682-695.

[6] Newson A, Almansa A, Fradet M, et al. Video inpainting of complex scenes[J]. SIAM Journal on Imaging Sciences, 2014, 7(4): 1993-2019.

[7] Huang J B, Kang S B, Ahuja N, et al. Temporally coherent completion of dynamic video[J]. ACM Transactions on Graphics (ToG), 2016, 35(6): 196.

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//31st International Conference on Neural Information Processing Systems, 4-9 December, Long Beach, CA, USA. Red Hook: Curran Associates Inc. , 2017: 6000-6010.

[9] Liu R, Deng H M, Huang Y Y, et al. FuseFormer: fusing fine-grained information in transformers for video inpainting[C]//2021 IEEE/CVF International Conference on Computer Vision, 10-17 October, 2021, Montreal, QC, Canada. New York, USA: IEEE, 2021: 14020-14029.

[10] Liu R, Deng H M, Huang Y Y, et al. Decoupled spatial-temporal transformer for video inpainting[J]. arXiv: 2104.06637, 2021.

[11] Li Z, Lu C Z, Qin J H, et al. Towards an end-to-end framework for flow-guided video inpainting[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 18-24 June, 2022, New Orleans, LA, USA. New York, USA: IEEE, 2022: 17541-17550.

[12] Song M Y, Zhang Y, Aydın T O. TempFormer: temporally consistent transformer for video denoising[C]//17th European Conference on Computer Vision, 23-27 October, 2022, Tel Aviv, Israel. Cham: Springer, 2022: 481-496.

[13] Teed Z, Deng J. RAFT: recurrent all-pairs field transforms for optical flow[C]//16th European Conference on Computer Vision, 23-28 August, 2020, Glasgow, UK. Cham: Springer, 2020: 402-419.

[14] Zhang K D, Fu J J, Liu D. Flow-guided transformer for video inpainting[C]//17th European Conference on Computer Vision, 23-27 October, 2022, Tel Aviv, Israel. Cham: Springer, 2022: 74-90.

[15] Gao C, Saraf A, Huang J B, et al. Flow-edge guided video completion[C]//16th European Conference on Computer Vision, 23-28 August, 2020, Glasgow, UK. Cham: Springer, 2020: 713-729.

责任编辑 李依诺