



注意力机制的秦腔视频去噪算法

师秦高雪^{1,2}, 杨超然^{1,2}, 刘鑫达^{1,2}, 耿国华^{1,2}

(1. 西北大学 文化遗产数字化国家地方联合工程研究中心, 陕西 西安 710127;

2. 西北大学 可视化技术研究所, 陕西 西安 710127)

摘要 秦腔, 作为中国传统戏曲艺术的瑰宝, 拥有深厚的历史底蕴。然而, 秦腔早期的影像资料常受噪声和失真影响, 导致画质不佳, 严重妨碍了秦腔数字档案的保存品质。目前应用的视频去噪技术在处理秦腔那色彩丰富、纹理复杂的服饰时, 往往没有充分利用视频帧序列的时间连贯性, 使得去噪效果并不理想, 难以有效保留视频帧的核心特征。基于注意力机制的秦腔视频去噪算法开展研究, 针对现有视频去噪算法忽略帧间时序相关性导致效果不佳的问题, 提出了一种新的视频去噪算法, 该算法利用双门控注意力机制进行时序信息的融合。首先, 通过时序融合模块, 将视频连续帧的时序信息进行有效整合; 其次, 利用双门控注意力去噪网络精确识别并消除时序上的噪声; 最后, 通过多头交互注意力精炼模块进一步细化特征, 以消除去噪过程中可能产生的伪影并恢复丢失的细节, 从而提升去噪后图像的质量。实验结果表明, 与 DVDNet、ViDeNN 以及 FastDVDNet 等现有方法相比, 该方法可以更好地利用视频的时序信息, 达到干净且高效的秦腔视频去噪效果。

关键词 秦腔; 视频去噪; 注意力机制; 时序融合

中图分类号: TP391 **DOI:** 10.16152/j.cnki.xdxzbzr.2025-01-014

Qin Opera video denoising algorithm based on attention mechanism

SHI Qingaoxue^{1,2}, YANG Chaoran^{1,2}, LIU Xinda^{1,2}, GENG Guohua^{1,2}

(1. National and Local Joint Engineering Research Center for Cultural Heritage Digitization, Northwest University,

Xi'an 710127, China; 2. Institute of Visualization Technology, Northwest University, Xi'an 710127, China)

Abstract Qin Opera, as a treasure of Chinese traditional theatre art, has a profound historical heritage. However, the early video materials of Qin Opera are often affected by noise and distortion, resulting in poor picture quality, which seriously hampers the preservation quality of Qin Opera digital archives. Currently applied video denoising techniques often do not make full use of the temporal coherence of the video frame sequence when dealing with the colorful and complex texture of Qin Opera's costumes, which makes the denoising effect unsatisfactory and makes it difficult to effectively retain the core features of the video frames. In this paper, we carry out research on the Qin Opera video denoising algorithm based on the attention mechanism,

收稿日期: 2024-10-15

基金项目: 国家自然科学基金(62271393); 文化和旅游部重点实验室项目(1222000812, cr2021K01); 西安市社会发展科技创新示范项目(2024JH-CXSF-0014)。

第一作者: 师秦高雪, 女, 从事虚拟现实、图像处理研究, shiqingaoxue@stumail.nwu.edu.cn。

通信作者: 耿国华, 女, 教授, 博士生导师, 从事智能信息处理、虚拟现实与可视化研究, ghgeng@nwu.edu.cn。

and the main research contents are as follows: Aiming at the existing video denoising algorithms ignoring the temporal correlation between frames which leads to the problem of poor effect, we propose a new video denoising algorithm, which makes use of the double gating attention mechanism for the fusion of the temporal sequence information. The algorithm firstly integrates the timing information of consecutive video frames effectively through the timing fusion module; then accurately identifies and eliminates the timing noise using the dual-gated attention denoising network; finally, the features are further refined through the multi-head interactive attention refining module to eliminate the artifacts that may be generated during the denoising process and recover the lost details, to enhance the quality of the denoised image. The experimental results demonstrate that compared with existing methods such as DVDNet, ViDeNN, and FastDVDNet, this method can make better use of the timing information of the video to achieve clean and efficient denoising of Qin Opera videos.

Keywords Qin Opera; video denoising; attention mechanism; temporal fusion

秦腔,作为中国戏曲艺术的重要组成部分,是中华文化宝库中的璀璨明珠。信息技术的应用极大地促进了秦腔等文化遗产的数字化记录,便于其再现与再利用。然而,秦腔数字化过程中面临诸多挑战,如老旧录音录像资料的噪声和失真问题。尽管秦腔拥有深厚的历史底蕴和丰富的艺术表现形式,但历史遗留下来的录音和录像资料往往因质量低下而受到噪声和失真的困扰。受限于存储与技术水平,部分秦腔资料已损坏或不完整,影响了数据的完整性和可用性。

随着科技的不断进步,秦腔这一传统艺术形式迎来了新的传播和发展空间。人工智能和数字化技术的应用,不仅为秦腔艺术的保护、展示和传承提供了有力支持,而且极大地提升了秦腔艺术的数字化展示效果。在这个过程中,确保秦腔数字资源的品质至关重要,是维护和传递秦腔文化的核心环节。在秦腔图像处理的领域中,去噪技术的应用成为了一个关键步骤。这一过程首先需要对秦腔表演的图像进行预处理,包括调整图像的灰度和尺寸等,以便更好地适应后续处理流程。其次,通过先进的图像处理技术,对图像中的噪点进行识别和分析,从而确定噪点的类型和具体位置。针对秦腔表演图像的独特性,选择恰当的去噪算法是至关重要的,不仅能够有效地消除图像中的噪点和干扰,还能确保图像的核心信息和细节得到保留。在去噪操作完成后,通过一些后处理步骤,例如图像锐化和对比度优化,可以进一步提升图像质量,从而获得清晰、生动的秦腔表演图像。

图像去噪在计算机视觉领域扮演着至关重要的角色,且致力于精确地消除图像中的噪声并恢复其原始风貌。该技术不仅能提升图像的视觉感

知品质,还能图像识别、分析和理解等后续处理任务提供更为清晰和精确的数据基础。图像去噪的应用增强了数字化图像记录的品质,确保了图像保存的清晰度和真实感,对于秦腔等艺术形式的记录和传播至关重要,有助于其艺术表现的长久保存与广泛传播。

随着数字化技术的发展,文化遗产的保护工作已经进入了一个新阶段。在这一领域,周明全等人在其著作中全面探讨了文化遗产数字化保护的技术和应用^[1],为该领域的研究者提供了宝贵的知识资源。本研究提出的秦腔视频去噪算法,不仅在视频处理领域展现出巨大的应用潜力,也为文化遗产的数字化保护与活化提供了强有力的技术支持。这一成果与耿国华等人所强调的观点相呼应,他们指出文化遗产的活化迫切需要创新技术的支撑^[2]。因此,针对现有视频去噪算法在保留秦腔图像重要细节等方面,未能有效利用帧之间高度相关的内容的问题,提出了一种基于双门控注意力的时序融合视频去噪算法(twin gate attention-temporal fusion network, TGA-TFNet)。该算法首先通过时序融合模块整合连续帧之间的信息,并利用时序上的连续性和冗余信息来提升去噪性能。其次,采用双门控注意力去噪网络,有效识别并去除时序上的噪声,同时保留重要的时序信息,例如运动细节和连贯性。最后,通过多头交互注意力精炼模块对特征进行再次细化,以消除去噪过程可能产生的伪影,并恢复因去噪导致的过度平滑而丢失的一些细节,从而提高去噪结果的质量。

