深度学习视频压缩简介

视频以及图像的有损压缩算法会造成较为严重的失真以及效应，比如，基于块的编码策略将会引起块效应；高频分量的缺失会造成压缩后的图像会更加模糊，还有振铃效应，颜色偏移等等。特别是在编码是在较差的编码配置下（低比特率）尤为明显。这些效应会严重降低用户体验，所以如何去除这些效应或者削弱这些效应的影响也就成为一个重要的问题。

在新一代视频编码标准 HEVC（High Efficiency Video Coding）中，采用两种环路滤波的方案来削弱这些效应：去块滤波器以及 SAO（样点自适应补偿）。从名字上来看，去块滤波器主要针对受损视频的块效应。而 SAO 则使用附加的偏置来补偿其他的效应，这个偏置通过编码器计算并且随着码流传输到解码器辅助解码说明 SAO 可以实现 BD-rate 下降。

随着人工智能近几年热度逐渐上升，其算法深度学习也在更广泛的领域中发挥作用。它采用深层神经网络来提取数据的表征，并且将其组合为高层语义特征，构造一个非线性映射。在计算机视觉领域，图像识别，图像标记，目标跟踪等高层次任务上已经得到了很好的效果，而在诸如图像的超分辨率以及降噪等低层次视觉任务中，深度学习也逐渐展现出其优越的性能。

1. 深度学习在视频后处理中的应用

目前深度学习在视频后处理上的应用可以分为两类，一类是环内滤波，一类是环外滤波。环内滤波指的是在 HEVC 编码环中，使用深度学习网络来替换原来的后处理模块来提升编码性能，如图 1。

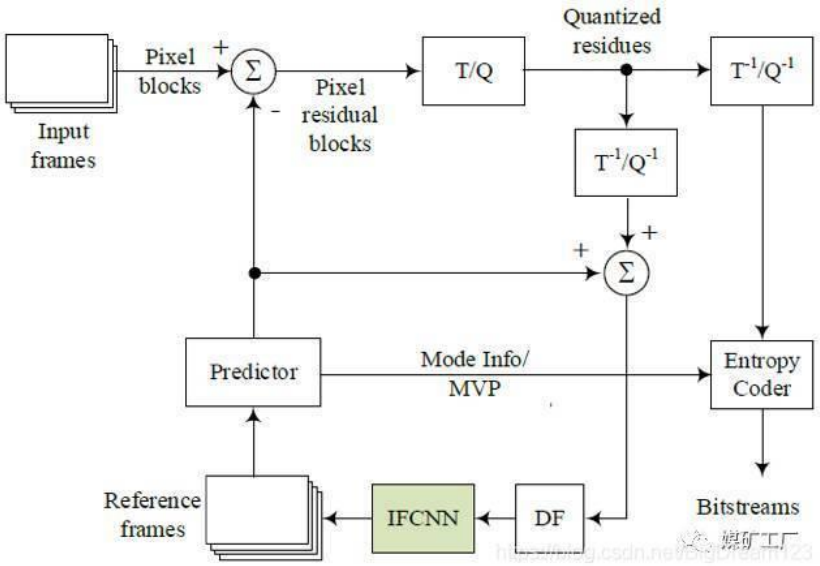

图 1 环内滤波示意

环外滤波则不需在 HEVC 编码环中进行替换，正常编码的码流在解码端解码完成后使用神经网络后处理滤波即可。在编码端也可以提供一些辅助解码的参数信息，作为边信息融入码流中进行传输。

1.1 环内滤波

1.1.1 IFCNN

Park 和 Kim 首先提出一种使用卷积神经网络来进行环内滤波的方法，具体的结构如图 1 所示，采用神经网络替换 HEVC 后处理技术中的 SAO。该网络整体由三个卷积层构成，引入残差网络的思想，使得神经网络不需要直接生成高质量的图像，而只需要学习高质量图像与压缩受损图像之间的残差即可，这样就加快了训练的速度，保证了收敛性。为了让神经网络的训练更加贴合编码本身，作者令视频序列通过关闭 SAO 的编码器，将重建的 YUV 文件以及其对应的 Ground_truth 组合作为训练集。

