论文阅读报告

论文信息

标题：《Perceptual Quality Improvement in Videoconferencing Using Keyframes-Based GAN》

作者：Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, Alberto Del Bimbo

论文：https://doi.org/10.1109/TMM.2023.3264882

关键词：图像复原、面孔、流媒体、可视化、图像编码、特征提取、视频会议、视频恢复、生成对抗性网络

摘要

这篇论文为视频会议中压缩视频的质量改善提供了一种创新的`GAN-based`技术解决方案，用于减少视频会议中的压缩伪影，并改善视频的视觉质量，特别强调了面部特征的恢复和关键帧的智能使用

1. 目的和背景：考虑到视频会议的重要性以及有损视频压缩算法减少了所需的带宽但同时降低了视觉质量，该研究旨在通过新技术改善压缩视频的视觉效果。
2. 方法特点：
    - 利用关键帧：利用视频流中高质量的 I 帧作为参考关键帧，这些帧在整个传输过程中保持不变，为视觉质量改善提供指导。
    - 更新策略：采用一种新的更新策略，维护并更新一个紧凑且有效的关键帧参考集。
3. 技术实现：
    - 多尺度特征提取：从压缩帧和参考帧中提取多尺度特征。
    - 特征组合：根据面部标志，以渐进方式结合这些特征，以恢复视频压缩后丢失的高频细节。
4. 实验结果：实验证明，所提出的方法在高压缩率下也能提高视觉质量并生成逼真的结果。

一、引言

1. 研究背景与需求：
    - 由于 COVID-19 大流行，视频会议已成为全球个人和商业通信的主要手段。这一背景强调了需要更高质量视频通信的必要性。
2. 现有技术的问题：
    - 使用有损视频压缩算法（如 `H.264` 和 `H.265`）可以减少视频传输所需的带宽，但同时会引入压缩伪影，降低视频流的感知质量。这种质量下降恶化了用户体验，甚至在某些情况下令人无法接受。
3. 研究动机和目标：
    - 鉴于这些问题，提高视频质量的研究构成了一个非常活跃的研究领域。特别是，生成对抗网络（GAN）已被证明是改善图像和视频处理任务中的一种有前景的强大工具，它们能够生成逼真和感知上令人满意的结果。
4. 研究贡献：
    - 本工作提出了一种新的基于 `GAN` 的方法，用于改善视频会议中的视觉质量。研究聚焦于改善画面中人物的表现，尤其是头部区域，因为这是人际沟通中最具表现力和重要的部分。利用视频压缩算法中的 I 帧作为高质量的参考关键帧，以帮助恢复视频压缩后丢失的高频细节。

