智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于光流和多尺度特征融合的视频去噪算法

孙立辉, 陈恒, 商月平

引用本文:
孙立辉, 陈恒, 商月平. 基于光流和多尺度特征融合的视频去噪算法[J]. 智能系统学报, 2024, 19(6): 1593-1603.
SUN Lihui, CHEN Heng, SHANG Yueping. Video denoising based on optical flow and multi-scale features[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(6): 1593-1603.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202306002

DOI: 10.11992/tis.202306002
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20240709.1127.013

基于光流和多尺度特征融合的视频去噪算法

孙立辉[1], 陈恒[1], 商月平[2]
(1. 河北经贸大学 信息技术学院, 河北 石家庄 050061; 2. 河北经贸大学 数学与统计学学院, 河北 石家庄 050061)

摘 要: 为有效地去除视频当中的噪声, 减少纹理细节丢失, 提出了一种基于光流和多尺度特征融合的级联视频去噪算法。通过分组策略对序列帧进行精准对齐, 然后送入集成残差细化和可选择性跳跃连接的多尺度架构, 实现细节特征的精确保留与高效融合, 进而采用非局部注意力机制以深入挖掘视频帧的时空特征, 重建高质量视频。同时为保留更多纹理细节, 提出一种联合感知损失的目标函数监督训练。实验结果表明, 所提算法的去噪结果可以保留更多的纹理特征, 更符合人眼视觉的习惯。该算法在强噪声下具备鲁棒性高、计算量小的特点, 可以满足实时去噪的要求。

关键词: 多帧去噪; 视频去噪; 光流对齐; 感知损失; 非局部注意力; 图像处理; 计算机视觉; 深度学习
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2024)06−1593−11

中文引用格式: 孙立辉, 陈恒, 商月平. 基于光流和多尺度特征融合的视频去噪算法 [J]. 智能系统学报, 2024, 19(6): 1593–1603.
英文引用格式: SUN Lihui, CHEN Heng, SHANG Yueping. Video denoising based on optical flow and multi-scale features[J]. CAAI transactions on intelligent systems, 2024, 19(6): 1593–1603.

Video denoising based on optical flow and multi-scale features

SUN Lihui[1], CHEN Heng[1], SHANG Yueping[2]
(1. School of Information Technology, Hebei University of Economics and Business, Shijiazhuang 050061, China; 2. College of Mathematics and Statistics, Hebei University of Economics and Business, Shijiazhuang 050061, China)

Abstract: To effectively eliminate noise from videos while preserving texture details, a cascade video denoising algorithm that integrates optical flow and multi-scale features is proposed. The process begins by accurately aligning sequence frames using a grouping strategy. These frames are then processed through a multi-scale architecture that combines residual refinement and selective skip connection. This approach not only preserves detailed features but also enhances alignment and fusion. Furthermore, a non-local attention mechanism is employed to deeply mine spatiotemporal features, enabling the reconstruction of high-quality videos. To preserve detailed textures, a target function supervision training method that combines perceptual loss is proposed. Experimental results show that the proposed algorithm retains more texture features and aligns well with human visual perception. It is also highly robust, has low computational complexity under strong noise, and meets real-time denoising requirements.

Keywords: multi-frame noise reduction; video denoising; optical flow alignment; perceptual loss; non-local attention; image processing; computer vision; deep learning

由于传感器内部电路、拍摄环境等内外界的因素影响[1], 拍摄的视频不可避免地会引入噪声, 降低采集质量, 弱化人们的观感效果, 影响后续检测、识别等视觉任务的处理性能。因此, 为获取高质量的视频, 需要对其进行去噪处理。

视频处理方法与专注于空间特征的图像去噪不同, 它在时间维度上会有冗余信息, 去噪时更多地依赖于对时序特征的精确提取与有效利用[2]。早期传统方法多数将其视为图像去噪的简单拓

收稿日期: 2023−06−01. **网络出版日期**: 2024−07−12.
基金项目: 河北省重点研发计划项目(20350801D).
通信作者: 孙立辉. E-mail: sun-lh@163.com.

展，忽略视频帧的时间相关性，容易出现不断闪烁、伪影和引入新的噪声等问题。

为解决上述问题，最近一系列基于深度学习的去噪算法被提出。根据处理视频帧方式的不同，主要分为 2 种：1）基于序列间非局部特性去噪，2）基于显隐式对齐去噪。前者通过利用序列的自相似性，在时空维度上识别并匹配相似图像块以实现去噪。在处理视频序列时，提升该方法性能需扩展搜索范围，以更精确地捕捉局部特征。为提升帧间序列一致性，后者序列对齐方式被进一步提出，其中，显式对齐依赖于光流法，而隐式对齐则利用可形变卷积或者 U-net 等神经网络学习实现。然而，由于不同帧之间存在运动模糊、遮挡或者光照变化，容易造成光流估计不准确，影响显式对齐的准确性，此外，尽管隐式对齐能够处理大视差问题，但对局部特征的敏感度不足，导致去噪后图像易出现平滑现象，影响纹理细节恢复，降低整体去噪效果。

基于上述分析，本文提出新颖的结合光流和多尺度融合的高效视频去噪算法 (efficient deep video denoising, EDVDNet)。其中，设计运动校正网络用于细化光流向量，网络采用归一化层捕捉光流向量的统计特性，提取精炼的运动信息；同时将可选择跳跃连接和空间残差细化集成到多尺度架构中，长短跳跃连接使得细节特征得以保留，最后通过非局部思想增强模型对时序信息的捕捉能力。实验结果表明该算法可以极大保留图像纹理细节，提升去噪性能的同时支持实时处理。

1　相关工作

1.1　图像去噪

图像去噪一直是研究者长期关注的课题，旨在从噪声图像中恢复出清晰的图像。传统方法受限于手工参数或对先验知识的依赖[3]，而基于深度学习的方法，特别是基于卷积神经网络 (convolutional neural networks, CNN)，通过学习噪声图像和干净图像之间的映射关系，实现新的性能突破。近些年，研究焦点转向提升深度学习模型的表达能力，例如，Zhang 等[4] 和 Guo 等[5] 通过深度卷积和多层网络训练学习空间的噪声分布，输出带有残差的干净图像。最近，为了使模型更具可解释性，Helou 等[6] 借鉴贝叶斯框架提出了一种盲通用图像融合降噪器。Ren 等[7] 使用展开策略设计网络结构，并提出了一种用于图像去噪的双元素注意机制网络。现阶段，许多研究者寻求去噪质量和效率之间的平衡，为进一步提高从真实噪声图像中提取特征的能力做出许多努力。