与将视频分割成单帧处理的现有方法相比,本研究能够更好地挖掘视频的时序信息,从而提高去噪性能。这一创新性方法不仅有助于数字化

技术在文化遗产保护中的应用,也为秦腔文化的数字化保护与传承开辟了新的可能性,为秦腔艺术的传承与发展注入了新的活力。

1 相关工作

视频去噪是计算机视觉领域中一个重要的研究方向,其目标是从含有噪声的视频中恢复清晰的图像。早期视频去噪技术主要使用帧内和帧间的滤波方法^[3],然而这些方法难以充分利用视频的时空信息。随着深度学习的兴起,视频去噪得到了新的解决方案,能够更有效地捕捉视频数据的复杂性。最近,自监督学习领域也取得了进展,例如 Liu 等人提出的 FeaSC 方法^[4],通过减少视图间的互信息,增强了自监督预训练的有效性。Maggioni 等人成功地利用视频的时空信息进行了去噪^[5],但是该方法在处理复杂场景时仍存在一定的限制。为了克服这一问题,Davy 等人提出了基于深度学习的时空去噪网络^[6],通过学习视频数据的深层特征来进行去噪,相比传统方法,该方法能够更好地恢复细节并减少模糊。然而,这种方法的网络参数较多,需要大量计算资源。而 Tassano 等人提出的 DVDNet 通过创新的双路径网络结构有效地融合了帧内和帧间信息^[7],用于视频去噪。但该方法的性能受限于高计算资源需求并且对相邻帧质量的依赖较高。另外,Xue 等人提出的 TOFlow 通过端到端可训练的网络^[8],将运动估计与视频处理集成在一起,显著提高了性能。尽管这种方法能够学习特定任务的运动表示,优于传统的光流方法,但可能需要较大的计算资源,并且可能受到实时应用的限制。Wang 等人提出的 FITVNet 通过先对单帧图像进行去噪^[9],然后应用时空去噪模块来处理整个视频,有效地解决了快速移动物体边界模糊的问题,尤其提高了物体边界处的去噪质量。然而,该方法的去噪能力在极端情况下有限,并且在处理非常复杂的场景时可能受到限制。Mehta 等人提出的 EVR-Net 采用轻量级的网络设计^[10],能够显著降低参数和计算成本,同时还能保持与同时期方法相竞争的性能。

Tassano 等人提出了 FastDVDNet^[11],结合了注意力机制和端到端学习,在无需显式运动补偿的情况下实现了实时视频去噪。该方法显著提高了视频去噪的效率并保持了良好的去噪效果。然

而,在极端噪声条件下,该方法的去噪效果仍有改进的空间。Vaksman 等人引入了补丁工艺帧的概念^[12],通过拼接匹配的补丁构建与真实帧相似的人工帧,并将视频序列与补丁工艺帧结合后送入 CNN,从而显著提高了去噪性能。然而,当处理大量数据以生成每个输出帧时,可能需要较大的计算资源,这可能限制了在资源受限环境中的应用。Maggioni 等人提出了 EMVD^[13],通过循环方式应用多个级联处理阶段,包括时间融合、空间去噪和时空细化。该方法递归地利用自然视频中固有的时空相关性,可以显著降低模型复杂性,同时不会严重影响性能。Song 等人提出了 TempFormer^[14],使用小波变换预处理降低视频分辨率以提高效率,并通过空间时间 Transformer 块和联合空间时间混合模块来学习空间和时间注意力。Li 等人提出的方法可以隐式地捕捉帧间的多帧聚合对应关系^[15],通过引入分组空间位移,获得广阔的有效感受野,并有效地聚合帧间信息。Liang 等人提出的 VRT 方法聚合视频序列的局部和全局特征信息,并对相邻帧进行融合^[16]。该方法具有平行帧预测和长距离时间依赖性建模的能力。

综上所述,尽管当前的视频去噪算法在提升视频流畅度方面取得了显著成效,但它们在处理特定类型的视频内容如秦腔视频时,仍然面临一系列独特的挑战。秦腔视频的去噪工作不仅要求提高视频的流畅度,更关键的是要消除伪影和恢复因过度平滑而丢失的细腻表情与动作细节,这些是秦腔艺术表现力的核心。秦腔视频因其独特的古典美感、丰富的表情变化和传统文化背景而备受关注,使得其去噪处理不仅要注重技术性能,还要考虑到文化传承的准确性。

秦腔视频的质量问题,往往与视频的动态范围、色彩饱和度和纹理细节的保留紧密相关。这些因素在传统的视频去噪算法中可能未被充分考虑,导致在去噪的同时损失了秦腔视频的艺术特色。因此,设计一种能够兼顾时序融合、细节保留、计算效率和适应性的秦腔视频去噪算法,成为本研究的核心目标。

本研究将深入分析秦腔视频的特性,如其特有的表演节奏、面部表情的微妙变化以及服饰的精细纹理,从而开发出更具针对性的去噪算法。通过这种针对性的设计,旨在提升秦腔视频质量的处理能力,确保在去噪过程中能够最大限度地保留秦腔的艺术魅力。该方法不仅能够提高视频

的观赏性,也为保护和传承这一珍贵文化遗产提供了技术支持。

2 TGA-TFNet

本文提出了一种基于双门控注意力的时序融合视频去噪算法 (twin gate attention-temporal fusion network, TGA-TFNet)。该算法能够有效捕捉相邻帧之间的相关性,保证去噪结果的时序一致性,并避免引入伪影以及修复因过度平滑而损失的细节,从而生成高质量的去噪视频。TGA-TFNet 主要由时序融合模块、双门控注意力去噪网络以及多头交互注意力精炼模块组成,这些组成

部分旨在解决秦腔视频去噪过程中的关键挑战。通过与秦腔视频特征的紧密结合,TGA-TFNet 在保持视频时序性的同时,充分保留了视频细节,提高了去噪效果,展现出对秦腔视频处理的显著优势。下面将详细介绍这些模块的功能和作用。

2.1 网络结构

本文提出的基于双门控注意力的时序融合视频去噪算法如图 1 所示。该方法的具体流程包括时序融合、去噪、精炼与细化 3 个步骤。输入包括当前噪声帧 $\text{noisy}(t)$ 和先前帧序列的特征融合特征 $\text{fused}(t-1)$,输出为当前帧以及当前帧的去噪与细化结果 $\text{refine}(t)$ 。

