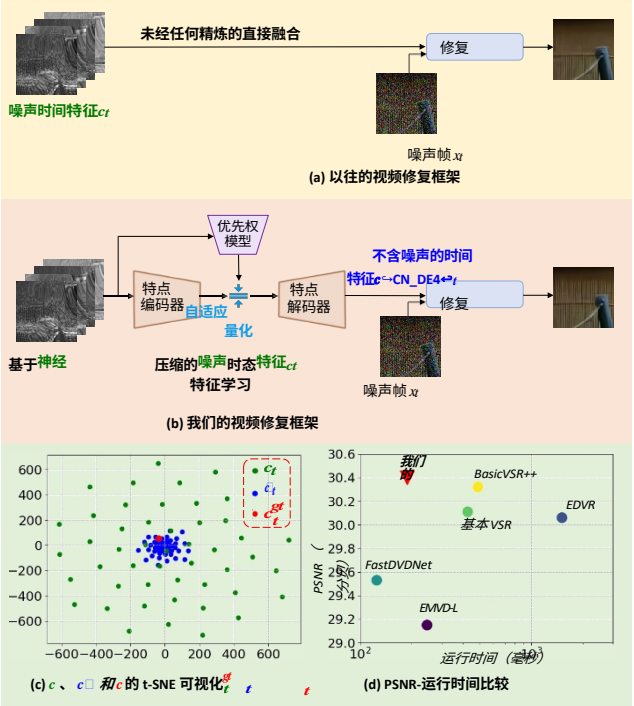
基于神经压缩的视频修复特征学习

Cong Huang *[1] Jiahao Li[2] Bin Li[2] Dong Liu[1] Yan Lu[2]
[1]中国科学技术大学[2] 微软亚洲研究院
hcy96@mail.ustc.edu.cn, dongeliu@ustc.edu.cn, {li.jiahao, libin, yanlu}@microsoft.com

摘要

如何有效利用时间特征对于视频修复来说至关重要，但也极具挑战性。时间特征通常包含各种噪声和非相关信息，它们可能会干扰当前帧的还原。本文提出通过学习噪声抑制特征表征来帮助视频还原。我们的灵感来自于神经编解码器是一种天然的去噪器。在神经编解码器中，为了节省比特率，更倾向于丢弃那些难以预测但耗费大量比特的噪声和不相关内容。因此，我们设计了一个神经压缩模块来过滤噪声，并保留特征中最有用的信息，用于视频修复。为了实现对噪声的鲁棒性，我们的压缩模块采用了空间通道量化机制，以自适应地确定潜像中每个位置的量化步长。实验表明，我们的方法可以显著提高视频去噪性能，与 BasicVSR++ 相比，我们只用了 0.23 倍 FLOPs 就提高了 0.13 dB。同时，我们的方法在视频去噪和去毛刺方面也取得了 SOTA 结果。



1. 导言

视频修复的目的是从质量下降的输入中恢复高质量的视频。典型的劣化包括各种噪音、雨水、雾霾等。它的应 用范围很广，但这一问题仍未得到充分探索。与注重单幅图像内在特征的图像复原不同[44]，视频修复更依赖于提取和利用时间特征以获得更好的质量。近期的视频修复方法主要侧重于网络结构设计，以更好地提取时间特征。例如，RViDeNet [43] 和 EDVR [36] 使用可变形卷积来对齐相邻帧的特征。BasicVSR [7] 设计了一个双向特征提取网络。BasicVSR++ [8]引入了二阶网格传播网络结构和流量引导的

*This work was done when Cong Huang was an intern at Microsoft Research Asia.

图 1. (a) 没有细化时间特征的先前框架。(b) 基于神经压缩的特征学习框架。(c) t-SNE [35] 可视化。c^{gt} 取自干净视频（*超平滑*，Set8 [34]）。对于 c^{t} 和 $c^{\tilde{}t}$，我们在相同的输入视频中添加不同的正态白高斯噪声（噪声 σ 相同，但噪声种子不同）来对这些特征点进行采样。结果表明，$c^{\cdot t}$ 对噪声的鲁棒性更强，更接近 c^{gt}。(d) 视频去噪性能比较（Set8，噪声 $\sigma = 50$）。

可变形配准网络。然而，这些方法直接使用提取的时间特征，而不进行任何再精细化处理。时间特征通常包含大量噪声和无关信息，会干扰当前帧的还原。本文以视频去噪为例，探讨如何有效利用提取的时间特征。

我们提出了一种新颖的基于神经压缩的解决方案，以完善特征并学习不含噪声的特征代表。

从神经编解码器的角度来看，噪声数据通常包含大量高频，难以预测。从神经编解码器的角度来看，噪声数据通常包含大量高频且难以预测。为了节省比特率，编解码器倾向于丢弃这些噪声和不相关的内容。这就促使我们设计一种神经压缩模块，以净化时间特征并过滤其中的噪声信息，从而实现视频还原。为了实现对噪声的鲁棒性，即让噪声扰动数据的表示与干净数据的量化表示具有高概率映射，需要正确设置量化步骤。然而，现有的神经压缩框架大多只支持固定的量化步长。这无法满足我们的目的，甚至会损害固有纹理。为了解决这个问题，我们为压缩模块设计了一种空间通道自适应量化机制，量化步长由我们的先验模型学习决定。我们的量化机制可以自适应地净化具有不同内容特征的特征。在训练过程中，交叉熵损失用于指导压缩模块的学习，并有助于保留最有用的内容。

图 1 显示了框架比较。从图 1 (c) 所示的 t- SNE [35]可视化效果来看，我们发现通过基于神经压缩的特征学习，特征对噪声的鲁棒性更强，更接近于从干净视频中生成的特征。图 1 (d) 是性能对比图。我们观察到，与之前的先进方法（SOTA）相比，我们的框架在噪声稳健特征表征的加持下，显著提高了修复质量。本文的主要贡献概述如下：