1.2　视频去噪

传统视频去噪算法多采用扩展和改进图像去噪技术，但是视频和图像本身属于两种不同媒介，视频中丰富的冗余特征对于去噪有着极大的帮助。近些年，随着非局部思想的提出，研究者们开始尝试将非局部补丁的概念引入到卷积神经网络框架。Davy 等[8] 提出专用于为每个图像块寻找相似块的网络层，随后，Vaksman 等[9] 提出 Patch-craft 概念，将每帧分割成独立且重叠的图像块再进行非局部搜索和去噪，充分利用了自然图像和序列的自相似性。当前，视频恢复等任务愈发强调对齐的重要性。随着研究的深入，序列对齐成为一大挑战[10]。对齐操作分化为显式与隐式两大研究路径，其中显式对齐主要通过计算并反向补偿相邻帧间的像素运动信息实现精准对齐。例如 Tassano 等[11] 和 Xue 等[12] 均使用光流对视频帧进行运动估计和补偿，其方法减少了闪烁现象；在视频恢复领域中的 BasicVSR[13]，光流作为双向循环神经网络 (recurrent neural networks, RNN) 传播特征信息的主要架构，也得到了优越的去噪性能，而 Li 等[1] 突破性提出前向循环模块和前瞻性循环模块，提升了光流的处理效率。隐式对齐则侧重于通过卷积网络在不同阶段学习和融合输入序列帧间的运动分量，例如 Tassano 等[14] 和 Xiang 等[15] 利用两阶段 U-net 网络学习噪声的映射，通过残差实现降噪，Wang 等[16] 通过结合金字塔和可形变卷积架构实现帧对齐。这些方法根据对输入序列的处理方式不同，可细分为多帧方法和循环方法[1]，各具特色地提升了视频去噪任务的性能。

上述方法虽取得了显著性能，但仍存在一些问题。一方面，非局部思想的性能提升依赖于先验和参数的设置。另一方面，显式对齐结构易受外界干扰，影响对齐质量；隐式结构则对局部特征不敏感，导致去噪性能受限。鉴于此，本文提出了一种结合光流与多尺度融合的视频去噪算法。设计的校正模块有效地弥补显式对齐的局限性，并将非局部思想融入多尺度架构，旨在精确保留噪声序列帧间的干净细节特征，进而实现高效且鲁棒的视频去噪性能。

2　本文方法

2.1　网络结构

EDVDNet 架构如图 1 所示，集成校正对齐模块 (refine align module, RAM)、多尺度细化模块 (multi-scale refinement module, MRM) 和时空融合

模块 (spatiotemporal fusion module, STFM), 分两阶段执行。基于相邻帧间具有较小的运动变化, 因此, 将输入的相邻 3 帧为 1 组。第 1 阶段, 分组序列经校正对齐模块实现预对齐, 后通过多尺度细化模块进一步优化对齐精度, 输出融合后的中间特征; 第 2 阶段通过基于非局部注意力的时空融合模块扩展时序特征, 捕捉远程依赖关系, 实现深度全局去噪效果。

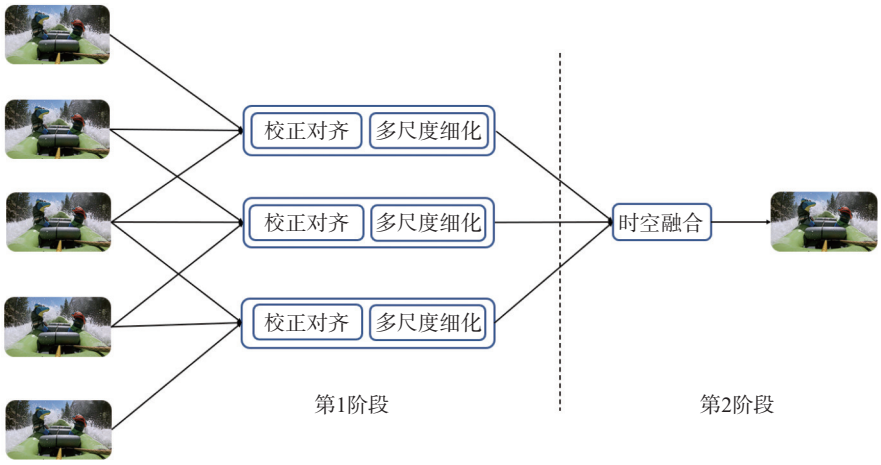
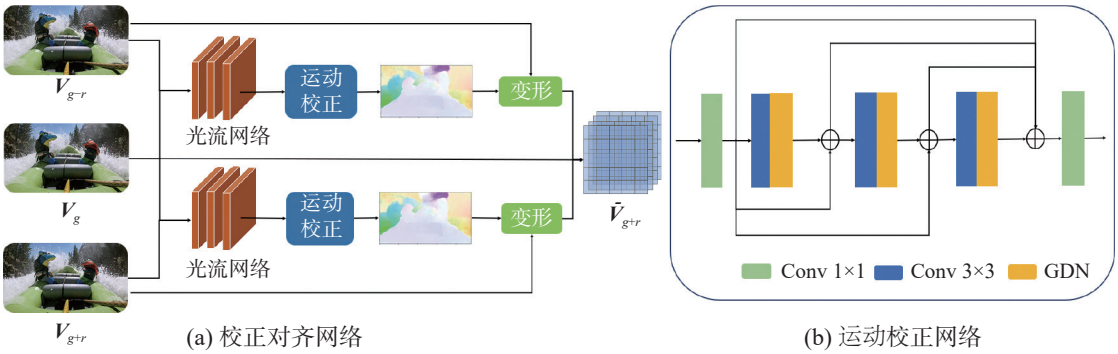


图 1　EDVDnet 架构

Fig. 1　Architecture of EDVDnet

2.2　校正对齐模块

由于视频场景未知复杂, 以往方法多数采用隐式对齐, 导致训练学习时可能过度依赖训练集的规模和质量, 限制了模型在实际环境中的应用能力。因此, 探索显式光流对齐机制成为提升模型泛化能力的关键研究方向, 传统光流对齐法[17]通过估计相邻帧和参考帧的运动向量场, 进而利用这些运动向量的反向运算实现视频帧的变形对齐。基于此, 本文提出如图 2 所示的校正对齐模块。



(a) 校正对齐网络　　(b) 运动校正网络

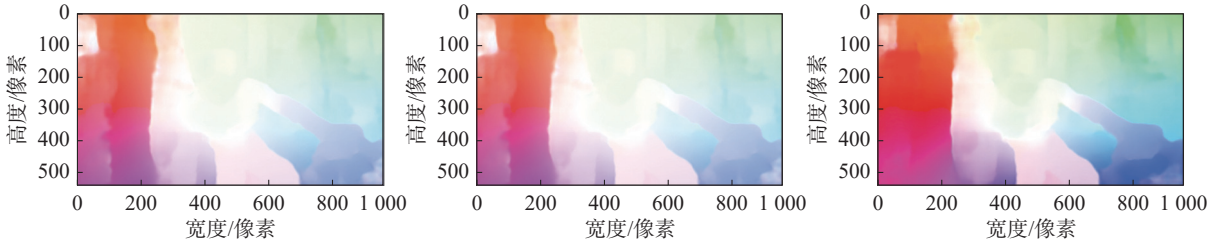
图 2　校正对齐模块（RAM）

Fig. 2　Refine align module(RAM)