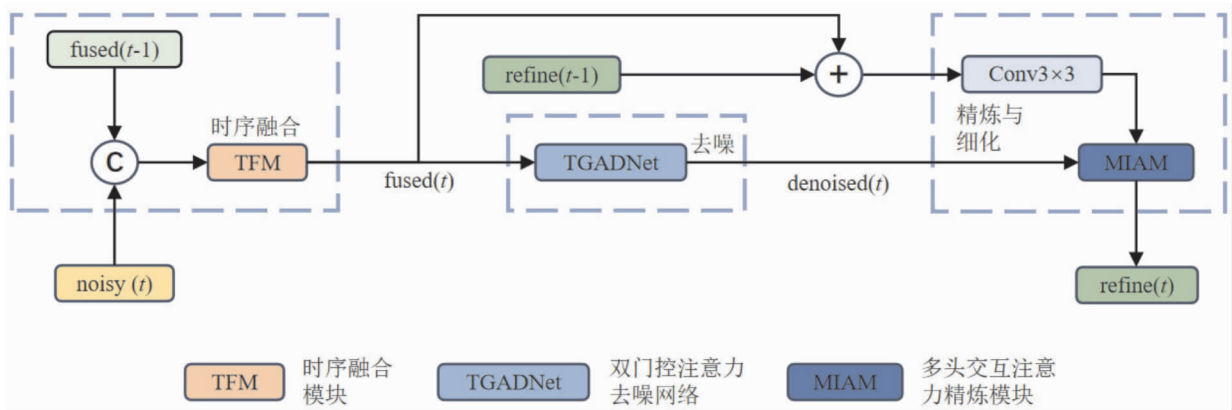


图 1 基于双门控注意力的时序融合视频去噪算法的网络结构

Fig. 1 Network architecture of twin gate attention-temporal fusion network

本文提出的基于双门控注意力的时序融合视频去噪方法,包括如下步骤。

1) 时序融合。将输入的两个帧 $\text{noisy}(t)$ 和 $\text{fused}(t-1)$ 传入时序融合模块,得到新的融合特征 $\text{fused}(t)$ 。该步骤的目标是利用视频中固有的时间相关性,最大程度地减少图像中的噪声,同时不引入任何时间伪影。

2) 去噪。将上一步骤得到的融合特征 $\text{fused}(t)$ 输入双门控注意力去噪网络,进行去噪处理,得到去噪后的视频帧 $\text{denoised}(t)$ 。此步骤的目的是利用融合特征 $\text{fused}(t)$ 中的时间冗余信息,精确且高效地去除噪声。

3) 精炼与细化。该步骤的输入由上两步骤生成的 $\text{denoised}(t)$ 和 $\text{fused}(t)$,以及先前帧序列的去噪与细化结果 $\text{refine}(t-1)$ 的卷积结果组成。将这两个输入送入多头交互注意力精炼模块,对去噪后的视频帧进行精炼和细化,得到经过去噪和细化处理的干净视频帧 $\text{refine}(t)$ 。该步骤的目的是消除去噪过程中引入的伪影,并恢复因过度

平滑而损失的细节。

2.2 时序融合模块

在视频去噪过程中,有效利用时序信息是提高去噪性能的关键。由于视频帧与帧之间的内容具有高度相关性,这种相关性可以用来增强单个帧中的信号并抑制噪声。然而,传统的视频去噪网络通常只关注单帧内的特征提取,而忽视了帧与帧之间的时序关联。为了克服这一限制,本文设计了一个时序融合模块 (temporal fusion model, TFM)。

该模块的主要目的是整合连续帧之间的信息,以增强当前帧的信号并减少噪声。通过引入先前帧的融合特征和当前的噪声帧,该模块旨在利用时序上的连续性和冗余信息来提升去噪性能。如图 2 所示,首先进行拼接操作,将当前的噪声帧 $\text{noisy}(t)$ 与先前帧中的融合特征 $\text{fused}(t-1)$ 结合,以整合时间维度上的信息。其次,利用两个连续的 3×3 卷积层 (使用 ReLU 激活函数) 对融合后的数据进行特征提取,从而增强有用信号并

抑制噪声。随后,采用另一个带有 Sigmoid 激活函数的 3×3 卷积层,动态地调整特征图中每个元素的重要性。接下来,将先前的融合结果与 Sigmoid 输出相乘,这一步骤使网络能够有选择地增强关键特征,得到的结果再与将先前的融合结果相加。最后,利用另一个 3×3 卷积层对特征进行整合和细化,得到当前帧的融合特征 $fused(t)$,为下一阶段的去噪网络提供保留更多时序特征的特征集。通过设计时序融合模块(TFM),该方法能够更有效地利用视频序列中的时序冗余性,不仅提高去噪网络对时间相关特征的捕捉能力,而且增强网络在处理高动态场景时的稳健性。

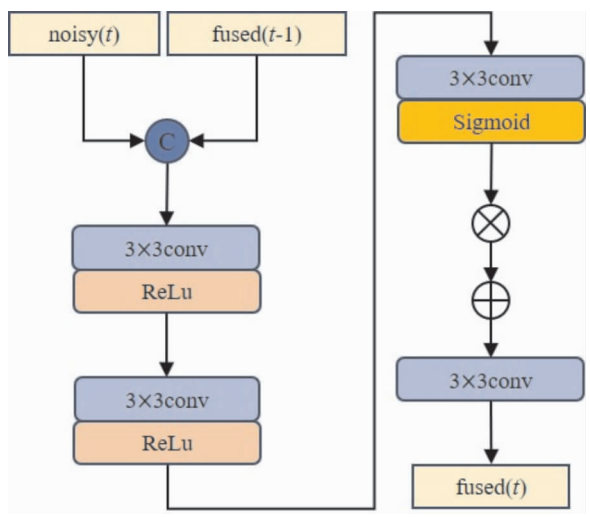


图 2 TFM 结构图
Fig. 2 TFM structure diagram

2.3 双门控注意力去噪网络

本小节提出了一种基于 HIN 的双门控注意力去噪网络 (twin gated attention denoising network, TGADNet),该网络采用 HIN 的整体结构,并引入双门控注意力模块 (twin gated attention module, TGAM),专门处理视频序列的时序特征。

HINet 利用半实例归一化 (half instance normalization, HIN) 结合了批量归一化 (batch normalization, BN) 和实例归一化 (instance normalization, IN) 的特性。通过在 BN 和 IN 之间建立平衡, HINet 能够保持图像内容的稳定性,同时增强网络对于图像细节和纹理的恢复能力。然而,这些方法仍然不足以满足视频去噪的需求。因此,引入了 TGAM 作为 HINet 编码器和解码器之间的桥梁,以帮助模型更好地理解 and 重建秦腔视频帧之

间的动态变化。TGAM 使网络能够识别和调节对空间噪声特征的关注,并根据视频帧之间的动态变化调整注意力分布。通过这样的设计,本小节提出的 TGADNet 在处理秦腔视频序列时能够有效地识别和去除时序上的噪声,同时保留重要的秦腔视频时序信息,例如运动细节和连贯性。

如图 3 所示, TGADNet 由两个子网络组成,这两个子网络通过跨阶段特征融合模块 (cross-stage feature fusion, CSFF) 和监督注意力模块 (supervised attention module, SAM) 进行连接。这两个模块的设计灵感来自于 Zamir 的方法^[17]。