• 我们提出了一种新颖的基于神经压缩的视频还原特征学习方法。经过我们的神经压缩模块处理后，这些特征对噪声具有更强的鲁棒性，从而提高了修复质量。

• 为了实现对噪声的鲁棒性，并自适应地净化具有不同内容特征的特征，我们在空间信道上设计了一种可学习的量化机制。

• 为了进一步提高性能，我们还设计了一个注意力模块来帮助特征学习，以及一个运动矢量细化模块来改进从嘈杂视频中估算出的不连续运动矢量。

• 我们提出了一种轻量级框架。与之前的 SOTA 方法相比，我们的方法在视频去噪、去链和去毛刺方面实现了更好的质量-复杂性权衡。

2. 相关工作

2.1. 视频修复

利用时间相关性的现有视频修复方法可分为两类：滑动修复法和时间相关性修复法。

基于窗口的方法和递归方法。

基于滑动窗口的方法将相邻的几个帧作为每个帧的输入。有些方法 [10, 34] 并不依赖于明确的运动对齐。VNLNet [10] 使用非本地模块搜索跨帧的相似补丁。FastDVDNet [34] 使用堆叠 U-Net [32] 来逐步融合未对齐的相邻帧。相比之下，ToFlow [39] 和 DVDNet [33] 则使用运动估计组件来明确对齐相邻帧。为了探索更多的时间相关性，RViDeNet [43] 和 EDVR [36] 提出了特征域对齐。它们对齐的是相邻帧的特征，而不是原始像素，最近的大多数方法都采用了这种机制。

基于滑动窗口的方法时间范围狭窄，无法利用滑动窗口外的信息。相比之下，递归方法可以在较长的时间范围内学习时间特征，从而获得更好的性能。EMVD [25] 将所有过去的帧作为辅助信息进行递归组合。Yan *等人*[40] 提出了一种不需要显式排列的递归特征传播框架。BasicVSR [7] 中的特征传播使用了显式对齐。最近，Ba- sicVSR++ [8] 通过使用二阶网格传播结构和流引导的可变形配准模块，实现了出色的性能。

2.2. 视频压缩

传统的视频编解码器，如 H.264 和 H.265，采用的是由预测、变换、量化、熵编码和循环滤波组成的混合框架。得益于神经图像压缩技术的进步[3,4,27]，神经视频压缩技术[2,17,19,20,24]近来也有了长足的发展。例如，Lu *等人*[24] 设计了 DVC 模型，该模型沿用了传统视频编解码器的框架，但使用神经网络来实现其中的所有模块。继 DVC 之后，Agustsson *等人*[2] 设计了一种更先进的尺度空间光流估计。最近，Li *等人*[17] 提出了一种基于条件编码的框架，取得了更好的性能。

3. 动机

我们的动机来自于视频压缩可以过滤噪音。视频压缩的目的是用最小的比特率成本来表示视频。对于传统编解码器来说，由于很难从参考帧中预测噪声内容，因此噪声内容的残差通常很大。这些残差包含大量高频，会消耗很多比特。为了达到节省比特率的目的，传统编解码器使用量化技术来剔除噪声内容的残差，尤其是其中的高频残差，这就像一个低通滤波器。我们使用传统编解码器 x265 [1] 进行分析实验，如图 2 所示。从图 2 (c)中，我们发现传统编解码器 x265

可以在一定程度上过滤噪声。

(a) 无噪音视频 (b) 作为编解码器输入的噪声视频 (c) 经 x265 压缩 BPP: 0.0598 (d) 经 x265 压缩 BPP: 0.1695 (e) 用神经编解码器压缩 BPP: 0.0746

图 2.神经编解码器 x265 [1] 和神经编解码器 [17] 在压缩噪声视频（加性白高斯噪声，$\sigma = 20$）时的比较。BPP 表示每像素比特数，用于衡量比特率成本。

大。图 2 (d) 显示，当分配到更多比特时，x265 会对噪声进行编码，但编码的方式要平滑得多。

与使用线性 DCT（余弦变换）的传统编解码器不同，神经编解码器将学习神经编码器将视频从像素域转换到潜在特征域。然后对潜在特征进行量化，并估计其分布以进行算术编码。这样可以更准确地预测分布，从而节省更多比特率。然而，噪声和不相关内容的分布很难很好地预测。因此，为了节省比特率，在交叉熵损失的引导下，这些内容更倾向于被丢弃。图 2 (e) 显示了神经编解码器 [17] 的功效（模型权重由 [17] 的作者提供）。特别是，与 x265 相比，神经编解码器能更好地去除其中的噪音，并保留更多的语义。

受这一分析的启发，我们提出利用神经编解码器来帮助视频修复。神经编解码器用于通过量化过滤特征中的噪声信息。如果量化步骤和数据分布学习得当，噪声扰动数据的表示将高概率地映射到与干净数据相同的量化表示。抗噪特征表示将提高最终的还原质量。

使用神经编解码器而非传统编解码器的另一个优势是，神经编解码器可进行端到端训练，在与其他修复模块联合训练时性能更佳。

4. 建议的方法

4.1. 框架概述

我们设计了一种基于神经压缩的视频修复框架。我们的框架包括三个部分：特征对齐、用于学习噪声稳健特征表征的特征细化以及特征融合。该框架如图 3 所示。

特征对齐。给定噪声帧 x_{t-1} 和 x_t，我们首先使用运动估计来估计运动矢量（MV）mv_t。然后，我们设计一个运动矢量细化模型 ule，以改进从噪声视频中估算出的不连续 MV mv_t。利用改进后的 MV mv_t，通过双线性经变函数得到粗特征 $\hat{c}_t$。

特征提纯。由于 $\hat{c}_t$ 包含一些噪声和非相关信息，我们提出了一种基于神经压缩的特征提纯方法来提纯特征。据悉，我们的特征提纯部分由两个模块组成。一个是注意力模块，另一个是用于噪声稳健特征学习的神经压缩模块。

特征融合。利用抗噪特征 $\tilde{c}_t$ 和当前帧 x_t，通过还原模块生成最终输出帧 y_t。除了 y_t，还原模块部分还将生成下一步使用的时间特征 c_t。

4.2. 特征对齐

为了将上一步的时间特征与当前帧对齐，我们需要预测运动轨迹。在本文中，我们使用预先训练好的光流估计网络 SPyNet [30] 作为运动估计模块。

然而，要从降级帧中估算出准确的 MV 是相当困难的。如图 4 (a)所示，未经任何处理的 MV 存在损坏和不连续的问题，与图 4 (c)中根据干净帧估算的 MV 相比不够准确。为了解决这个问题，我们建议使用 MV 精化模块来改进 MV。MV 精炼模块采用轻量级自动编码器结构。它将损坏的 MV 编码为紧凑的表示，然后解码为精炼的 MV。详细的网络结构见补充材料。如图 4 (b) 所示，经过我们的 MV 精炼后，MV 更加干净，与干净帧中的 MV 更加相似。