传统光流的计算过程可能会受到噪声干扰, 导致估算出的向量准确性降低, 如图 3 所示, 对比清晰帧 (图 3(b)) 与噪声帧的光流结果 (图 3(c)), 噪声明显削弱了物体边缘特征, 对后续的变形工作造成不利影响。为应对这一挑战, 本文在变形前提出了一个运动校正网络, 旨在提升光流法在估计相邻帧与参考帧运动向量时的精度。结构如图 2(b) 所示, 该结构由若干卷积和 GDN[18] (generalized divisive normalization) 实现, 融入残差可以更好重建边缘和细节向量。Ballé等[18]提出的 GDN 原适用于图像重建, 将其应用于光流校正, 能够有效捕捉光流图像的统计特性, 从而提取更加精确的运动向量。图 3(a) 是经过运动校正后输出的光流图, 展现出更高的清晰度和准确性, 更贴近原始光流的真实情况。最后, 通过校正的运动向量实现帧对齐:

$$\overline{\boldsymbol{V}}_{g+r} = W(\boldsymbol{V}_{g+r}, R(F(\boldsymbol{V}_g, \boldsymbol{V}_{g+r}))) \tag{1}$$

式中: g 表示每组序列中参考帧索引, $g \in [t-1, t+1]$, $r \in [-1, 1]$; t 表示 V_{g+r} 的中心帧索引; r 是相邻帧索引; F 表示光流对齐; R 表示运动校正网络; W 表示对齐操作[1], 输出对齐帧 $\overline{\boldsymbol{V}}_{g+r}$ 后送入下一阶段。



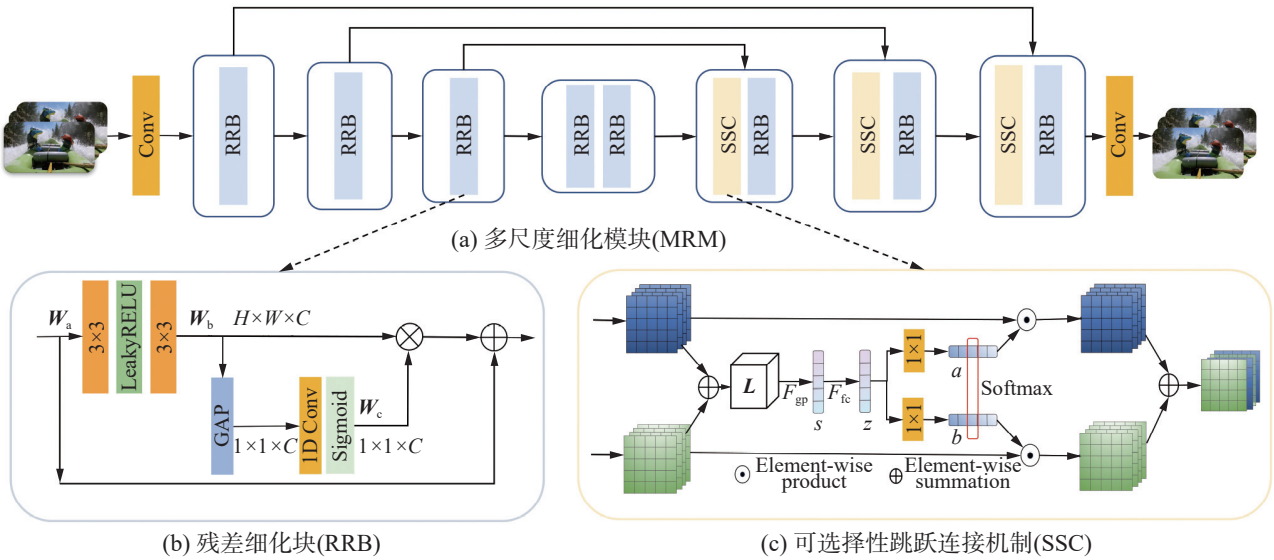
(a) 相应噪声帧经运动校正后的光流　(b) 清晰帧得到的光流　(c) 相应噪声帧得到的光流

图 3　使用运动校正前后的光流对比

Fig. 3　Comparison of optical flow before and after motion refine

2.3　多尺度细化模块

鉴于光流计算中严格的约束条件可能导致对齐后细节处出现模糊和伪影现象，本文构建一种多尺度细化模块 (MRM)，实现对预对齐特征进行空间去噪和再校正。多尺度细化模块结构如图 4 所示。



(a) 多尺度细化模块(MRM)

(b) 残差细化块(RRB)　(c) 可选择性跳跃连接机制(SSC)

图 4　多尺度细化模块（MRM）

Fig. 4　Multi-scale refinement module (MRM)

级联卷积由于其感受野的限制，不足以捕捉序列中细节特征间的相互关系。相对地，多尺度架构能够提供更为综合的特征提取能力。因此，多尺度细化模块基于 U-net 架构，为了补偿光流估计的不确定性并丰富语义信息，对 U-net 网络做出如下优化：

1）在多尺度架构中，针对特征冗余和噪声传播问题，本文依据文献 [19-21] 的理论，在跳跃连接中引入可选择性跳跃连接机制 (selective skip connection, SSC) 架构。如图 4(c) 所示，跳跃连接时，对低层次语义信息 $\boldsymbol{L}_1$ 和高层语义信息经上采样后的特征 $\boldsymbol{L}_2$ 进行选择性整合。这一整合过程通过门控机制实现，得到通道特征 $\boldsymbol{L}$，使每一个分支携带不同信息进入下一层神经元；对通道特征进行低维嵌入获得通道上的全局信息 $\boldsymbol{S}\in\mathbf{R}^{1\times C}$，经全连接操作生成特征图 $\boldsymbol{Z}\in\mathbf{R}^{d\times 1}$，再通过 Softmax 获得通道权重，不同通道可以根据权重引导选择有用特征，聚合成新特征图 $\boldsymbol{V}$ 作为解码器的输入。可选择性机制的公式定义为

$$\boldsymbol{S}=\mathcal{F}_{\mathrm{gp}}(\boldsymbol{L})=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{L}(i,j) \tag{2}$$

$$\boldsymbol{Z}=\mathcal{F}_{\mathrm{fc}}(\boldsymbol{S})=\gamma(\boldsymbol{W}\boldsymbol{S}) \tag{3}$$

$$a_c=\frac{\mathrm{e}^{\boldsymbol{A}_c z}}{\mathrm{e}^{\boldsymbol{A}_c z}+\mathrm{e}^{\boldsymbol{B}_c z}},b_c=\frac{\mathrm{e}^{\boldsymbol{B}_c z}}{\mathrm{e}^{\boldsymbol{A}_c z}+\mathrm{e}^{\boldsymbol{B}_c z}} \tag{4}$$