首先,将带有时序信息的融合特征送入第一个子网络,对视频帧进行第一阶段的去噪。在去噪过程中,每一层所产生的特征以及第一阶段的去噪结果分别通过跨阶段特征融合模块 (cross-stage feature fusion, CSFF) 和监督注意力模块 (supervised attention module, SAM) 传递到下一个阶段进行聚合。CSFF 模块用于丰富下一阶段的多尺度特征,而 SAM 模块则用于强调重要特征并抑制次要信息。接下来,第二个子网络接收第一个子网络提供的输入,进行第二阶段的去噪,并输出去噪后的特征 denoised。

每个子网络都采用 U-Net 结构。对于每个阶段的 U-Net,首先通过单层卷积获取待去噪特征图的浅层特征,然后将特征送入编解码架构中 (4 个下采样 + 4 个上采样)。编码器部分使用 HIN Block 来提取每个尺度的特征,并在下采样过程中增加通道数。编码器最底部的输出被送入一组 TGAM 中,以从编码器特征中获取关键的去噪线索。TGAM 将经过门控注意力加权的特征送入解码器中。解码部分通过 Pixel Shuffle 进行上采样,以避免传统上采样方法引入伪影或模糊问题,并恢复或增加图像的细节和清晰度。同时,解码部分使用 ResBlock 来提取高级特征,并与 HIN Block 提取的特征进行融合,以补偿反复采样导致的信息损失。最终,输出每个阶段的去噪结果。

自注意力 (self-attention, SA) 模块对于去噪很有帮助,因为它可以捕获远程依赖关系,从而增加接受野。然而,由于 SA 模块需要计算序列中每对元素之间的相互作用,因此在处理高分辨率图像时,这些模块通常具有很高的计算复杂性。门控注意力 (gate attention, GA) 是对传统自注意力机制的一种改进。

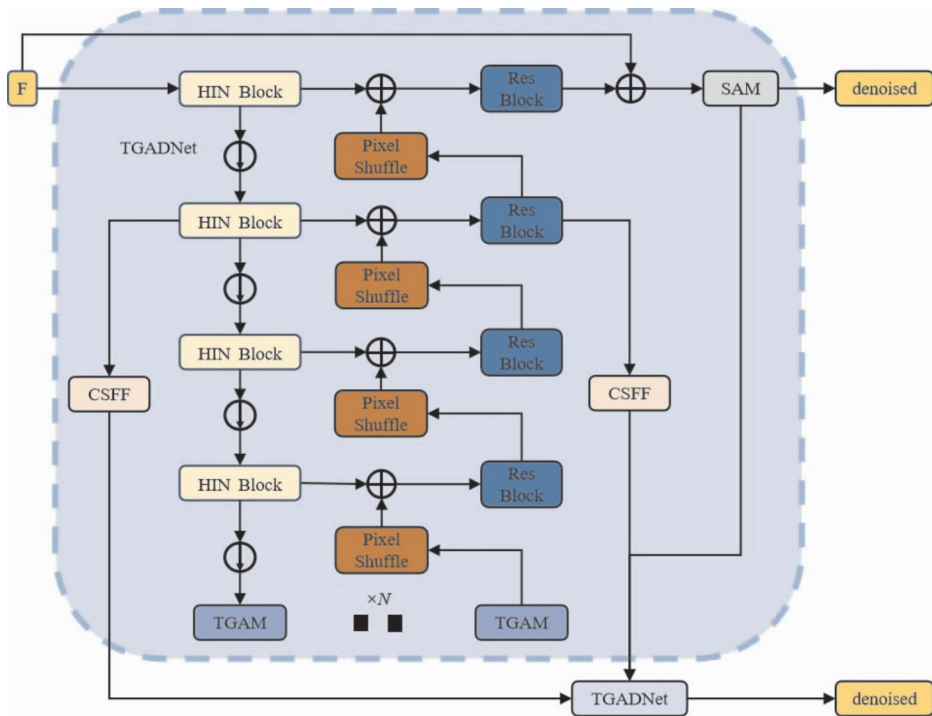


图 3 TGADNet 结构图

Fig.3 TGADNet structure diagram

为了在减少计算负担的同时提高去噪过程的效率,本小节采用了 TGAM。TGAM 通过引入额外的门控机制来动态调整注意力的聚焦程度,以调节不同元素之间注意力权重的分配。这样可以使模型更加专注于噪声密集或重要区域,从而提高去噪的效率和性能。此外, TGAM 通过扩展特征图的通道并将其分流送入可学习矩阵,以对频域中的依赖性进行建模,从而捕获特征中的长距离依赖关系,并大大降低了计算成本。

TGAM 的结构图如图 4 所示。首先,输入特征图 F_{in} 经过 LayerNorm 层进行归一化。其次,通过 1×1 卷积将特征图的通道数从 c 扩展到 $(2r_1 + r_2)c$, 其中 $r_1 \leq 1$ 和 $r_2 \leq 1$ 是用于控制通道冗余的缩减因子的超参数。 $(2r_1 + r_2)c$ 通道被分为 3 个流, 每个流具有 r_1c , r_1c 和 r_2c 数量的通道, 以捕获远程依赖关系。接下来, 在两个流中分别两次捕获频域中的远程依赖关系, 得到 F_0 和 F_1 。最后, 通过一系列点乘和 1×1 卷积的降维操作, 并与 F_{in} 进行残差连接, 得到经过门控注意力加权的特征图 F_{out} 。上述提到的两次捕获频域中的远程依赖关系是通过学习一个 $H \times W \times C$ 的矩阵 ω (与特征映射的大小相同) 来建模的。与普通卷积相比, 这样的计算成本更低。以图 4 中的 F_1 流为例, 首先使用二维快速傅里叶变换^[18] (2D FFT) 将特征映射转换

到频域,得到 F_{FFT} ;其次,将学习到的矩阵 ω 与 F_{FFT} 相乘,以捕捉频域的依赖关系,得到特征 F'_{FFT} ;最后,将 F'_{FFT} 作为来自 r_2c 通道的特征映射的门控信号,执行门控注意力并最终输出 F_1 。相比于普通卷积^[19],这种方式将计算复杂度从卷积运算的 $O(N^2)$ 降低到元素矩阵乘法的 $O(N)$,其中 N 表示特征映射中的所有像素点。

2.4 多头交互注意力精炼模块

任何去噪方法都有可能引入伪影和图像细节的丢失,特别是当输入图像的信噪比较差或模型的复杂性受到明显约束时。因此,为了恢复因去噪而丢失的精细细节和纹理,并进一步提高去噪效果和精炼特征表示,提出了一种多头交互注意力精炼模块(multi-head interaction attention model, MIAM)。MIAM 综合考虑了仍带有噪声的图像、无噪声的去噪图像以及经过精炼的上一帧图像。它将这些图像结合起来,用于精炼去噪后的特征。该模块通过整合当前帧的去噪特征和时间维度上的信息(噪声图像特征和上一个细化帧的融合特征),旨在增强特征表示并实现更好的时序一致性。使用 MIAM 模块的目的是在特征级别上提高去噪效果,同时恢复丢失的细节和纹理。这种方法能够处理复杂的噪声情况,并提供更准确的图像去噪结果。