作者对 ALL-Intra 模式以及 LDP、RA 模式分别训练网络模型，然后将其整合到 HEVC 参考软件 HM 16.0 中，实验结果如表 1 所示。

表 1 IFCNN 客观性能测试（与原始编码算法相比）

Sizes	Seq.	All Intra	LDP-Case I	LDP-Case II	RA-Case I	RA-Case II
		BDBR (%)	BDBR (%)	BDBR (%)	BDBR (%)	BDBR (%)
832×480	BD	-10.1	-5.3	-3.0	-6.0	-6.7
	BQM	-3.7	-3.0	-2.4	-2.4	-2.9
	PS	-2.7	-2.0	-1.2	0.0	-1.1
	BDT	-7.6	-3.5	-2.4	-4.3	-4.9
416×240	BP	-3.3	-2.8	-1.5	-0.6	-1.1
	BQS	-2.4	-3.3	-2.9	1.4	-0.8
	B	-3.4	-2.3	-2.6	0.0	-1.4
	RH	-4.9	-0.4	0.6	-1.2	-1.6
	Avg.	-4.8	-2.8	-1.9	-1.6	-2.6

这项技术一个缺陷是训练和测试集都选自同一视频序列，虽然取不同帧，但是由于一组序列帧与帧间的内容和分布很相似，所以训练处模型的推广能力不足，不过也证明了深度学习在视频后处理这个领域的极大潜力。

1.1.2 VRCNN

Dai 等在 IFCNN 的基础上，提出一种多滤波器尺寸的卷积神经网络结构来进行后处理，使用网络模型完全替换后处理模块来提升编码性能。作者参考 GoogleNet[4]的思想，增加网络深度的同时，也在网络宽度上进行扩展，即使用多个小尺寸的卷积窗的并行组合来替换单个大尺寸卷积核，不同尺寸的卷积核可以提取到不同层次的图像特征，因此使用这种方法，可以在一层中整合图像的多种特征，有利于图像的重建。网络结构如图 2。

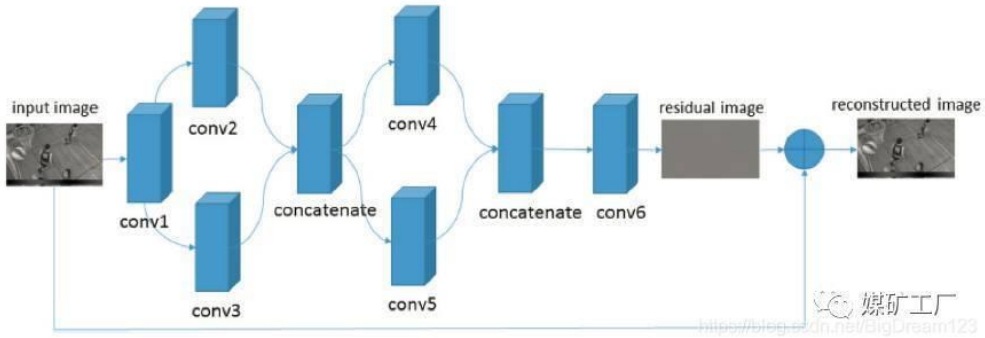

图 2 VRCNN 网络结构

在此结构中，第二层和第三层网络均使用两个并行且尺寸不同的卷积核对特征图进行提取，并且按通道整合在一起。这里仍然使用残差连接的思想，神经网络只

需学习受损图像与 Ground_Truth 之间的残差，从而加速网络收敛，削弱过拟合的影响。为了提升模型的推广能力，在此作者使用自然图像作为训练集，自然图像具有较为广泛的统计特性与特征，因此可以覆盖绝大多数的视频情景。将图片输入关闭去块滤波器和 SAO 的 HM 编码器，得到的重建码流即可作为训练输入数据。这项技术中，只对 ALL-Intra 的编码模式进行测试，同时每个序列只对第一帧进行测试，客观测试性能见表 2。