式中：H 和 W 分别为输入特征的高和宽，F_{gp} 表示全局池化，F_{fc} 表示全连接层，其中 γ 表示激活和归一化操作，$\boldsymbol{W}\in\mathbf{R}^{d\times c}$ 为权重矩阵，矩阵维度将根据输入通道数量实现自适应变化；a_c 和 b_c 对应两路分支 $\boldsymbol{L}_1$ 和 $\boldsymbol{L}_2$ 的注意力值，其中 $\boldsymbol{A}_c$ 代表第 C 行，a_c 代表 a 的第 C 个元素，$\boldsymbol{B}_c$ 和 b_c 同理，最终得到特征图 $\boldsymbol{V}$ 中每个通道表示为

$$\boldsymbol{V}_c=a_c\cdot\boldsymbol{L}_{1_c}+b_c\cdot\boldsymbol{L}_{2_c},\quad a_c+b_c=1 \tag{5}$$

2）在提取和融合阶段，各通道对序列对齐的贡献度各异，本文引入了基于注意力[22] 的残差细

化块 (residual refine block, RRB) 以动态分配学习权重至各通道, 筛选并强化关键序列特征, 从而实现精确对齐。如图 4(a) 所示, 输入特征 $\boldsymbol{W}_{\mathrm{a}}$ 首先经卷积和激活层操作后得到 $\boldsymbol{W}_{\mathrm{b}} \in \mathbf{R}^{H\times W\times C}$, 后经 GAP 操作使每个通道的特征信息被平均化, 通过 1 维卷积生成新的特征 $\boldsymbol{W}_{\mathrm{c}} \in \mathbf{R}^{1\times 1\times C}$, 其中一维卷积核大小 k 决定相邻通道数大小, 随着通道数量增加, k 由输入通道数 C 决定, 最后采用 Sigmoid 函数得到权重特征 $\boldsymbol{W}_{\mathrm{c}}$, 与操作前的特征 $\boldsymbol{W}_{\mathrm{b}}$ 对位相乘, 将重要程度放回原先的通道中。

多尺度细化模块结合了残差短连接与跳跃机制的长连接, 有利于参数在训练过程中分布得更加均匀[23], 同时残差连接有助于整体架构捕获序列细节, 抑制噪声传播。

2.4　时空融合模块

第 1 阶段尽管实现了精确对齐和初步降噪, 但尚未充分利用多帧之间的时序特征。受文献 [12] 的启发, 提出时空融合模块, 旨在通过整合多组序列的中间特征, 扩展时间维度上的感受野, 以增强模型对时序信息的捕捉能力。时空融合模块与多尺度细化模块相似, 但是本文区别在于, 将残差细化块替换为非局部残差融合块 (non-local residual fusion block, NRFB)。

依据图像和视频序列的自相似性原理, 同时受非局部和视觉处理任务结合[8,19-21] 的思想, 非局部残差融合结构如图 5 所示。首先, 特征图 $\boldsymbol{F}_{\mathrm{a}}$ 经卷积和 LeakyRelu 提取特征后生成 $\boldsymbol{F}_{\mathrm{b}} \in \mathbf{R}^{H\times W\times C}$, 随后依次经 1×1 卷积、变形和 Softmax 操作, 得到注意力权值 $\boldsymbol{F}_{\mathrm{c}} \in \mathbf{R}^{1\times 1\times HW}$; 然后改变 $\boldsymbol{F}_{\mathrm{b}}$ 维度得到新的 $\boldsymbol{F}_{\mathrm{b}} \in \mathbf{R}^{1\times HW\times C}$, $\boldsymbol{F}_{\mathrm{b}}$ 和 $\boldsymbol{F}_{\mathrm{c}}$ 相乘聚合所有位置的特征, 获得全局上下文特征 $\boldsymbol{F}_{\mathrm{d}} \in \mathbf{R}^{1\times 1\times C}$; 紧接着采用 1×1 卷积和激活实现特征转换得到通道级的相关性, 最后通过像素级相加将全局特征聚合到每个位置上, 同时引入残差连接保留更多视频帧细节输出 $\boldsymbol{F}_{\mathrm{e}}$。与第 1 阶段相比, 非局部残差融合块可增强全局特征的建模能力, 并有效捕捉帧间的长距离依赖, 从而优化帧间的细节特征表达。

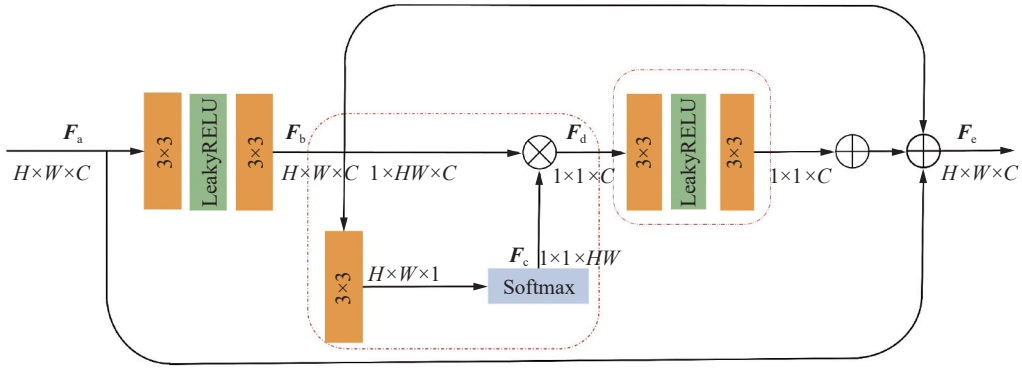


图 5　非局部残差融合块 (NRFB)

Fig. 5　Non-local residual fusion block (NRFB)

2.5　损失函数

目前基于深度学习的视频去噪算法中通常采用全 MAE (mean absolute error) 或者全 MSE (mean squared error) 作为损失函数, 然而该函数和客观评价中的指标存在一定的数学转换关系[24], 可能导致数值虚高。同时峰值信噪比作为逐像素比较的计算方法, 易导致细节丢失, 造成视觉上的平滑和模糊。为解决该问题, 本文采用感知损失[25] 和 MSE 损失双重监督策略, 利用感知损失的特征比较, 更贴近人类视觉感知。损失函数 L 公式为

$$L = \delta L_{\mathrm{perceptual}} + L_{\mathrm{MSE}} \tag{6}$$

式中: $L_{\mathrm{perceptual}}$ 表示感知损失; L_{MSE} 表示均方差损失; δ 为权重系数, 基于经验 δ 设定为 0.1。

感知损失函数利用预训练的 VGG16[26] 网络提取数据特征, 基于高层语义计算特征的 MSE 损失, 以指导网络训练。本文采用该方法, 通过只计算前 14 层提取的高层语义特征, 其中每层语义特征为

$$l_{\mathrm{feat}}^{\phi,j}(\hat{\boldsymbol{y}},\boldsymbol{y}) = \frac{1}{C_j H_j W_j}\left|\phi_j(\hat{\boldsymbol{y}}) - \phi_j(\boldsymbol{y})\right|_2^2 \tag{7}$$