4.3. 通过神经压缩完善特征

以往的递归方法直接融合当前帧和对齐的时间特征，而不进行任何再精细化处理。实际上，时间特征可能仍然包含一些不相关的噪声信息，从而干扰了当前帧的还原。

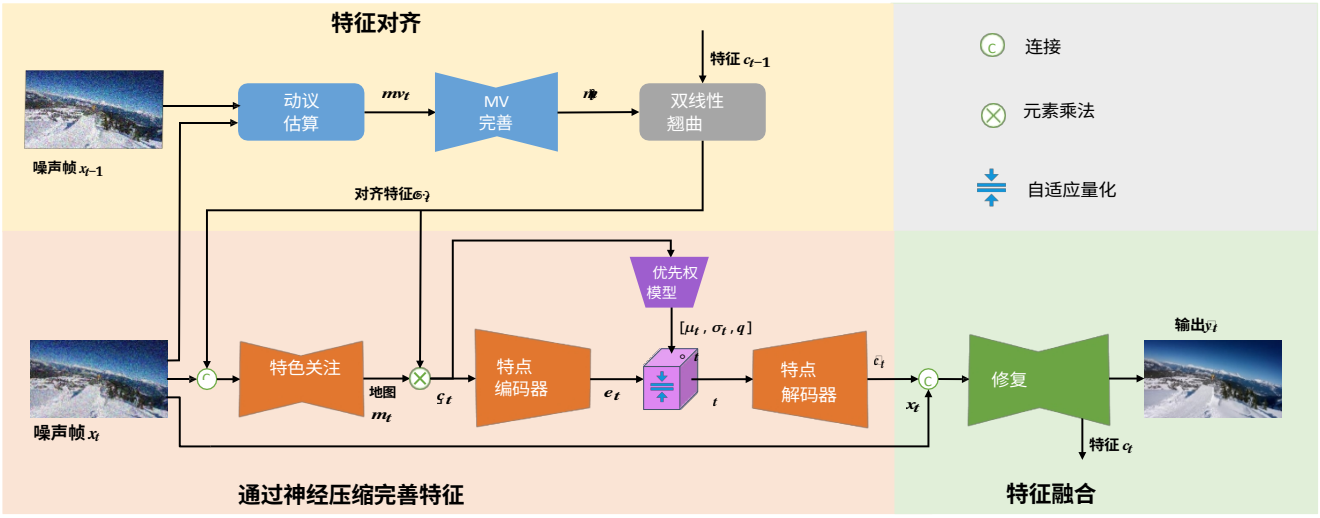


图 3.我们方法的整体框架。各模块的详细网络结构见补充材料。

(a) 精修前的 MV (b) 精修后的 MV (c) 清洁帧的 MV

图 4.MV 比较示例。我们发现，ME 改进后的 MV 更清晰，与来自干净帧的 MV 更相似。放大查看更清晰。

为了解决这个问题，我们提出了一个特征细化过程，以学习噪声抑制特征表征。这一过程由两个模块组成，即注意力模块和神经压缩模块。关于注意力机制，许多论文[14, 26, 28, 46]都对其进行了研究，并证明了它的有效性。因此，我们设计了一个注意力模块来扩展时态特征，以帮助特征学习。为了在性能和复杂性之间取得良好的平衡，我们设计了一个基于自动编码器的注意力网络，其详细的网络结构可参见补充材料。

在注意力模块之后，时间特征 $c^{\vee}_t$ 将通过拟议的神经压缩模式进行净化。

ule。按照神经图像/视频压缩的设计思路[4, 17, 24]，我们的神经压缩模块由特征编码器-解码器、量化过程和先验数据组成。

模型

首先，通过特征编码器将时间特征 $c^{\vee}_t$ 编码为紧凑的潜码 e_t：

$$e_t = \text{编码器}(c^{\vee}_t)。 \quad (1)$$

为了实现对噪声的鲁棒性，对 e_t 进行了量化。$e_t \in [s_k, s_{k+1})$ 被量化为值 $\frac{s_k+s_{k+1}}{2}$，其中 s_k 和 s_{k+1} 表示数值范围。

$e^{\cdot\cdot}_t = Encoder(c^{\cdot\cdot}_t)$ 和 $e_t = Encoder(c^{\vee}_t)$ 如果量化步长 $s_{k+1} - s_k$ 比较大，那么它们将很有可能位于同一区域 $[s_k, s_{k+1})$ 中，具有相同的量化值。也就是说，量化表示对噪声输入具有鲁棒性。然而，决定鲁棒性的前提条件是数据分布和量化步骤都是经过正确学习的。

现有的神经图像/视频压缩量化方案大多只使用固定的量化步长。事实上，内容特征在空间上存在很大差异。固定的量化步长无法很好地处理各种复杂的内容。例如，固定的小量化步长无法去除噪声信息。固定的大量化步长反而会造成较大的信息损失（即内在量化噪声）。因此，我们提出了一种自适应量化机制，即学习量化步长。示意图如图 5 所示。首先，e_t 除以学习到的量化步长 q_t，然后再对学习到的平均值 μ_t 进行分拖。然后将商数四舍五入为最接近的整数。最后，通过相反的运算得到量化的区域编码 $e^{\wedge}_t$。括号

提法是

$$e_q = \left\lfloor \frac{e_t - \mu_t}{q_t} \right\rfloor * + \mu_{\cdot t} \quad (2)$$

$\lfloor \cdot \rfloor$ 为整数舍入运算。有了量化的时间码 $e^{\wedge}_t$，然后通过特征解码器对噪声时间特征 $c^{\sim}_t$ 进行解码：

$$c^{\sim}_t = Decoder(e^{\wedge}_t). \quad (3)$$

如前所述，需要对数据分布和量化步骤进行适当的学习，以实现对噪声的控制。

让 $c^{\breve{}}_t = \quad c^{\breve{}}_t + \epsilon$ 为噪声特征，噪声ϵ。非假设*编码器*是 Lipschitz 连续的，

不稳定性实际上，我们并不知道数据的分布情况、因此我们使用先验模型对其进行估计，然后使用

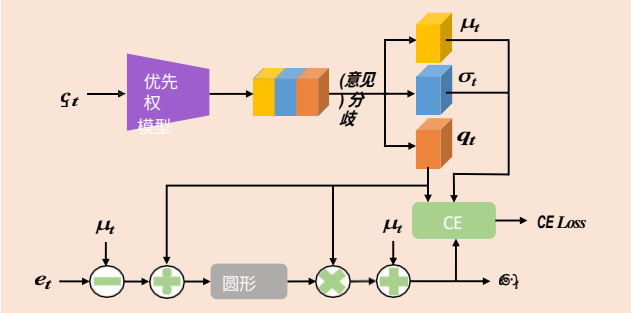


图 5 自适应量化机制图自适应量化机制示意图。CE 指交叉熵。

交叉熵损失用于指导数据分布和量化步骤的学习。交叉熵损失的计算公式为

$$\text{损失}_{CE} = E_{\hat{e}_t}\ [-\log p_{\hat{e}_t}(\hat{e}_t)], \quad (4)$$

其中 $p_{\hat{e}_t}(\hat{e}_t)$ 是潜码 $\hat{e}_t$ 的估计概率质量函数。本文沿用 [15, 17] 的假设，即 $p_{\hat{e}_t}(\hat{e}_t)$ 遵循拉普拉斯分布。由神经网络组成的先验模型用于估计分布参数。先验模型的详细结构见补充材料。但是