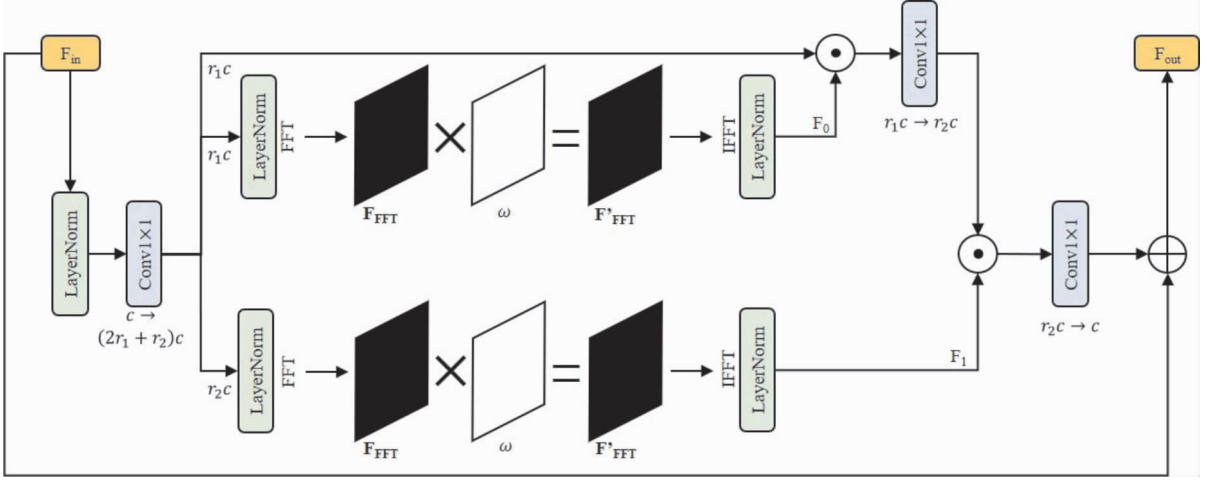


图 4 TGAM 结构图

Fig. 4 TGAM structure diagram

交互注意力 (interaction attention, IA) 中将去噪后的特征作为支持特征 $F^S \in \mathbf{R}^{N \times C}$, 携带着噪声特征和细化特征帧的融合特征, 作为参考特征 $F^R \in \mathbf{R}^{N \times C}$, 其中 N 表示通道内的像素数。参考特征的查询 Q^R 、支持帧的键 K^S 和值 V^S 通过线性层表示为

$$Q^R = F^R P^Q, K^S = F^S P^K, V^S = F^S P^V \quad (1)$$

式中: $P^Q, P^K, P^V \in \mathbf{R}^{C \times C}$ 是线性层。接下来的 IA 注意力计算公式为

$$IA(F^R, F^S) = \text{Softmax}\left(\frac{Q^R (K^S)^T}{\sqrt{C}}\right) V^S \quad (2)$$

受多头自注意力的启发, 将输入分组, 生成多组查询、键、值; 然后将多组查询、键、值并行地进行注意力计算, 将结果汇聚拼接; 最后, 经过一个全连接层, 最终得到输出。以这样的方式交互注意力成为多头交互注意力 (multi-head interaction attention, MIA), 这样网络可以同时关注输入的不同部分, 从而更好地捕捉输入特征中的信息。

多头交互注意力精炼模块的工作流程表示为式(3), F_1 表示当前需要细化的特征, F_2 则是上一个细化帧与噪声图像特征的融合结果。MIA(F_1, F_2) 代表了通过对齐操作, 利用 F_2 对 F_1 进行多头交互注意力的处理。Concat 表示通道拼接操作, $MLP(x)$ 表示多层感知机 (multilayer perceptron), 而 $LN(x)$ 表示层归一化 (layer normalization) 操作。首先, 对 F_1 和 F_2 进行两次多头交互注意力计算, 将得到的注意力结果分别表示为 Y_1 和 Y_2 。其次, 将 Y_1 和 Y_2 进行通道拼接操作, 将其输入到多层感知机中, 以进行通道缩减。将多层感知机的输出结果与 F_1 进行残差连接, 得到特

征 F_3 。最后, 对 F_3 进行层归一化操作, 将其输入到另一个多层感知机中, 进行进一步的特征转换。将多层感知机的输出结果与输入的 F_3 进行残差连接, 即可获得经过细化和精炼后的视频帧 F_{out} 。

$$Y_1, Y_2 = MIA(F_1, F_2), MIA(F_2, F_1)$$

$$F_3 = MLP(\text{Concat}(Y_1, Y_2)) + F_1 \quad (3)$$

$$F_{out} = MLP(LN(F_3)) + F_3$$

通过 MIAM, 网络能够利用融合特征中的噪声图像信息, 更精细地区分噪声和有用信号, 从而在去噪后的视频中保留更多的细节和纹理信息, 恢复因去噪而模糊或丢失的图像细节。同时, 通过融合特征中上一帧的特征, 网络可以更好地理解帧与帧之间的关系, 能够在保留重要细节的同时, 识别并减少去噪过程中引入的不自然伪影。此外, 在去噪过程中还能保持视频的时间连贯性, 避免去噪后视频出现抖动或不自然的过渡。

2.5 损失函数

本小节旨在通过使用峰值信噪比 (PSNR) 和时间一致性损失作为优化目标, 来提高视频序列中图像的质量, 并确保帧与帧之间的时间连续性和一致性。这不仅有助于减少噪声, 还有效避免对秦腔表演中精细的面部表情和复杂的服饰纹理造成损害, 具体描述如下。

2.5.1 峰值信噪比 (PSNR)

峰值信噪比 (PSNR) 是最常用的衡量图像重建质量的指标之一。它通过比较原始图像与去噪图像之间的最大可能像素值差异和实际像素值误差的比率来评估图像的质量。PSNR 的计算公式为

$$L_{\text{PSNR}} = 10 \log_{10} \left(\frac{I_{\text{max}}^2}{e_{\text{MSE}}} \right) \quad (4)$$

式中: $f_{\max}^2$ 表示图像可能的最大像素值,对于 8 位图像,该值通常为 255; e_{MSE} (均方误差)是原始图像和去噪图像之间平均误差的平方。PSNR 值越高,表示去噪图像与原始图像越接近,图像质量越好。在本小节中,通过最大化 PSNR 值来优化图像的视觉质量,对于秦腔这类传统艺术形式来说尤为重要,因为它们通常包含丰富的细节和纹理,需要在去噪的同时保留其独特的艺术特点。

2.5.2 时间一致性损失

在处理秦腔这类传统戏曲艺术视频时,除了在单个图像帧上进行高质量重建之外,视频去噪还需要考虑帧与帧之间的时间一致性。为此引入了时间一致性损失函数,以确保相邻帧之间的去噪结果在视觉上保持连续性和一致性。这在秦腔视频去噪中尤为重要,因为秦腔的表演艺术强调唱、念、做、打的连贯性,任何帧与帧之间的不连贯都可能破坏其艺术表现力。时间一致性损失通过比较连续帧的去噪输出与原始视频帧之间的差异来计算,其计算公式为

$$L_{\text{temp}} = \frac{1}{N-1} \sum_{i=1}^{N-1} \|f(d_i) - f(d_{i+1}) - (o_i - o_{i+1})\|_2 \quad (5)$$