式中: j 为 VGG16 网络的第 j 层, ϕ_j 为第 j 层卷积计算得到的特征图, $\boldsymbol{y}$ 和 $\hat{\boldsymbol{y}}$ 分别为原始输入和网络输出的特征, 特征图的尺寸分别为 C_j、H_j、W_j。

3　实验结果与分析

3.1　数据集

真实环境中噪声来源复杂, 由多种噪声复合形成, 根据概率分布中随机变量原则, 高斯函数是模拟真实噪声最有效的方法之一, 因此, 本文通过高斯噪声来合成实验数据, 以便在受控的噪

声水平下进行比较分析。

本文从公开 DAVIS2017[27] 数据集中随机选择 80 个视频序列，每段视频约 50~100 帧，引入具有不同方差的高斯噪声以构建训练集。剩余 10 个视频序列作为验证集，用于观察模型拟合效果。测试集使用 DAVIS 测试集和 Set8 数据集；验证集和训练集均使用相同加噪方法，使用多个测试数据以防止过拟合现象。实验设置 4 组不同方差 σ=10,20,40,50，以模拟不同程度的噪声条件。

3.2　实验设置

本文通过 PyTorch 实现，实验采用的 CPU 为 Intel(R) Xeon(R) Silver 4210R, GPU 为 NVIDIA GeForce RTX2080SUPER。在训练中，图像块的分辨率设定为 96 像素×96 像素，通过从连续 5 帧视频序列中提取相同位置的图像块来构建数据集。为增强模型泛化性，采用了包括翻转和旋转在内的数据增强技术，使训练集达到 256 000 个图像块。实验中，光流计算使用预训练模型，其他网络层则采用 Kaiming[28] 初始化。同时对几个参数进行训练验证，确定如下设定值：批次 batch 为 48，训练 60 个 epoch，初始学习率设定为 10^{-4}，逐步衰减到 10^{-6}，其他网络使用默认参数的 Adam 算法[29] 优化。

本文将采用 2 种视频恢复中常用的指标即峰值信噪比 (peak signal-to-noise ratio, PSNR) 和结构相似性 (structured similarity, SSIM) 作为定量标准，其中 SSIM 取值范围为 0~1，两项指标越大说明其图像质量越高，与原图更加接近。为验证本方法的优越性，对所提算法与目前先进的算法进行对比，分别为：基于自相似原理中具有代表性的 PaCNet[9] 和 VNLnet[8]、基于光流对齐中的 DVDnet[11] 和 ToFlow[12]、基于隐式对齐的 FastDVDnet[14] 和 ReMoNet[15]6 种方法。

3.3　实验结果与分析

本方法和 6 种对比方法均在 DAVIS 测试集和 Set8 数据集上进行推理。其中，FastDVDnet 和 DVDnet 均采用默认参数在同一数据集上进行训练；ToFlow 输入序列统一为 5 帧，根据默认参数训练；VNLnet 和 PaCNet 采用原作者提供的权重进行推理；ReMoNet 数据指标由原论文提供。

表 1 对比了 DAVIS 测试集上 6 种算法与本研究方法在 4 种噪声水平下的 PSNR 和 SSIM 性能。结果表明，本方法在 2 个评价指标上均优于 FastDVDnet 基线，平均 PSNR 提升约 0.8 dB，SSIM 提升约 0.05。与最新基线 PaCNet 相比，本方法表现相当，定性指标相差无几。特别是当噪声强度较高时（大于 30 dB），本方法的增益更为显著，这表明本方法在处理高噪声序列时更具有明显的优势。此外，PaCNet 在低噪声条件下评价指标表现更好与其在损失函数设计上的考量和参数搜索范围的广泛性有关。

表 1　在 DAVIS 测试集上推理的客观评价指标

Table 1　Objective evaluation indicators for reasoning on DAVIS test set

噪声水平/dB	指标	ToFlow[12]	VNLnet[8]	DVDnet[11]	FastDVDnet[14]	ReMoNet[15]	PaCNet[9]	EDVDnet(本文)
10	PSNR/dB	35.71	35.53	37.94	38.70	38.97	**39.64**	39.35
	SSIM	0.9194	0.9148	0.9370	0.9531	0.9672	**0.971**	0.9697
20	PSNR/dB	33.46	34.49	35.2	35.57	35.77	**36.31**	36.27
	SSIM	0.9002	0.9105	0.9132	0.9169	0.9380	0.9352	**0.9438**
40	PSNR/dB	29.89	32.32	32.41	32.51	32.64	**33.32**	**33.32**
	SSIM	0.8123	0.8698	0.8712	0.8886	0.8872	0.8979	**0.9029**
50	PSNR/dB	28.51	31.43	31.45	31.48	31.65	32.20	**32.33**
	SSIM	0.7583	0.8169	0.8211	0.8365	0.8651	0.8743	**0.8837**

注：加粗代表最优，下划线表示次优。

为验证本方法在视觉细节上的去噪效果，选取了方差 σ 为 20、40 的噪声图像进行可视化分析，并对局部细节进行放大，如图 6、7 所示。结果表明，在图 6 中，尽管 ToFlow 和 DVDNet 利用对齐技术，但它们未能充分利用序列特征，导致去噪后的图像序列出现模糊。VNLNet 在去除字母边缘的噪声时仍有残留。相比之下，FastDVDnet 虽然取得了较好的去噪效果，但其隐式架构对细节特征捕捉不足，导致去噪后的图像仍有模糊。PaCNet 虽然通过精确搜索相似区域获得了优异的定量指标，但其结果在视觉上边缘过于平滑，颜色偏亮。而本方法结合校正后的光流和多尺度