与[15,17]不同的是，[15,17]只估算了分布 $_{ttt}$有了（μ_t, σ_t, q_t），就可以计算出 $p_{\hat{e}_t}$（$\hat{e}_t$）的概率估计值：

$$p(\hat{e}_t) = \prod_i \left(L(\mu_{t,i}, \sigma^2_{t,i}) * U\left(-\frac{q_{t,i}}{2}, \frac{q_{t,i}}{2}\right)\right)(\hat{e}_{t,i}), \quad (5)$$

其中 i 表示每个元素在 $\hat{e}_t$ 中的空间位置。根据公式 5 中的概率质量函数，我们可以通过公式 4 计算出交叉熵损失。交叉熵损失会引导压缩模块学习适当的数据分布和量化步骤，从而实现对噪声的鲁棒性。

在我们的框架中，量化步骤 q_t 是可以从空间通道角度学习的。它可以适应具有不同内容特征的区域。图 6 是量化步长图的一个通道示例。像素强度代表量化步长的大小。像素强度越大，表示需要消除的噪声信息越多。如图 6 所示，平滑区域的量化步长通常较大，因为其中的噪声信息更容易去除。相比之下，纹理区域（如背景海报中的表格，实际上存在许多去尾信息）的量化步长通常较小。

4.4.特征融合

特征融合部分包含一个还原模块。它将把具有抗噪能力的时间特征 $\tilde{c}_t$ 与当前帧 x_t 融合，然后生成最终输出帧 $\tilde{y}_t$。

(a) 当前帧　　(b) 量化步长图

图 6.输入帧和相应量化步长图的可视化示例。该帧来自 HEVC E 级数据集 [5]。

值得注意的是，除了最终输出帧 $\tilde{y}_t$ 之外，修复模块还能生成用于下一步的时间特征 c_t，如 [7]。我们的还原模块基于轻量级 U-Net [32]。详细的网络结构见补充材料。

4.5.损失函数

在我们的方法中，损失函数包括两个项目：

$$\text{损失} = \sum_{t=1}^{n} \text{损失}_{L2}(y_t, \tilde{y}_t) + \lambda \cdot \text{损失}_{CE}(\hat{e}_t). \quad (6)$$

y_t 和 $\tilde{y}_t$ 分别为干净帧和估计帧。

损失_{L2} 是 L2 *损失*，损失_{CE} 是交叉熵损失。为了学习去噪特征表示，并让它帮助最终的重建，我们采用了两阶段训练方案，详情见补充材料。

5. 实验

我们在几项视频修复任务中对我们的方法进行了评估，包括去噪、去毛刺和去马赛克。

5.1. 数据集

视频去噪。我们测试了合成数据集和真实世界数据集。对于合成数据集，我们沿用了 FastDVDNet [34] 中的设置。训练使用包含 90 个视频的 DAVIS2017 train-val 集。Set8 用于测试。我们在干净视频中添加加性白高斯噪声（AWGN），合成噪声视频。测试了五种噪声水平，即 $\sigma = 10$、20、30、40、50。对于真实世界的数据集，我们沿用了 EMVD [25] 中的设置，并使用了 RViDeNet [43] 中的数据集。它包括一个捕获的原始视频数据集（CRVD）和一个合成的原始视频数据集（SRVD）。按照 EMVD 和 RviDeNet 的方法，我们使用 CRVD 场景 *1~6* 加上 SRVD 进行训练，并使用 CRVD 场景 *7~11* 进行测试。

视频衍生。根据文献[42]，我们在 RainSynComplex25 [21] 和 RainSynAll100 [42] 数据集上测试了我们的方法。

RainSynComplex25 包含 190 个训练视频和 25 个测试视频。RainSynAll100 包含 900 个训练视频和 100 个测试视频。

σ	VNLnet [10]	DVDNet [33]	FastDVDNet [34]	EMVD-L [25]	EMVD-S [25]		EDVR [36]	BasicVSR[7]	BasicVSR++[8]	。	我们的
1037	.10/0.9637	36.08/0.9592	36.44/0.9624	36.56/0.9624	35.01/0.944237		.16/0.9658	37.12/0.9674	37.27//0.9682	37.17/0.9684	
2033	.88/0.9360	33.49/0.9307	33.43/0.9334	33.27/0.9320	31.65/0.892734		.09/0.9379	3 4 .13/0.9397	34.25/0.9411	34.22/0.9437	
3031	.95/0.9096	31.79/0.9023	31.68/0.9066	31.40/0.9032	29.94/0.867832		.31/0.9125	32.33/0.9157	32.55/0.9168	32.57/0.9184	
4030	.55/0.8814	30.55/0.8745	30.46/0.8812	30.05/0.8761	28.64/0.832831		.02/0.8887	31.05/0.8929	31.28/0.8936	31.39/0.8970	
5029	.47/0.8561	29.56/0.8480	29.53/0.8573	29.15/0.8528	27.83/0.808230		.06/0.8660	3 0 .11/0.8690	30.32/0.8696	30.45/0.8770	
FLOP (G)	-	-	665	1106	5		3089	2947	3402	771	

表 1.在合成数据集 Set8 上与 SOTA 视频去噪方法的 PSNR/SSIM 比较。最佳性能突出显示在**红色**（第一名）和**蓝色**（第二名）。在所有噪声水平下，我们的方法都能获得最佳的 SSIM 值。

	F astDVDNet [34]	EDVR [36]	RViDeNet [43]	EMVD-L [25]	EMVD-S [25]	基本 VSR [7]	BasicVSR++ [8]	我们的	我们的-L
PSNR	44.30	44.71	44.08	44.48	42.63	44.80	**44.98**	44.72	**45.09**
SSIM	0.9881	0.9902	0.9881	0.9895	0.9851	0.9903	0.9903	**0.9906**	**0.9909**
运行时间（毫秒）	132	1511	1254	246	59	425	488	188	275