表 2 VRCNN 的客观测试结果（与原后处理算法）

Class	Sequence	BD-rate		
		Y (%)	U (%)	V (%)
Class A	Traffic	−5.6	−3.5	−4.1
	PeopleOnStreet	−5.4	−5.9	−5.7
	Nebuta	−0.9	−4.9	−4.1
	SteamLocomotive	−1.9	−0.5	−0.3
Class B	Kimono	−2.5	−1.5	−1.4
	ParkScene	−4.4	−3.3	−2.5
	Cactus	−4.6	−3.9	−6.3
	BasketballDrive	−2.5	−3.7	−5.3
	BQTerrace	−2.6	−3.3	−3.0
Class C	BasketballDrill	−6.9	−5.8	−6.8
	BQMall	−5.1	−5.3	−5.3
	PartyScene	−3.6	−4.4	−4.4
	RaceHorses	−4.2	−6.7	−11.0
Class D	BasketballPass	−5.3	−4.4	−6.5
	BQSquare	−3.8	−4.2	−6.4
	BlowingBubbles	−4.9	−8.4	−7.9
	RaceHorses	−7.6	−8.5	−11.5
Class E	FourPeople	−7.0	−5.3	−5.2
	Johnny	−5.9	−5.0	−5.5
	KristenAndSara	−6.7	−6.1	−6.2
Class Summary	Class A	−3.5	−3.7	−3.6
	Class B	−3.3	−3.2	−3.7
	Class C	−5.0	−5.5	−6.9
	Class D	−5.4	−6.4	−8.1
	Class E	−6.5	−5.5	−5.6
Overall	**All**	−4.6	−4.7	−5.5

1.1.3 MMS-Net

ICIP2017 中，Kang 和 Kim 等提出一种多模型/多尺度的卷积模型来提升后处理性能，多尺度的 CNN 结构能够有效提高图像的重建性能。另外编码视频中的 CTU（编码树单元）信息可以指导网络正确检测和去除分块伪影，作者也使用 CU

（编码单元）和 TU（变换单元）信息这类编码参数来对重建进行辅助。网络模型见图 3。

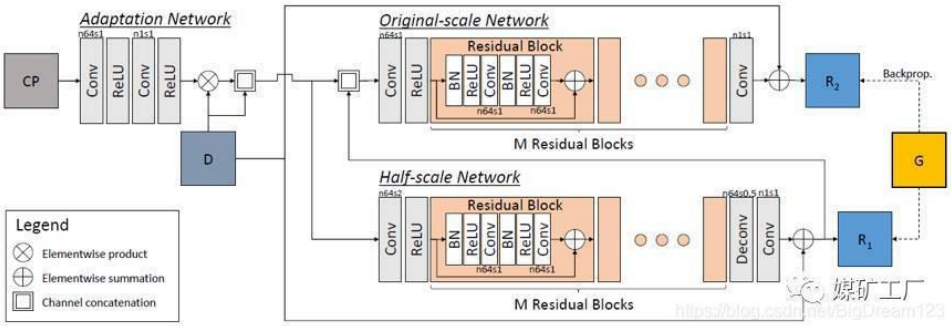

图 3 MMS-Net 模型

结构图中 CP 指的是编码参数，文章中具体位置 CU 和 TU 信息。D 指的是受损的图像，Rk 代表第 k 个尺度模型恢复的图像，G 则代表 Ground_truth。提取到的编码参数首先需要经过预处理，将 CU（或 TU）边界像素值设为 2，非边界区域像素设为 1，如图 4。之后将处理后的 CP 图输入到一个自适应网络中（见图 3 左上角），将 CP 信息转换为图像的特征空间，并投影到单通道特征图中。该特征图与输入的受损图像逐元素相乘，作为旁路信息输入到多尺度网络中。