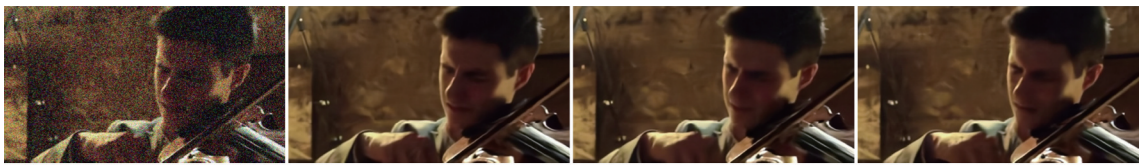
细化策略, 显著减少了噪声对细节特征的负面影响, 从而在运动物体的局部特征还原上更为清晰。在低照度环境下, 噪声强度显著增加, 导致传统自相似特性和对齐技术难以区分噪声和细节, 如图 7 所示, 去噪方法均丢失了背景纹理和人物手部特征, 去噪结果过于平滑。FastDVDnet 在土色背景的纹理恢复上表现不佳, 而 PaCNet 虽然改善了这一问题, 但未能精确恢复人物手部的边缘细节, 且处理时间较长。相比之下, 本研究提出的复合对齐和非局部自相似方法, 能快速准确地搜索并恢复图像块, 有效恢复人物细节和背景纹理特征。

(a) 噪声图　(b) ToFlow　(c) VNLNet　(d) DVDNet

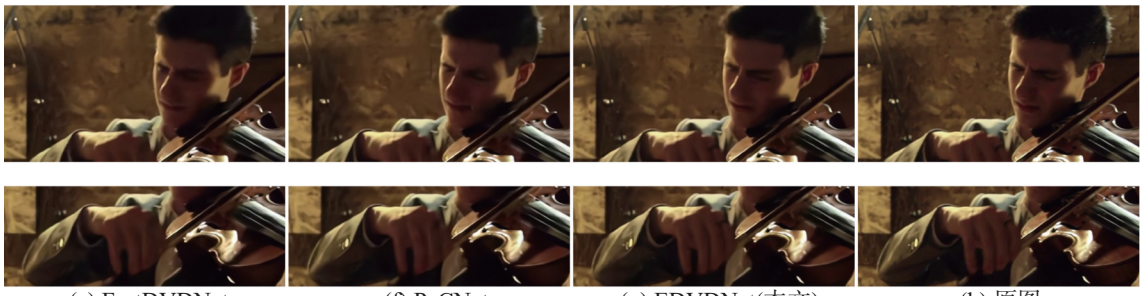
(e) FastDVDNet　(f) PaCNet　(g) EDVDNet(本文)　(h) 原图

图 6　在 DAVIS 测试集上主观视觉效果（σ=20）

Fig. 6　Subjective visual effect on DAVIS test set (σ=20)



(a) 噪声图　(b) ToFlow　(c) VNLNet　(d) DVDNet



(e) FastDVDNet　(f) PaCNet　(g) EDVDNet(本文)　(h) 原图

图 7　在 DAVIS 测试集上主观视觉效果（σ=40）

Fig. 7　Subjective visual effect on DAVIS test set (σ=40)

表 2 给出了 Set8 数据集上各模型的客观评价指标。结果显示, 本方法在 PSNR 上比 FastDVDnet 高出约 0.5 dB, 在 SSIM 上取得了约 0.03 的提升, 整体性能超越了现有先进算法, 同时运行时间显著减少。与最佳基线 PaCNet 相比, 在弱噪声条件下两者表现相当, 但在强噪声条件下本方法表现出更好的稳健性和鲁棒性。

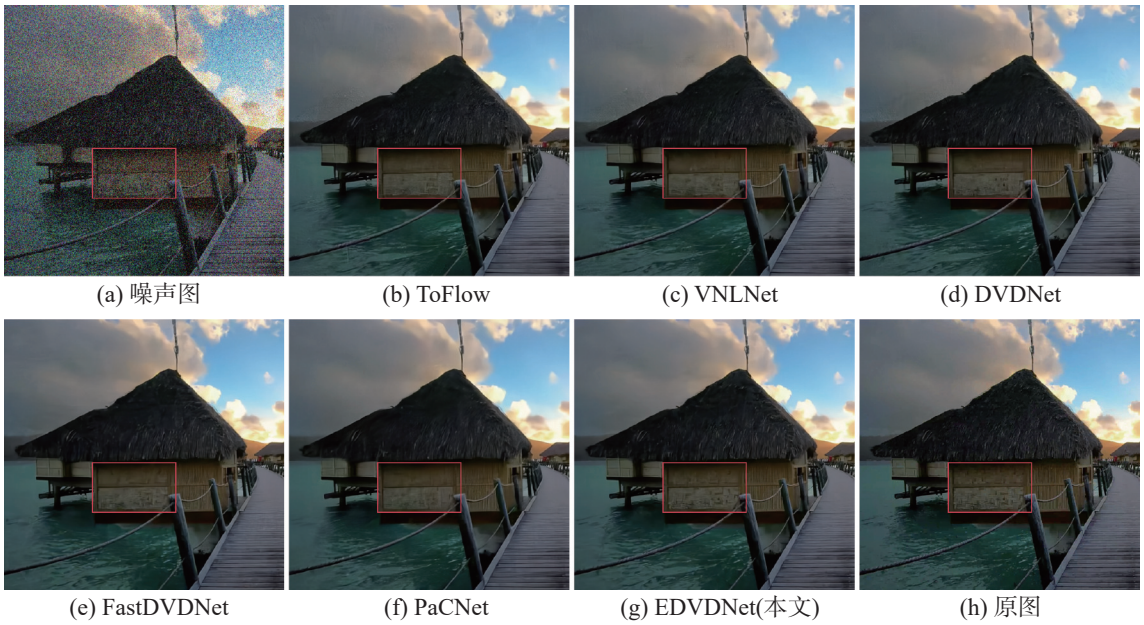
为验证本方法在视觉细节上的去噪效果, 选取噪声方差 σ 为 40 时部分图像进行可视化。如图 8 所示, 云朵区域本身缺乏复杂纹理, 加之摄像机快速移动, 导致去噪困难。VNLNet 和 DVDnet 在左上角云朵和房屋边缘的去噪结果中残留了噪声和伪影, 这可能是由于对齐不精确导致边缘细节的丢失。FastDVDnet 同样未能清晰呈现云朵和房屋边缘。相比之下, PaCNet 虽然改善了边缘特征, 但在红色区域的草屋编织墙处产生了平滑效果, 这是由于算法在处理具有显著运动变化的序列时, 难以准确搜索到相似区域。本方法利用运动校正网络和细化模块, 结合多尺度架构中的长短跳跃连接, 有效恢复了边缘和纹理细节。结合表 2 和图 8, 本算法不仅在其他数据集上展现了良好泛化能力, 而且在细节恢复上也更为清晰。