表 2.在实际数据集 CRVD [43] 上与 SOTA 视频去噪方法的比较。在默认设置下，我们的方法优于其他快速方法，并接近慢速方法。如果使用功能更强大的修复模块（即 "Ours-L"），我们可以在 PSNR 和 SSIM 方面达到 SOTA 性能。运行时间是整个数据集在 P100 GPU 上的平均帧运行时间。

视频去雾化。我们使用 REVIDE [45] 数据集，该数据集通过采集系统捕捉同一场景中的成对雾霾视频和相应的无雾霾视频。该数据集包含 42 个训练视频和 6 个测试视频。

5.2. 视频去噪结果

我们将我们的方法与这些基线进行了比较：VNL- Net [10]、DVDNet [33]、FastDVDNet [34]、EMVD [25]、EDVR [36]、BasicVSR [7]、BasicVSR++ [8] 和 RVi-DeNet [43]。EMVD 有多种复杂程度不同的网络结构配置。我们测试了大型模型（EDVR-L）和小型模型（EMVD-S）（有关配置的更多详情，请参阅补充材料）。最初的 BasicVSR/BasicVSR++ 是双向方法，可同时利用未来帧和过去帧的时间特征。为了更公平地与其他方法进行比较，我们将 BasicVSR/BasicVSR++ 修改为单向方法，只使用过去帧的时间特征。

定量比较。我们使用峰值信噪比（PSNR）和结构相似性指数（SSIM）作为定量评估指标。表 1 列出了合成噪声视频的结果，表 2 列出了真实世界噪声视频的结果。对于合成视频，如表 1 所示，我们的方法在所有噪声水平上都获得了最佳的 SSIM。在 PSNR 方面，当噪声水平较高（$\sigma = 40$ 或 50）时，我们的方法比第二好的方法 BasicVSR++ [8] 至少高出 0.11dB。此外，我们发现当噪声水平较高时，BasicVSR++ 的质量改进幅度更大。这证明我们提出的神经压缩模块可以有效地过滤噪声。值得注意的是，我们的方法的 FLOPs 仅为 BasicVSR++ 的 0.23 倍，这表明我们的方法在质量和复杂度之间实现了更好的权衡。与低复杂度方法 FastDVDNet [34] 和 EMVD- L [25] 相比，我们的方法显著提高了质量。对于真实世界中的高噪声视频，应该承认我们的方法在默认设置下目前还不能

就 PSNR 而言，我们的方法优于 BasicVSR 和 BasicVSR++，但就 SSIM 而言，我们的方法优于它们。与低复杂度方法 FastDVDNet 和 EMVD-L 相比，我们的方法能获得最佳质量。此外，如果我们将类似 U-Net [32] 的还原网络改为类似 W-Net [38] 的更复杂的还原网络（更多细节见补充材料），在表 2 中表示为 "Our-L"，我们可以同时获得最佳的 PSNR 和 SSIM，但复杂度仍然远低于 BasicVSR 和 BasicVSR++。

质量比较。图 7 显示了视觉质量比较。如图 7 所示，没有进行特征对齐的 FastDVDNet 在文本区域出现严重失真。BasicVSR++ 的结果由于时间特征中传播的噪声而变得非常模糊。相比之下，我们基于神经压缩的方法可以学习不含噪声的特征，并能还原出更加清晰的纹理。更多可视化比较见补充材料。

5.3. 视频提取结果

我们将我们的方法与之前的 SOTA 视频去雨方法进行了比较，包括 MS-CSC [18]、SE [37]、Spac- CNN [9]、FastDerain [16]、J4RNet-P [21]、FCRVD [41]、RMFD [42] 和 BasicVSR++ [8]。由于 RainSynAll100 使用积雨降解来生成雨天视频，包括 SE、MS-CSC、SpacCNN 和 FastDerain 在内的部分基准方法无法处理这种降解，因此使用 MRF [6] 作为后处理。更多详情可参见 [42]。FCRVD、RMFD、Ba- sicVSR++ 和我们的方法无需额外的后处理就能处理这种劣化。如表 3 所示，在 RainSynComplex25 上，Ba-sicVSR++ 的 PSNR 和 SSIM 优于 RMFD，但在 RainSynAll100 上，Ba-sicVSR++ 的 SSIM 略逊于 RMFD。相比之下，我们的方法在两个数据集上使用ro- bust 时间特征时，PSNR 和 SSIM 都达到了最佳水平。在 RainSynAll100 上，我们的方法带来了 0.44 dB 的 PSNR 增益和 0.0063 的 SSIM 增益。我们还测试了 RainSynLight25 [21] 和 NTURain [9]。它们的

		MS-CSC [18]	SE [37]	SpacCNN [9]	FastDerain [16]	J4RNet-P [21]	FCRVD [41]	RMFD [42]	BasicVSR++ [8]	我们的
RainSynAll100	PSNR	16.19	15.29	18.39	17.09	19.26	21.06	25.14	**27.67**	**28.11**
	SSIM	0.5078	0.5053	0.6469	0.5824	0.6238	0.7405	**0.9172**	0.9135	**0.9235**
RainSynComplex25	PSNR	16.96	16.76	21.21	19.25	24.13	27.72	32.70	**33.42**	**34.27**
	SSIM	0.5049	0.5273	0.5854	0.5385	0.7163	0.8239	0.9357	**0.9365**	**0.9434**

表 3.在 RainSynComplex25 [21] 和 RainSynAll100 [42] 上与 SOTA 视频推导方法的比较。我们使用与我们相同的设置训练 Ba- sicVSR++ [8]。其他基线结果由 RMFD [42] 论文提供。

	DCP [12]	GDNet [22]	DuRN [23]	KDDN [13]	MSBDN [11]	论坛渔业局 [29]	VDN [31]	EDVR [36]	CG-IDN [45]	BasicVSR++ [8]	我们的
PSNR	11.03	19.69	18.51	16.32	22.01	16.65	16.64	21.22	**23.21**	21.68	**23.63**
SSIM	0.7285	0.8545	0.8272	0.7731	0.8759	0.8133	0.8133	0.8707	**0.8836**	0.8726	**0.8925**

表 4.在 REVIDE [45] 测试集上与 SOTA 视频去毛刺方法的比较。我们使用与我们相同的设置训练 BasicVSR++ [8]。其他基线结果由 CG-IDN [45] 论文提供。

	M_a	M_b	M_c	M_d
MVR		✓	✓	✓
NCFL			✓	✓
FA				✓
PSNR	29.75	29.87	30.29	30.45

表 5.不同模块的消融研究。在 Set8（$\sigma = 50$）上进行了测试。MVR 是 MV 精炼，NCFL 是基于神经压缩的特征学习。FA 是特征关注。