式中: d_i 和 d_{i+1} 表示连续的去噪帧; o_i 和 o_{i+1} 表示对应的原始帧; $f(x)$ 表示特征提取函数; N 是视频序列中的总帧数。该损失项旨在鼓励模型生成的去噪帧在时间上保持一致性,减少视觉抖动。

2.5.3 总损失

首先,采用 PSNR 来调整每个完整图像的像素值差异,以提高单帧去噪图像的质量,并保留图像中的颜色和纹理特征,这对于秦腔表演中丰富的表情和服饰细节尤为重要。其次,使用时间一致性损失函数聚焦于视频的时序特征,以确保去噪过程不会引入任何不自然的帧间跳动或抖动,从而保持视频播放的平滑性和连贯性,这对于秦腔表演的流畅性和整体艺术表现至关重要。总损失函数表示为

$$L_{\text{total}} = \alpha L_{\text{PSNR}} + \beta L_{\text{temp}} \quad (6)$$

式中: L_{PSNR} 和 L_{temp} 分别代表 PSNR 损失和时间一致性损失; α 和 β 是用于平衡两种损失贡献的权重参数。

3 实验与分析

在提升秦腔视频的去噪品质中,TGA-TFNet

方法凸显了其卓越性。该方法在确保去噪后视频序列的时间连贯性方面表现突出,同时避免了伪影的产生,并有效恢复了因过度平滑而损失的细节。为了全面评估 TGA-TFNet 在秦腔视频去噪方面的表现,本节将此方法与其他去噪算法在多个数据集上进行了对比分析。实验涵盖了详尽的实验设置、数据集的选择、评估指标的确定、对比实验的执行以及消融实验的探究。通过这些严谨的实验流程,能够确保秦腔视频中的艺术性和细节得到最大程度的保留,同时去除噪声,使得秦腔的韵味和艺术表现力在数字化处理中得以传承和发展。

3.1 实验设置

本实验使用 Python 3.8 和 PyTorch 1.7,以及 NVIDIA 3090 显卡进行训练。在实验中,利用 DAVIS 训练数据集,该数据集中添加了标准差为 [5,50] 的高斯噪声,并构建了带噪干净对,作为网络模型的训练数据。于每个训练周期中,提取了 128 000 个训练样本进行训练,每个训练样本是一个 96×96 的 7 帧序列块,批处理大小设置为 32。在训练过程中,使用 Adam 优化器对网络参数进行优化。网络模型总共进行了 35 个周期的训练。前 15 个周期的学习率设置为 10^{-3} ,接下来的 10 个周期学习率变为 10^{-4} ,最后剩余的 10 个周期的学习率设置为 10^{-6} 。

3.2 数据集

深度学习任务需要大量的数据样本来对模型进行训练。为了获取丰富的样本,将使用添加了高斯噪声的 DAVIS^[20]数据集作为实验的训练集,用于对模型进行训练。同时,为了验证所提出的视频去噪方法的有效性,将使用添加了不同强度噪声的 DAVIS 测试数据集以及秦腔戏曲数据集对模型进行测试。接下来,将介绍 DAVIS 数据集和自建数据集的详细情况。

DAVIS 数据集是专为视频分析和理解而设计的重要资源,在计算机视觉研究领域得到广泛应用。该数据集旨在提供一个标准化的平台,用于评估和比较不同视觉算法在处理动态场景中的性能。DAVIS 数据集提供了一系列高质量的视频,每个视频都附带了精细的像素级手工标注,描述了视频中的主要对象和场景。数据集的设计考虑了多样性和复杂性,包含了从简单到复杂的各种场景,以模拟计算机视觉系统在现实世界中可能面临的挑战。自首次发布以来,DAVIS 数据集

经历了多次更新和扩展,每个新版本在视频数量、场景复杂度和标注质量方面都有所提升。例如,数据集从最初的单对象跟踪和分割扩展到了多对象场景,以适应计算机视觉领域的发展需求。DAVIS 数据集的一个显著特点是其高质量的手工标注,为研究人员提供了一个准确的基准来评估他们的算法。这些标注覆盖了视频的每一帧,为各种视觉任务(如对象检测、跟踪、分割等)提供了实验基础。由于其丰富的场景和精确的标注,DAVIS 数据集适用于广泛的计算机视觉任务,包括但不限于视频分割、对象跟踪、场景理解和动作识别。它为算法开发者和研究人员提供了一个共同的基准,以验证和比较他们的技术在真实世界条件下的性能。

秦腔戏曲视频数据集是通过收集现场演出的视频、访问数字档案和图书馆资源整理而成的,该数据集包含了 15 个秦腔戏曲视频。

3.3 消融实验

为了验证本文方法的有效性,本小节进行了消融实验,分别针对时序融合模块、双门控注意力模块以及多头交互注意力精炼模块进行了实验。接下来将详细介绍这些实验以及实验结果。

3.3.1 TFM 的消融实验

时序信息的充分利用可聚焦于视频帧之间高度相关的内容。相较于单层 3×3 卷积,TFM 能更充分地利用时序上的连续性和冗余信息,以提升去噪性能。为了验证 TFM 的有效性,在 DAVIS

数据集上分别测试了以 TFM 作为融合模块的去噪网络和以单层 3×3 卷积作为融合模块的去噪网络,相关实验结果如表 1 所示。从实验结果中可以明显观察到,TFM 作为融合模块的去噪网络相较于单层 3×3 卷积,更加专注于视频帧之间高度相关的内容,从而实现更佳的去噪性能。

3.3.2 TGAM 的消融实验

为了验证 TGAM 的有效性,在 DAVIS 数据集上添加了 $\sigma = 50$ 的高斯噪声,并进行了相应的实验。为了避免冗余卷积,引入了可学习的 $H \times W \times C$ 矩阵 ω ,用以替代传统的卷积操作。在相同参数数量的情况下,这一改进将 PSNR 从 31.09 dB 提高到 31.94 dB,实验结果如表 2 所示。此外,当在频域中捕获两次远程依赖关系时,与仅捕获一次依赖关系相比,在相同参数数量下 PSNR 进一步提高了 0.46 dB,表明双重相互作用在捕获远程依赖关系方面具有优势。此外,在计算速度方面,由于通过学习矩阵 ω 对频域中的依赖性进行建模,成功捕获两次远程依赖关系,将计算时间从 169 ms 缩短到 143 ms。

同时,为了研究通道数对性能的影响,对两个超参数 r_1 和 r_2 进行了调优。实验结果如表 3 所示,当 $r_1 = 0.5, r_2 = 1$ 时,PSNR 达到最佳。而当 $r_1 = 1$ 和 $r_2 = 1$ 时,PSNR 下降了 0.51 dB,说明增加信道数会降低性能,因为不同信道上的特征存在冗余。