表 2　在 Set8 数据集上推理的客观评价指标

Table 2　Objective evaluation indicators for reasoning on Set8 dataset

噪声水平/dB	指标	ToFlow[12]	VNLnet[8]	DVDnet[11]	FastDVDnet[14]	ReMoNet[15]	PaCNet[9]	EDVDnet(本文)
10	PSNR/dB	34.34	36.54	36.20	36.25	36.29	**37.06**	36.78
	SSIM	0.924 1	0.948 6	0.951 0	0.950 1	0.952 8	**0.960 6**	0.952 9
20	PSNR/dB	31.44	33.43	33.45	33.23	33.34	**33.94**	33.83
	SSIM	0.867 5	0.914 3	0.912 9	0.911 2	0.917 9	**0.924 7**	0.918 6
40	PSNR/dB	28.33	30.35	30.43	30.46	30.37	30.70	**31.03**
	SSIM	0.764 4	0.837 4	0.841 2	0.845 4	0.857 0	0.862 3	**0.876 4**
50	PSNR/dB	27.26	28.52	28.87	29.15	29.44	29.66	**29.87**
	SSIM	0.718 3	0.793 8	0.811 1	0.815 4	0.830 8	0.834 9	**0.845 6**

注：加粗代表最优，下划线代表次优。



(a) 噪声图　(b) ToFlow　(c) VNLNet　(d) DVDNet

(e) FastDVDNet　(f) PaCNet　(g) EDVDNet(本文)　(h) 原图

图 8　在 Set8 数据集上主观视觉效果（σ=20）

Fig. 8　Subjective visual effect on Set8 dataset (σ=20)

3.4　消融实验

为探究 EDVDnet 中各子模块的功能，本文进行了消融研究。实验保持训练参数不变，随机选取 DAVIS 测试集的 10 个序列（DAVIS-10）和 Set8 作为测试基准，每序列分析前 20 帧。

1）校正对齐模块

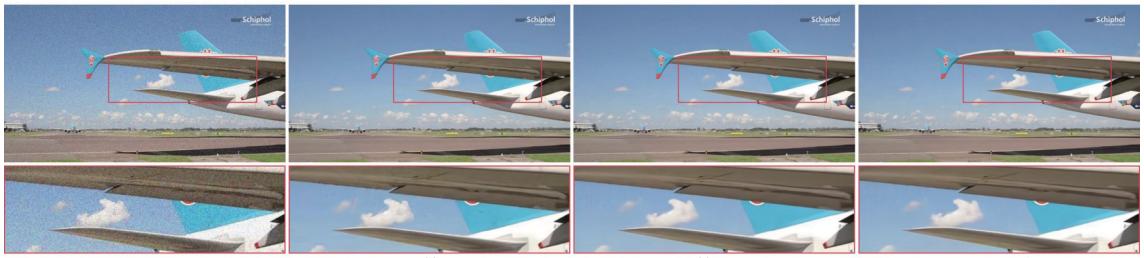
EDVDnet 设计校正对齐模块作为初始阶段，为验证该模块的必要性，以原架构为基准，分别训练未使用和使用校正对齐模块的 2 种网络。如图 9 所示，原始帧 (图 9(d)) 在加入 σ=20 噪声后 (图 9(a))，在不使用校正对齐模块的情况下 (图 9(b))，由于飞机起飞时相邻帧之间的显著位置变化，导致机翼边缘出现严重伪影并残留较多噪声。相比之下，使用校正对齐模块后的输出 (图 9(c)) 伪影显著减少，边缘和纹理细节恢复。这一结果验证了光流对齐模块在预处理序列帧中的有效性，为后续的纹理和细节捕捉与校正提供了便利。

2）多尺度细化模块

为验证多尺度细化模块 (MRM) 的性能，本节以文献 [14] 的去噪块 (denoising block) 作为基准，以连续 3 帧作为输入，经光流对齐后送入该模块，输出中间帧去噪结果，分别测试使用去噪块和使用多尺度细化模块 (本文提出) 的 2 种架构。如表 3 所示，使用本方法获得整体平均 PSNR 指标均高于使用去噪块的指标约 1.2 dB，表明多尺度细化模块所引入的可选择跳跃机制和残差细化结构能够更好地恢复序列特征，同时泛化性能也有所提升。

3）时空融合模块实验

为验证改进多尺度细化模块 (MRM)，提出时空融合模块 (STFM) 的必要性，第 1 阶段网络不变，分别使用 2 种模块作为第 2 阶段架构。



(a) 噪声图　(b) 不使用RAM　(c) 使用RAM　(d) 原图

图 9　使用光流对齐模块前后对比图

Fig. 9　Comparison before and after using FAM

表 3　多尺度细化模块消融实验对比

Table 3　Comparison of MRM ablation experiments　dB

噪声水平σ	方法	DAVIS-10	Set8	平均PSNR
σ=20	Denoising Block	32.10	31.05	31.58
	MRM	**33.32**	**31.96**	**32.64**
σ=40	Denoising Block	30.68	29.16	29.92
	MRM	**31.44**	**30.15**	**30.79**

注: 加粗代表最优。

表 4 的数据显示, 提出时空融合模块作为网络的时空域处理阶段, 平均 PSNR 提升了约 0.5 dB, 这表明在 MRM 中引入 NRFB 进行改进是有效的。进一步的去噪效果分析通过选取测试集中的“skate-jump”序列进行可视化, 原始帧与噪声帧分别如图 10(d) 和 10(a) 所示。MRM 处理的结果 (图 10(b)) 在恢复面部特征时丢失了右眼的细节信息, 并且面部颜色偏暗; 而 STFM 处理的结果 (图 10(c)) 更接近原始图像, 这证实了 NRFB 在通过上下文建模和特征转换捕捉长距离依赖方面的优势, 有助于重建图像的局部细节。与表 3 对比可以发现, 采用 STFM 作为第 2 阶段网络时, 平均 PSNR 提升了约 2 dB。这一增益明确说明, 两阶段网络结构在扩充时间维度特征方面具有显著优势, 从而验证了本方法级联模型的合理性。

表 4　时空融合模块消融实验对比

Table 4　Comparison of STFM ablation experiments　dB

噪声水平σ	方法	DAVIS-10	Set8	平均PSNR
σ=20	MRM	35.30	33.10	34.20
	STRM	**35.72**	**33.83**	**34.77**
σ=40	MRM	31.69	30.73	31.21
	STRM	**32.27**	**31.03**	**31.65**

注: 加粗代表最优。

(a) 噪声图　(b) 使用MRM

(c) 使用STFM　(d) 原图

图 10　时空融合模块优化对比

Fig. 10　STRM optimization comparison

3.5　去噪速度对比

表 5 给出了本方法和其他方法在同一平台推理单个序列时间上的对比。测试视频的分辨率为 854 像素×480 像素，序列长度固定为 10 帧。对比实验中计入了光流计算的时间，PaCNet 中仅使用 GPU 搜索 Patch。数据结果显示，本方法在处理速度上较基于非局部自相似和光流的算法快 4~10 倍，得益于其级联多尺度架构和优化的长短跳跃连接，显著降低了计算成本。与 PaCNet 相比，本方法快约 200 倍，归因于避免了在序列间进行复杂的相似 Patch 搜索和计算。尽管参数量较大，本方法的渐进式架构允许有效利用前序特征，缩短了运行时间。最终，本方法平均处理时间接近最优，每帧仅需约 0.30 s，支持实时去噪处理。