	NCFL-AdapQ	NCFL-FixedQ	NCFL-NoQ
PSNR	30.29	29.86	29.98

表 6.量化消融研究。在 Set8（$\sigma = 50$）上测试。NCFL-AdapQ 是自适应量化的默认模型，即表 5 中的 M_c。NCFL-FixedQ 表示我们使用现有神经视频编解码器中的固定量化步骤。NCFL-NoQ 取消了量化，只是一个普通的自动编码器。

结果见补充材料。图 7 还显示了视觉质量对比。我们可以看到，我们的模型可以很好地去除雨痕，并产生更清晰、更美观的结果。

5.4. 真实世界视频去噪结果

表 4 显示了我们的方法与先前的 SOTA 真实世界视频去毛刺方法的比较：DCP [12]、GDNet [22]、DuRN [23]、KDDN [13]、MSBDN [11]、FFA[29]、VDN[31]、EDVR[36]、CG-IDN[45]和 Ba- sicVSR++[8]。如表 4 所示，BasicVSR++ 优于 EDVR，但不如 MBSDN 和 CG-IDN，后者是专门为去毛刺任务设计的。相比之下，我们的方法与排名第二的 CG-IDN 相比，PSNR 提高了 0.42 dB，SSIM 提高了 0.0089。此外，我们的方法的参数为 16M，只有
是拥有 2300 万个参数的 CG-IDN 的 0.70 倍。如图 7 所示，我们的方法得到的结果更加直观。

5.5. 消融研究

本文提出了三个关键模块：用于改进 MV 的 MV 精化（MVR）、基于神经压缩的特征学习（NCFL）和自适应

量化和特征关注 (FA)。我们研究了这些模块的效果，结果见表 5。在没有 MVR、NCFL 和 FA 的情况下，基线模型只包含运动估计模块、双线性扭曲和恢复模块。

MV 精炼（MVR）。如表 5 所示，基线模型 M_a 的 PSNR 仅为 29.75dB。它的问题在于根据噪声视频估算的 MV 不连续。当使用我们的 MVR 时，MV 得到了改进，M_b 达到了 PSNR
29.87 分贝。我们的 MVR 将 PSNR 提高了 0.12 分贝。

基于神经压缩的特征学习（NCFL）。
如果进一步结合 NCFL 和 MVR，M_c 将达到 PSNR
30.29 dB，与 M_b 相比提高了 0.42 dB。显著的改进验证了 NCFL 的有效性。此外，我们还研究了 NCFL 的两个变体。如表 6 所示，没有量化的普通自动编码器（即 NCFL-NoQ）的 PSNR 降至 29.98dB。这说明 NCFL 带来的改进主要来自自适应量化机制，而不是模型参数的增加。此外，我们还测试了 NCFL-FixedQ，它与许多现有的神经视频编解码器一样使用固定的量化步骤。NCFL- FixedQ 的 PSNR 下降到 29.89 dB。其性能甚至不如 NCFL-NoQ。这表明，固定量化步骤反而会丢失一些有用的信息，在学习噪声抑制表征方面也会失败。相比之下，可学习的空间信道量化步骤可以自适应地过滤噪声，净化具有不同内容特征的时间特征，这一点相当重要。

特征关注（FA）。本文还提出了一个 FA 模块，以进一步帮助特征学习。如表 5 所示，M_d 的 PSNR 为 30.45 dB。FA 使 PNSR 提高了 0.16 dB，可见其有效性。

5.6. 关于不同退化的 NCFL

表 5 和表 6 列出了 AWGN 退化条件下的 NCFL。然而，我们的 NCFL 并不局限于 AWGN。它对其他复杂降解也非常有效，例如实词去噪、去链和去色。表 7 显示了对多种降解的综合研究。例如，M_1 与 M_3

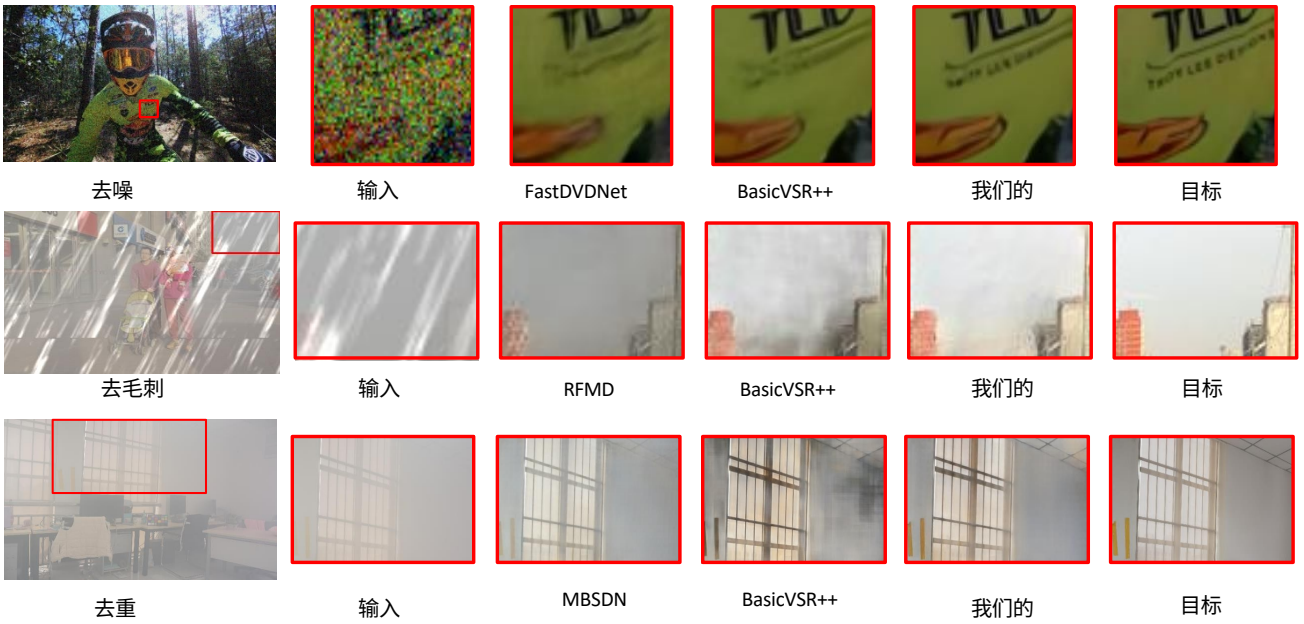


图 7 去噪去噪：Set8 测试集中的*摩托车*视频，噪声方差为 50。去污：RainSynAll100 测试集中的 *0985* 视频。去毛刺REVIDE 数据集中的 *L006* 视频。