表 1 不同融合方式的实验结果
Tab. 1 Experimental results of different fusion methods

融合模块	$\sigma = 10$	$\sigma = 20$	$\sigma = 30$	$\sigma = 40$	$\sigma = 50$
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
TFM	39.00/0.966 8	35.93/0.940 0	34.23/0.914 7	32.97/0.893 2	31.94/0.874 9
单层 3×3 conv	8.67/0.957 6	35.53/0.932 4	33.97/0.902 3	32.59/0.882 1	31.67/0.862 7

表 2 不同捕获方式的实验结果
Tab. 2 Experimental results of different capture methods

捕获方式	一次捕获	两次捕获	PSNR	Time/ms
Conv	—	✓	31.09	169
ω	✓	—	31.38	103
ω	—	✓	31.94	143

表 3 超参数 r_1 与 r_2 的实验结果
Tab. 3 Experimental results of hyperparameters r_1 and r_2

r_1	r_2	PSNR	Time/ms
0.5	0.5	31.68	121
0.5	1	31.94	143
1	1	31.43	179

3.3.3 MIAM 的消融实验

通过对去噪后的图像进行细化处理,可以恢复因去噪而被去除的精细细节和纹理。为了验证 MIAM 的有效性,在 DAVIS 数据集上添加了 $\sigma = 50$ 的高斯噪声,并进行了相应的实验。实验结果如表 4 所示,在引入 MIAM 后,PSNR 从 31.66 dB 提高到 31.94 dB,SSIM 从 0.866 1 提高到了 0.874 9。

表 4 MIAM 的消融实验结果

Tab.4 Ablation experimental results of MIAM		
MIAM	PSNR/dB	SSIM
有	31.94	0.874 9
无	31.66	0.866 1

3.4 对比实验

本小节将对 TGA-TFNet 视频去噪算法进行了全面的评估。实验结果表明,TGA-TFNet 在 DAVIS 测试集上对不同强度的高斯噪声具有卓越的去噪性能。如表 5 所示,TGA-TFNet 在 PSNR 和 SSIM 这两个关键客观评价指标上均优于 DVDNet^[7]、FastDVDNet^[11] 和 ViDeNN^[21] 方法。

此外,为了直观展示 TGA-TFNet 的性能,随机选取了 DAVIS 数据集中两个视频序列,并展示了它们经过不同去噪方法处理后的结果(见图 5)。从定性的角度分析,相较于其他去噪方法,TGA-TFNet 在保留细节信息和防止图像过度平滑化方面的优越性。

为了验证 TGA-TFNet 在特定领域如秦腔戏曲视频数据集上的应用效果,对其进行了测试,测试视频被添加了 $\sigma = 50$ 的高斯噪声。测试结果

表 5 不同方法在 DAVIS 数据集上的结果

Tab.5 Results of different methods on the DAVIS dataset					
噪声强度	评价指标	DVDNet ^[7]	ViDeNN ^[21]	FastDVDNet ^[11]	Ours
10	PSNR	38.45	37.27	38.85	39.00
	SSIM	0.965 7	0.956 7	0.966 4	0.966 8
20	PSNR	35.89	34.73	35.76	35.93
	SSIM	0.937 2	0.925 7	0.938 9	0.940 0
30	PSNR	34.20	33.07	33.99	34.23
	SSIM	0.914 1	0.896 4	0.914 2	0.914 7
40	PSNR	32.94	31.86	32.74	32.97
	SSIM	0.892 7	0.869 7	0.891 9	0.893 2
50	PSNR	31.92	30.90	31.78	31.94
	SSIM	0.874 5	0.845 3	0.871 5	0.874 9

如图 6 所示,TGA-TFNet 在秦腔视频的去噪结果上,不仅视觉上更为舒适,而且在色彩保真度上也有出色表现。即便在复杂光照条件下,TGA-TFNet 去噪后的秦腔视频依然能够保持色彩的鲜艳度和亮度的一致性,没有出现色彩偏移或明显的亮度变化。

3.5 实验结果与分析

本文旨在解决现有视频去噪算法忽略运动细节和时序连贯性等时序信息的问题,并提出了一种基于门控注意力的视频去噪算法(TGA-TFNet)。该方法主要由时序融合模块(TFM)、双门控注意力去噪网络(TGADNet)和多头交互注意力精炼模块(MIAM)组成,本节对这 3 个部分的结构和原理进行了详细的阐述。为了验证 TGA-TFNet 方法的有效性,本节对 TGA-TFNet 方法在 DAVIS 数据集和秦腔戏曲数据集上进行了实验对比。实验结果表明,TGA-TFNet 在这些数据集上取得了出色的去噪性能。此外,本节还通过消融实验验证了 TFM 模块、TGADNet 模块和 MIAM 模块对 TGA-TFNet 方法性能提升的重要作用。

本研究考虑了图像增强技术和视频修复技术在处理秦腔视频时可能产生的效果,这些技术虽然在提升视频视觉效果方面具有潜力,但它们与去噪算法的主要区别在于,它们更侧重于改善视频的视觉表现或修复物理损伤,而不是减少噪声。在处理秦腔视频时,这些技术需要与去噪算法协同工作,以确保在提升视频质量的同时,不破坏秦腔表演的艺术性和传统特色。

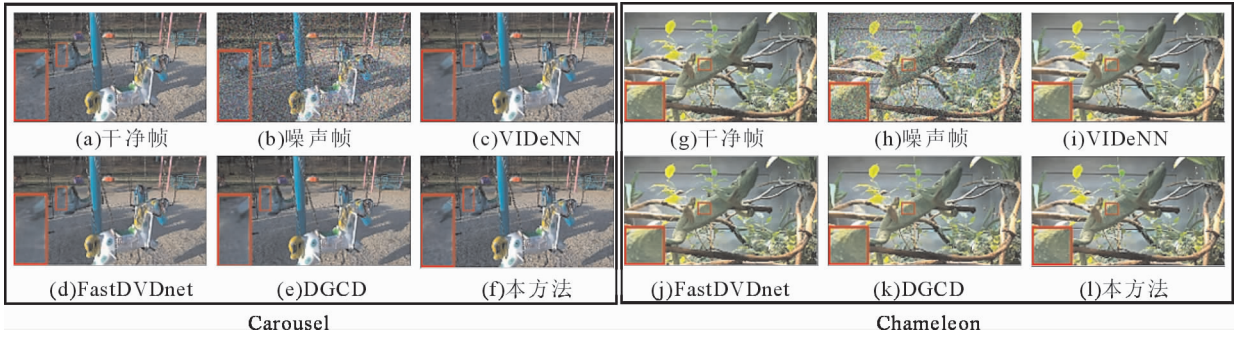


图 5 不同去噪方法在 Carousel 和 Chameleon 视频序列上的去噪结果

Fig. 5 Denoising results of different denoising methods on the Carousel and Chameleon video sequences