表 5　推理单个视频序列的平均运行时间对比

Table 5　Comparison of average running time for processing video sequences

对比类型	ToFlow	VNLnet	DVDnet	FastDVDnet	ReMoNet	PaCNet	EDVDnet(本文)
运行时间/s	1.25	1.45	2.90	**0.24**	—	79.27	**0.35**
参数量/10^3	1 440	1 420	1 330	**2 480**	**804**	—	3 120

注：加粗代表最优。

4　结束语

本文研究了当前视频去噪技术中常见的纹理细节损失与效果模糊问题，提出了一种基于光流和多尺度融合的显式视频去噪算法。针对显式对齐的局限性，设计了适用于视频帧恢复的校正网络，提升了对齐的精度；针对特征冗余和噪声传播问题，提出了融入非局部思想的多尺度架构，进而实现高效且鲁棒的视频去噪性能。实验结果显示，所提方法在去噪性能上表现优异，主观视觉质量上呈现出更清晰的图像与纹理特征。未来工作将致力于优化算法设计，并探索针对真实世界视频噪声特性的去噪方法，以进一步提升去噪性能与实用性。

参考文献：

[1] LI J, WU X, NIU Z, et al. Unidirectional video denoising by mimicking backward recurrent modules with look-ahead forward ones [C]//European Conference on Computer Vision. Berlin: ECCV,2022: 592−609.

[2] ZHAO Yaping, ZHENG Haitian, WANG Zhongrui, et al. Manet: improving video denoising with a multi-alignment network[C]//2022 IEEE International Conference on Image Processing. Bordeaux: IEEE, 2022: 2036−2040.

[3] 刘迪, 贾金露, 赵玉卿, 等. 基于深度学习的图像去噪方法研究综述[J]. 计算机工程与应用, 2021, 57(7): 1−13.
LIU Di, JIA Jinlu, ZHAO Yuqing, et al. Overview of image denoising methods based on deep learning[J]. Computer engineering and applications, 2021, 57(7): 1−13.

[4] ZHANG Kai, ZUO Wangmeng, CHEN Yunjin, et al. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising[J]. IEEE transactions on image processing: a publication of the IEEE signal processing society, 2017, 26(7): 3142−3155.

[5] GUO Shi, YAN Zifei, ZHANG Kai, et al. Toward convolutional blind denoising of real photographs[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 1712−1722.

[6] HELOU M, SUSSTRUNK S. Blind universal Bayesian image denoising with Gaussian noise level learning[J]. IEEE transactions on image processing: a publication of the IEEE signal processing society, 2020, 29: 4885−4897.

[7] REN Chao, HE Xiaohai, WANG Chuncheng, et al. Adaptive consistency prior based deep network for image denoising[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 8592−8602.

[8] DAVY A, EHRET T, MOREL J M, et al. A non-local CNN for video denoising[C]//2019 IEEE International Conference on Image Processing. Taipei: IEEE, 2019: 2409−2413.

[9] 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: 2137−2146.

[10] MAGGIONI M, HUANG Yibin, LI Cheng, 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: 3465−3474.

[11] TASSANO M, DELON J, VEIT T. DVDNET: a fast network for deep video denoising[C]//2019 IEEE International Conference on Image Processing. Taipei: IEEE, 2019: 1805−1809.

[12] XUE Tianfan, CHEN Baian, WU Jiajun, et al. Video enhancement with task-oriented flow[J]. International journal of computer vision, 2019, 127(8): 1106−1125.

[13] CHAN K C K, WANG Xintao, YU Ke, et al. BasicVSR: the search for essential components in video super-resolution and beyond[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 4945−4954.

[14] 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: 1351−1360.

[15] XIANG Liuyu, ZHOU Jundong, LIU Jirui, et al. ReMoNet: recurrent multi-output network for efficient video denoising[C]//The Association for the Advancement of Artificial Intelligence. Vancouver: AAAI, 2022: 2786−2794.

[16] WANG Xintao, CHAN K C K, YU Ke, et al. EDVR: video restoration with enhanced deformable convolutional networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach: IEEE, 2019: 1954−1963.

[17] SUN Deqing, YANG Xiaodong, LIU Mingyu, et al. PWC-net: CNNs for optical flow using pyramid, warping, and cost volume[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8934−8943.

[18] BALLÉ J, LAPARRA V, SIMONCELLI E P. Density modeling of images using a generalized normalization transformation [C]//4th International Conference on Learning Representations. Puerto Rico: ICLR, 2016: 100−112.

[19] LI Xiang, WANG Wenhai, HU Xiaolin, et al. Selective kernel networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 510−519.

[20] CAO Yue, XU Jiarui, LIN S, et al. GCNet: non-local networks meet squeeze-excitation networks and beyond[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop. Seoul: IEEE, 2019: 1971−1980.

[21] HU Jie, SHEN Li, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 42(8): 2011−2023.

[22] ZAMIR S W, ARORA A, KHAN S, et al. Learning enriched features for fast image restoration and enhancement[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(2): 1934−1948.

[23] WANG Haonan, CAO Peng, WANG Jiaqi, et al. UCTransNet: rethinking the skip connections in U-net from a channel-wise perspective with transformer[C]//The Association for the Advancement of Artificial Intelligence. Vancouver: AAAI, 2022: 2441−2449.

[24] 申屠敏健. 基于先验信息和卷积神经网络的视频去噪算法研究[D]. 成都: 电子科技大学, 2021.
SHENTU Minjian. Video denoising based on prior information and convolutional neural network[D]. Chengdu: University of Electronic Science and Technology of China, 2021.

[25] JOHNSON J, ALAHI A, LI Feifei. Perceptual losses for real-time style transfer and super-resolution[C]//European Conference on Computer Vision. Cham: Springer, 2016: 694−711.

[26] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//3rd International Conference on Learning Representations. San Diego: ICLR, 2015: 1−14.

[27] PERAZZI F, PONT-TUSET J, MCWILLIAMS B, et al. A benchmark dataset and evaluation methodology for video object segmentation[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 724−732.

[28] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1026−1034.

[29] KINGMA D P, BA J L. Adam: a method for stochastic optimization[C]//3rd International Conference on Learning Representations. San Diego: ICLR, 2015: 131−142.

作者简介：

孙立辉，教授，河北经贸大学管理科学与信息工程学院院长，主要研究方向为图像处理、目标检测。获得国家发明专利授权 4 项，发表学术论文 30 余篇。E-mail：sun-lh@163.com。

陈恒，硕士研究生，主要研究方向为图像处理。E-mail：chenh@hueb.edu.cn。

商月平，副教授，主要研究方向为统计与数据分析。E-mail：stshangyueping@hueb.edu.cn。