	去毛刺	去光晕	RWD	A W G N
M_1 : w/o NCFL	27.30	23.07	44.48	29.99
M_2 : w/ NCFL (w/o CE)	27.64	23.30	44.56	30.20
M_3 : w/ NCFL	28.11	23.63	44.72	30.4 5

表 7. 不同降解类型下的 NCFL 研究。CE 指交叉熵损失。RWD 指真实世界去噪。AWGN 表示加性白高斯噪声。

方向	方法		PSNRFLOPs (G)
Uni-direction	BasicVSR	30.11	2947
Uni-direction	BasicVSR++	30.32	3402
统一方向	我们的	30.45	771
双向	基本 VSR	30.68	5855
双向	BasicVSR++	31.10	7097
双向	我们的	31.21	1522

0.11 dB。

结果表明，NCFL 可以为 deraining 带来 0.81 dB 的增益。这些实质性的改进验证了我们的 NCFL 的有效性。此外，M_2 和 M_3 之间的比较表明，交叉熵损失可以有效地指导 NCFL 在多重退化条件下的学习。

5.7. 双向视频去噪

在之前的实验中，我们主要关注的是时间特征仅来自过去时间的单向设置。而在双向环境下，时间特征既可以来自过去时间，也可以来自未来时间。我们方法的一个优势是可以轻松扩展到双向环境。我们测试了 BasicVSR [7]、BasicVSR++ [8] 和我们的双向模型。PSNR 和复杂度比较如表 8 所示。如表 8 所示，双向设置为 BasicVSR 带来了 0.59 dB 的增益，BasicVSR++ 带来了 0.78 dB 的增益，而我们的方法带来了 0.76 dB 的增益，复杂度约为 2 倍。在双向设置下，我们的方法仍以 0.21 倍的 FLOPs 优于 BasicVSR++

表 8.Set8 的双向视频去噪，$\sigma = 50$。

6. 结论和限制

在本文中，我们设计了一种基于神经压缩的视频修复框架。我们从神经视频编解码器可以自然过滤噪声这一事实中得到启发，进而提出利用神经压缩来纯化时间特征并学习抗噪特征表示。为了解决固定量化步骤会损害固有纹理的问题，我们提出了一种可学习的空间通道量化机制，以实现对噪声的鲁棒性。同时，我们还提出了注意力模块和 MV 细分模块，以进一步提高性能。实验结果表明，与之前的 SOTA 方法相比，所提出的方法实现了更好的质量-复杂度权衡。

虽然我们的方法比之前的大多数 SOTA 方法更快，但我们方法的推理速度仍不能满足实时场景的要求。未来，我们将继续提高我们的方法在实时视频还原中的效率。

参考资料

[1] Ffmpeg. https://www.ffmpeg.org/.2, 3

[2] Eirikur Agustsson、David Minnen、Nick Johnston、Johannes Balle、Sung Jin Hwang 和 George Toderici。用于端到端优化视频压缩的规模空间流。In *Pro- ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8503-8512, 2020.2

[3] 约翰内斯-巴勒（Johannes Balle´）、瓦莱罗-拉帕拉（Valero Laparra）和埃罗-P-西蒙切利（Eero P Simoncelli）。端到端优化图像压缩》，*arXiv preprint arXiv:1611.01704*, 2016.2

[4] Johannes Balle´, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston.使用尺度超优先级的变分图像压缩。*ArXiv 预印本 arXiv:1802.01436*, 2018.2, 4

[5] Frank Bossen 等：《通用测试条件和软件参考配置》。见 *JCTVC-L1100*，第 12 卷，2013 年。5

[6] 蔡博伦、徐向民、陶大成。基于时空 mrf 的实时视频消隐。*环太平洋多媒体会议*，第 315-325 页。Springer, 2016.6

[7] Kelvin CK Chan、Xintao Wang、Ke Yu、Chao Dong 和 Chen Change Loy。BasicVSR: The search for essential com- ponents in video super-resolution and beyond.*In Proceed- ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4947-4956, 2021.1, 2, 5, 6, 8

[8] Kelvin CK Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy.BasicVSR++: Improving Video Super- Resolution with Enhanced Propagation and Alignment. *ArXiv preprint arXiv:2104.13371*, 2021.1, 2, 6, 7, 8

[9] Jie Chen, Cheen-Hau Tan, Junhui Hou, Lap-Pui Chau, and He Li. 在 cnn 框架下进行鲁棒视频内容对齐和雨水去除补偿。In *Proceedings of the IEEE conference on computer vision and pattern recogni- tion*, pages 6286-6295, 2018.6, 7

[10] Axel Davy、Thibaud Ehret、Jean-Michel Morel、Pablo Arias 和 Gabriele Facciolo.通过 cnn 进行非局部视频去噪。*ArXiv 预印本 arXiv:1811.12758*, 2018.2, 6

[11] Hang Dong、Jinshan Pan、Lei Xiang、Zhe Hu、Xinyi Zhang、Fei Wang 和 Ming-Hsuan Yang.密集特征融合的多尺度提升去雾网络。*IEEE/CVF 计算机视觉与模式识别会议论文集》*，第 2157-2167 页，2020 年。7

[12] 何开明、孙健、唐晓鸥利用暗通道先验去除单幅图像雾度。*电气和电子工程师学会图像分析与机器智能交易》*，33（12）：2341-2353，2010 年。7

[13] 洪明、谢源、李翠华和曲艳云。利用异构任务模仿进行图像去毛刺。In *Pro- ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3462-3471, 2020.7

[14] Jie Hu, Li Shen, and Gang Sun. 挤压与激发网络工程。In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132-7141, 2018.4

[15] 胡志浩 胡 Pytorch 视频 压缩，https://github.com/zhihaohu/pytorchvideocompression，2020。5

[16] Tai-Xiang Jiang（蒋泰祥）、Ting-Zhu Huang（黄廷柱）、Xi-Le Zhao（赵喜乐）、Liang-Jian Deng（邓良建）和 Yao Wang（王耀）。Fastderain：使用方向梯度先验的新型视频雨痕去除方法。*IEEE 图像处理论文集*》，28（4）：2089-2102，2018.
6, 7

[17] Jiahao Li、Bin Li 和 Yan Lu.深度上下文视频通信。*神经信息处理系统进展*》，34，2021。2, 3, 4, 5

[18] Minghan Li, Qi Xie, Qian Zhao, Wei Wei, Shuhang Gu, Jing Tao, and Deyu Meng.通过多尺度卷积稀疏编码去除视频雨痕。 In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni- tion*, pages 6644-6653, 2018.6, 7