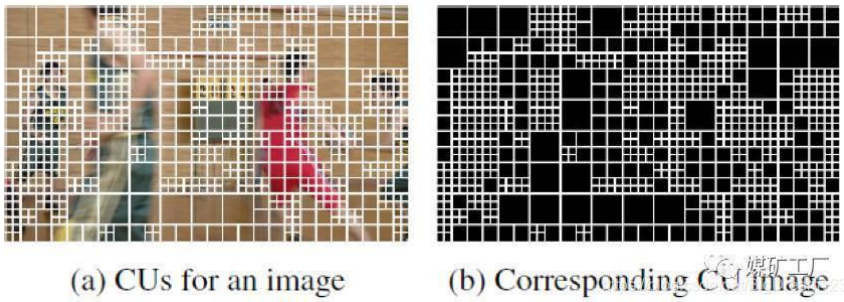

图 4 提取 CU 信息的一个实例

多尺度图像复原可以看为在多尺度空间中的分层处理，可以使恢复后的图像在精细尺度上保留一些较小的细节也可以在较粗的尺度上保留长时依赖。在图 4 中可以看到模型有两个尺度的连续网络。粗尺度网络从半尺寸输入图像中恢复损伤图片，再通过精细网络进一步修饰与恢复。在图 3 半尺度网络（Half-scale Network）中，输入帧是通过步长为 2 的卷积层进行下采样而不需要在网络外部对图像尺寸进行调整，最后再使用反卷积层将其上采样至原始图像尺寸，嵌入在网络中的

插值结构也简化了整个系统的处理过程。网络的主体参照了 Res-Net，采用多个残差块级联的方法加速收敛。

作者使用 Xiph.org Video Test Media 中的 28 个 HD 视频作为训练集，在全帧内模式下的重构序列作为输入数据，全帧内条件下测试性能见表 3。

表 3 MMS-Net 及其他网络性能的比较

		Average BD-rate		
Testset	Network	Y	U	V
classC	VDSR	-4.4%	-5.1%	-5.7%
	VRCNN	-4.3%	-6.0%	-6.9%
	MMS-net ($M = 5$)	-7.7%	-9.7%	-11.3%
	MMS-net ($M = 15$)	**-9.3%**	**-13.1%**	**-15.7%**
classD	VDSR	-4.1%	-5.1%	-6.5%
	VRCNN	-3.6%	-5.8%	-7.3%
	MMS-net ($M = 5$)	-6.5%	-8.6%	-10.9%
	MMS-net ($M = 15$)	**-7.7%**	**-10.9%**	**-15.7%**

1.2 环外滤波

上面简述的三种环内滤波方法均是在 HEVC 编码环中代替部分或者全部后处理模块，这种方法编码出的码流在解码时，需要再通过神经网络进行处理。下面将介绍一种环外滤波的思路。

Wang 和 Chen 等提出在解码器后加入神经网络模型来提升视频的重建质量。同时因为 HEVC 支持多种尺寸的变换单元（TU），使用码流中的 TU 信息来选择训练集的图像块大小，而作者也说明这种方法比统一采样的训练数据得到的结果更具鲁棒性。

参考深层超分辨率网络 VDSR，该网络堆叠 10 层卷积并且只使用 ReLU 作为激活单元，每一层的卷积核则为 3*3。与 VRCNN 一样使用 400 张自然图片作为训练集，使用 HM16.0 压缩后作为输入训练数据。训练好的模型在 AI、LDP、LDB 以及 RA 四种模式下进行测试，测试结果见表 4。