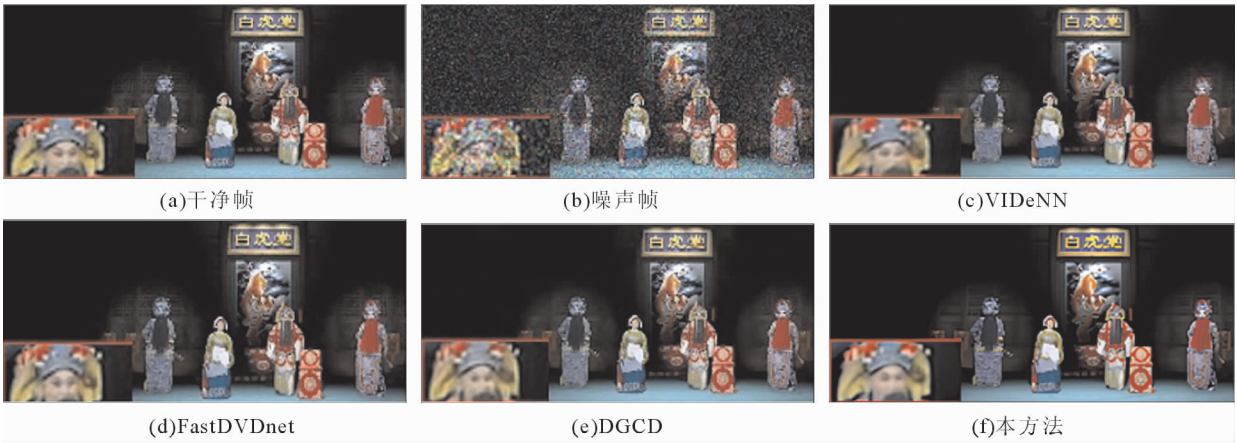


图 6 不同去噪方法在秦腔戏曲视频序列上的去噪结果

Fig. 6 Denoising results of different denoising methods on the Qin Opera video sequence

4 结语

在深入研究视频去噪技术的基础上,针对秦腔视频这一特定类型的视频材料,本文提出了改进和优化方法。秦腔作为中国西北地区广为流传的传统戏曲艺术形式,其视频资料的清晰度和流畅度对于传承和推广这一非物质文化遗产至关重要。因此,本文提出了一种基于双门控注意力的时序融合视频去噪算法(TGA-TFDNet),特别适用于秦腔视频的去噪处理。该算法利用帧与帧之间的相关性,避免了重复计算,在提高了去噪效率的同时,又保留了去噪结果中的重要细节,这对于秦腔表演中精细的面部表情和复杂的服饰纹理的保留尤为重要。通过采用 TFM 来利用时序上的连续性和冗余信息,以提升去噪性能,同时引入了 TGADNet 以实现双重门控机制,不仅能识别和调节对空间噪声特征的关注,而且能够通过 MIAM 利用帧与帧之间的注意力交互来消除去噪过程中产生的伪影,并恢复因过渡平滑而损失的细节,从而生成高质量的去噪视频帧序列。实验结果表

明,相较于目前先进的视频去噪算法,该方法在去噪质量和计算效率上具有强大的竞争力。然而,在极端噪声或低光照条件下,视频序列中的有效信息会大大减少,使得去噪过程更加困难。因此,未来将改进现有的双门控注意力机制,以提高算法在极端噪声条件下的鲁棒性。同时,还计划扩充秦腔视频数据集,并继续进行实验,以更充分验证该方法在秦腔视频去噪任务上的有效性。

参考文献

[1] 周明全,耿国华,武仲科. 文化遗产数字化保护技术及应用[M]. 北京:高等教育出版社,2011.

[2] 耿国华,何雪磊,王美丽,等. 文化遗产活化关键技术研究进展[J]. 中国图象图形学报,2022,27(6): 1988-2007.

GENG G H, HE X L, WANG M L, et al. Research progress on key technologies of cultural heritage activation[J]. Journal of Image and Graphics, 2022, 27(6): 1988-2007.

[3] TICO M. Multi-frame image denoising and stabilization[C]//2008 16th European Signal Processing Conference. Lausanne: IEEE, 2008: 1-4.

- [4] LIU X, ZHU Y, LIU L, et al. Feature-suppressed contrast for self-supervised food pre-training[C]//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM, 2023: 4359-4367.
- [5] MAGGIONI M, HUANG Y, LI C, et al. Efficient multi-stage video denoising with recurrent spatio-temporal fusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE, 2021: 3466-3475.
- [6] DAVY A, EHRET T, MOREL J M, et al. A non-local CNN for video denoising[C]//2019 IEEE International Conference on Image Processing (ICIP). Taipei: IEEE, 2019: 2409-2413.
- [7] TASSANO M, DELON J, VEIT T. Dvdnet: A fast network for deep video denoising[C]//2019 IEEE International Conference on Image Processing (ICIP). Taipei: IEEE, 2019: 1805-1809.
- [8] XUE T, CHEN B, WU J, et al. Video enhancement with task-oriented flow[J]. International Journal of Computer Vision, 2019, 127(8): 1106-1125.
- [9] WANG C, ZHOU S K, CHENG Z W. First image then video: A two-stage network for spatiotemporal video denoising[EB/OL]. (2020-01-22) [2024-06-20]. <https://arxiv.org/abs/2001.00346v2>.
- [10] MEHTA S, KUMAR A, REDA F, et al. Evrnet: Efficient video restoration on edge devices[C]//Proceedings of the 29th ACM International Conference on Multimedia. Virtual: ACM, 2021: 983-992.
- [11] TASSANO M, DELON J, VEIT T. Fastdvdnet: Towards real-time deep video denoising without flow estimation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 1354-1363.
- [12] VAKSMAN G, ELAD M, MILANFAR P. Patch craft: Video denoising by deep modeling and patch matching[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 2157-2166.
- [13] MAGGIONI M, HUANG Y, LI C, et al. Efficient multi-stage video denoising with recurrent spatio-temporal fusion[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3466-3475.
- [14] SONG M, ZHANG Y, AYDIN T O. Tempformer: Temporally consistent transformer for video denoising[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 481-496.
- [15] LI D, SHI X, ZHANG Y, et al. A simple baseline for video restoration with grouped spatial-temporal shift[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 9822-9832.
- [16] LIANG J, CAO J, FAN Y, et al. Vrt: A video restoration transformer[J]. IEEE Transactions on Image Processing, 2024, 33: 2171-2182.
- [17] ZAMIR S W, ARORA A, KHAN S, et al. Multi-stage progressive image restoration[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 14821-14831.
- [18] COOLEY J W, TUKEY J W. An algorithm for the machine calculation of complex Fourier series[J]. Mathematics of Computation, 1965, 19(90): 297-301.
- [19] LECUN Y, BOSER B, DENKER J S, et al. Back-propagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.
- [20] PONT-TUSET J, PERAZZI F, CAELLES S, et al. The 2017 DAVIS challenge on video object segmentation[EB/OL]. (2018-03-01) [2024-06-20]. <https://arxiv.org/abs/1704.00675v3>.
- [21] CLAUS M, VAN GEMERT J. Videnn: Deep blind video denoising[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach: IEEE, 2019: 1843-1852.

(编辑 张欢)