[19] Jianping Lin, Dong Liu, Houqiang Li, and Feng Wu.M-LVC：用于学习视频压缩的多帧预测。*IEEE/CVF 计算机视觉与模式识别会议论文集*，2020 年。2

[20] Jerry Liu, Shenlong Wang, Wei-Chiu Ma, Meet Shah, Rui Hu, Pranaab Dhawan, and Raquel Urtasun.用于高效视频压缩的条件编码（Conditional en- tropy coding for efficient video compression）。*ArXiv 预印本 arXiv:2008.09180*, 2020.2

[21] 刘佳莹、杨文涵、杨帅、郭宗明。视频中的深度联合递归雨点去除和重构。In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3233- 3242, 2018.5, 6, 7

[22] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen.Grid-dehazenet：基于注意力的多尺度图像去毛刺网络。In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7314-7323, 2019.7

[23] Xing Liu、Masanori Suganuma、Zhun Sun 和 Takayuki Okatani.利用配对操作潜力进行图像复原的双残差网络。In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7007-7016, 2019.7

[24] Guo Lu、Wanli Ouyang、Dong Xu、Xiaoyun Zhang、Chunlei Cai 和 Zhiyong Gao。DVC：端到端深度视频通信框架。In *Proceedings of the IEEE/CVF Con- ference on Computer Vision and Pattern Recognition*, pages 11006-11015, 2019.2, 4

[25] Matteo Maggioni、黄一斌、李成、肖帅、傅中谦、宋凤龙。利用递归时空融合实现高效多级视频去噪。*IEEE/CVF 计算机视觉与模式识别会议（CVPR）论文集*》，第 3466-3475 页、
2021 年 6 月2, 5, 6

[26] Yiqun Mei, Yuchen Fan, Yulun Zhang, Jiahui Yu, Yuqian Zhou, Ding Liu, Yun Fu, Thomas S Huang, and Humphrey Shi.用于图像复原的金字塔注意网络。*ArXiv 预印本 arXiv:2004.13824*, 2020.4

[27] David Minnen, Johannes Balle´, and George Toderici.用于学习图像压缩的联合自回归和分层先验。*ArXiv 预印本 arXiv:1809.02736*, 2018.2

[28] 牛犇、文伟磊、任文奇、张祥德、杨连平、王淑珍、张开浩、曹晓春、沈海峰。通过整体注意力网络实现单图像超分辨率*欧洲计算机视觉会议*》，第 191-207 页。Springer, 2020.4

[29] Xu Qin、Zhilin Wang、Yuanchao Bai、Xiaodong Xie 和 Huizhu Jia.FFA-Net：用于单张图像去毛刺的特征融合注意力网络。In *Proceedings of the AAAI Con- ference on Artificial Intelligence*, volume 34, pages 11908- 11915, 2020.7

[30] Anurag Ranjan 和 Michael J Black.使用空间金字塔网络进行光流估计。In *Proceedings of the IEEE conference on computer vision and pattern recogni- tion*, pages 4161-4170, 2017.3

[31] Wenqi Ren, Jingang Zhang, Xiangyu Xu, Lin Ma, Xiaochun Cao, Gaofeng Meng, and Wei Liu.利用语义分割进行深度视频去毛刺。*IEEE Transactions on Image Processing*, 28(4):1895-1908, 2018.7

[32] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox：用于生物医学图像分割的卷积网络。*医学图像处理和计算机辅助干预国际会议*，第 234-241 页。
Springer, 2015.2, 5, 6

[33] Matias Tassano、Julie Delon 和 Thomas Veit。DVDnet：用于深度视频去噪的快速网络。In *2019 IEEE In- ternational Conference on Image Processing (ICIP)*, pages 1805-1809.IEEE, 2019.2, 6

[34] Matias Tassano、Julie Delon 和 Thomas Veit。Fastdvdnet：实现无流量估计的实时深度视频去噪。In *CVPR*, pages 1354-1363, 2020.1, 2, 5, 6

[35] Laurens Van der Maaten 和 Geoffrey Hinton.使用 t-SNE 实现数据可视化》。*机器学习研究期刊*》，9（11），2008 年。1, 2

[36] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy.EDVR：利用增强型可变形卷积网络进行视频修复。In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pat- tern Recognition Workshops*, pages 0-0, 2019.1, 2, 6, 7

[37] Wei Wei、Lixuan Yi、Qi Xie、Qian Zhao、Deyu Meng 和 Zongben Xu.我们应该将视频中的雨条纹编码为去终结性还是随机性？在 *IEEE 全国计算机视觉会议论文集上*，第 2516-2525 页，2017 年。6, 7

[38] Xide Xia 和 Brian Kulis.W-net：用于完全无监督图像分割的深度模型。*ArXiv preprint arXiv:1711.08506*, 2017.6

[39] 薛天帆、陈百安、吴佳俊、魏东来、William T Freeman。面向任务流的视频增强。*国际计算机视觉杂志*》，127（8）：1106-1125，2019.2

[40] Bo Yan, Chuming Lin, and Weimin Tan.帧和特征上下文视频超分辨率。*美国人工智能学会会议论文集*》，第 33 卷，第 5597-5604 页，2019 年。2

[41] 杨文汉、刘家英、冯嘉仕。双层流的帧一致递归视频推演。In *Proceedings of the IEEE/CVF Conference on Computer Vi- sion and Pattern Recognition*, pages 1661-1670, 2019.6, 7

[42] 杨文汉、谭罗比、冯嘉仕、王诗琪、程斌、刘佳颖。递归多帧

*模式分析与机器智能》*期刊，2021 年。5, 6, 7

[43] Huanjing Yue, Cong Cao, Lei Liao, Ronghe Chu, and Jingyu Yang. 用动态场景基准数据集监督原始视频去噪。*IEEE/CVF 计算机视觉与模式识别会议论文集》*，第 2301-2310 页，2020 年。1, 2, 5, 6

[44] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. 多阶段渐进式图像复原。In *Pro- ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14821-14831, 2021.1

[45] 张欣怡、董航、潘金山、朱超、戴莹、王成杰、李吉林、黄飞跃和王飞。学习还原雾霾视频：一个新的真实世界数据集和一种新方法。*IEEE/CVF 计算机视觉与模式识别大会论文集》*，第 9239-9248 页，2021 年。6, 7

[46] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. 用于图像修复的残留非局部注意力网络。In *7th International Conference on Learning Represen- tations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.4