表 4 DCAD 在四种编码模式下的测试结果（与 HEVC baseline 相比）

Sequences		AI			LP		
		Y	U	V	Y	U	V
Class B	Kimono	-3.7%	-2.6%	-2.5%	-5.4%	-8.6%	-7.2%
	ParkScene	-4.6%	-3.9%	-2.9%	-4.2%	-5.3%	-4.0%
	Cactus	-4.3%	-4.5%	-9.1%	-7.0%	-10.7%	-14.4%
	BasketballDrive	-2.8%	-7.7%	-9.6%	-4.9%	-13.6%	-14.9%
	BQTerrace	-1.8%	-4.5%	-5.0%	-4.8%	-10.0%	-9.9%
Class C	BasketballDrill	-7.8%	-11.2%	-14.4%	-6.5%	-11.6%	-14.3%
	BQMall	-4.8%	-5.7%	-6.0%	-6.0%	-8.6%	-8.8%
	PartyScene	-2.3%	-4.5%	-5.3%	-2.1%	-5.1%	-5.7%
	RaceHorses	-3.6%	-7.2%	-11.8%	-4.6%	-11.9%	-16.9%
Class D	BasketballPass	-5.0%	-7.1%	-9.8%	-5.1%	-7.7%	-11.3%
	BQSquare	-3.3%	-3.2%	-6.2%	-3.7%	-4.5%	-9.4%
	BlowingBubbles	-4.2%	-8.7%	-8.7%	-4.1%	-9.4%	-9.0%
	RaceHorses	-8.4%	-10.7%	-14.3%	-8.0%	-12.5%	-15.6%
Class E	FourPeople	-8.3%	-6.6%	-7.2%	-12.1%	-13.5%	-14.5%
	Johnny	-7.3%	-9.0%	-8.1%	-11.5%	-18.0%	-15.2%
	KristenAndSara	-7.7%	-7.5%	-8.1%	-11.7%	-14.6%	-15.6%
Overall		**-5.0%**	**-6.5%**	**-8.1%**	**-6.4%**	**-10.3%**	**-11.7%**

Sequences		LB			RA		
		Y	U	V	Y	U	V
Class B	Kimono	-4.2%	-7.3%	-5.9%	-4.4%	-6.4%	-4.9%
	ParkScene	-3.9%	-4.4%	-3.1%	-4.3%	-3.6%	-1.3%
	Cactus	-6.0%	-9.4%	-12.8%	-6.9%	-9.5%	-11.7%
	BasketballDrive	-3.6%	-11.2%	-13.1%	-4.1%	-10.2%	-12.3%
	BQTerrace	-1.7%	-5.5%	-5.3%	-3.1%	-5.0%	-4.3%
Class C	BasketballDrill	-5.6%	-10.7%	-13.5%	-6.1%	-11.2%	-14.4%
	BQMall	-5.6%	-7.7%	-7.8%	-5.8%	-7.0%	-7.0%
	PartyScene	-1.3%	-4.2%	-4.6%	-1.1%	-3.5%	-4.0%
	RaceHorses	-3.9%	-10.9%	-15.9%	-4.3%	-11.6%	-16.0%
Class D	BasketballPass	-4.8%	-7.4%	-11.0%	-4.8%	-7.7%	-10.9%
	BQSquare	-1.6%	-2.9%	-7.5%	-1.4%	-1.6%	-6.6%
	BlowingBubbles	-3.5%	-8.7%	-8.6%	-3.3%	-8.7%	-7.9%
	RaceHorses	-7.5%	-12.0%	-15.1%	-8.0%	-12.3%	-15.5%
Class E	FourPeople	-11.3%	-12.5%	-13.4%	-11.1%	-10.4%	-10.8%
	Johnny	-9.4%	-15.2%	-13.2%	-8.8%	-13.5%	-11.1%
	KristenAndSara	-10.6%	-13.1%	-14.0%	-10.3%	-11.[illegible]	-[illegible]0%
Overall		**-5.3%**	**-8.9%**	**-10.3%**	**-5.5%**	**-8.3%**	**-9.4%**

上海交通大学研究团队同样提出一种环外后处理滤波的编解码结构，与 DCAD 相似，使用 VDSR 作为网络模型训练处理。但在此基础上，于编码器前加入一个分类的模块，用来提取每一帧图像的统计信息，并使用 K-means 算法将其分类。考虑到 CNN 在图像重建上的应用本质上是对图像底层统计特征提取并且重组的过程，这里预先对输入序列进行统计分类是合理的。分类信息也将作为辅助信息嵌入到编码码流之中。而后处理模块也将提取这些辅助信息选用不同的模型进行处理。

1.3 总结

深度学习模型通过学习受损图像与 ground_truth 之间的端对端映射，对压缩的视频进行有效的滤波处理。通过上述方案的描述，我们可以看出深度学习在视频滤波这一领域上的极大潜力。但是当前提出的技术主要在全帧内模式下发挥作

用，一旦开启码率控制，性能就会变得不稳定。因此一方面应考虑更适合帧间编码的训练策略或者网络结构，另一方面应注意图像视频本身的统计特性，将其加入到模型中辅助重建。