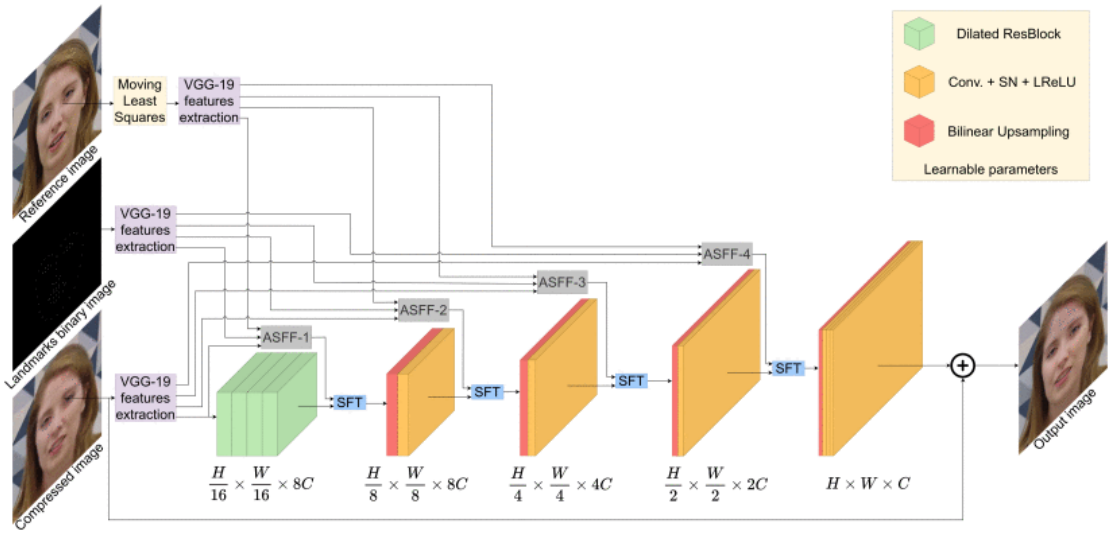
二、相关工作

本文回顾了视频编码、视频质量改善、面部质量改善等方面的相关工作，并指出现有方法的局限性。例如，一些方法需要学习压缩，这在实际使用中存在较高的市场门槛。

三、提出的方法

架构

本文提出了一种新颖的 `GAN` 架构，如下图所示：

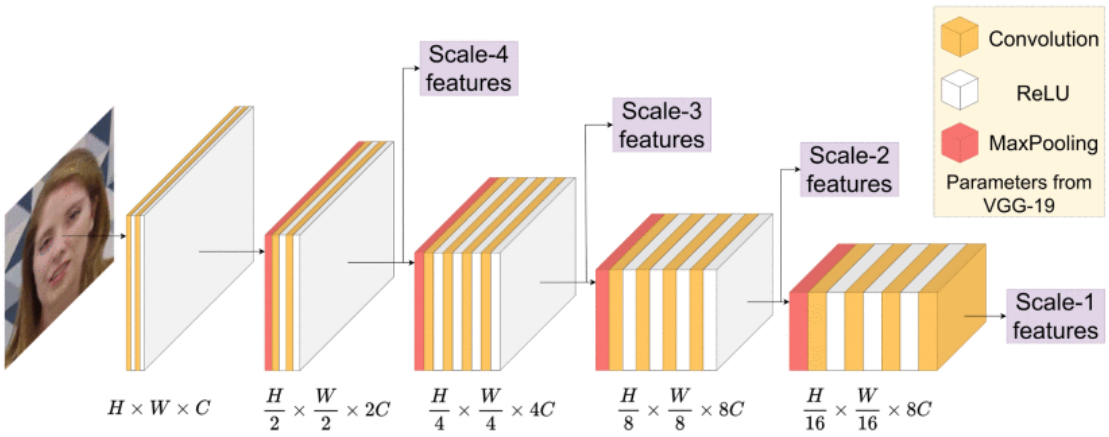


与参考工作的不同之处：直接对参考图像进行变形而不是仅对其特征进行变形，并且以渐进方式在多个尺度上提取和融合特征，以帮助网络恢复从粗糙到细致的细节。这种方法强调了在恢复视频细节时，如何有效地利用高质量关键帧以及动态更新策略。

图像恢复模型

图像恢复模型是在上述架构基础上实际进行图像处理和细节恢复的具体应用，目的是从压缩图像生成恢复后的图像，流程如下：

1. 输入图像处理：
    - 使用一个压缩图像作为输入，该图像在传输或存储过程中遭受了质量损失。
    - 使用一个高质量的参考图像（通常是视频中未经压缩的 I 帧）。
    - 利用面部特征的二值图像，这些特征在压缩图像上被标记出来。
2. 特征提取：
    - 使用预训练的 `VGG-19` 模型从上述三个输入图像（退化图像、参考图像和地标二值图像）中提取不同尺度的特征。
    - 将参考图像根据面部特征使用移动最小二乘法变形，以与压缩图像更好地对齐。
    - 特征是在 `VGG-19` 网络的不同层次（relu_2_2, relu_3_4, relu_4_4 和 conv_5_4）中提取的。



3. 特征对齐：
    - 应用自适应实例归一化（AdaIN）技术，调整参考特征以减少风格和光照差异。
    - 我们用 F^d 和 F^g 表示退化特征和引导特征。`AdaIN` 可以写成 $F^{g,a} = \sigma(F^d)\left(\frac{F^g - \mu(F^g)}{\sigma(F^g)}\right) + \mu(F^d)$，其中 σ(·) 和 μ(·) 分别代表平均值和标准偏差。
4. 特征上采样和特征融合：
    - 在经过多个扩张残差块处理后，压缩（退化）的特征被逐步上采样，即空间分辨率被放大，同时通道数被减少。这一步骤有助于从低分辨率的特征中恢复出更细节的图像内容。
    - 自适应空间特征融合（ASFF）：这一技术用于结合来自不同源的特征（如退化特征和参考特征），优化特征组合，以恢复图像质量。
    - 空间特征变换（SFT）：SFT 模块利用从 ASFF 块输出的特征生成仿射变换参数（缩放因子 α 和位移参数 β），用于空间特征调制。输出的 SFT 公式为 $SFT = \alpha \cdot F^r + \beta$，其中 · 表示元素级乘积，$F^r$是恢复的特征，即从退化特征中恢复出来的特征。
5. 残差学习和跳跃连接：
    - 文章根据 Increasing Video Perceptual Quality with GANs and Semantic Coding | Proceedings of the 28th ACM Int 中的方法，训练网络学习退化图像与恢复输出之间的残差图像，这有助于减少训练时间并提高网络训练的稳定性。
    - 跳跃连接（skip connection）：在退化输入图像和恢复输出之间设置跳跃连接，有助于保持图像的基本结构和信息，从而提高恢复质量。

ASFF 块

在基于示例的方法中，参考图像和退化图像的特征融合是核心部分，因为它允许充分利用参考图像提供的信息。而且简单的串接(concatenation-based)方法不能充分利用

参考特征。

文章强调，在他们的多尺度架构中，他们依赖于多个自适应空间特征融合（ASFF）块。尽管参考图像通常包含更多的高质量细节，但在重建整体面部组成部分时，降级图像应具有更大的权重。例如，如果参考图像的嘴部是闭合的，而压缩图像的嘴部是开放的，那么牙齿的重建应主要基于从降级图像恢复的特征。为此，ASFF 块产生一个基于降级图像面部标记的注意力掩码，以指导引导特征和恢复特征的融合。

这段内容突出这种方法的一种重要思路，那就是利用两种类型的图像——高质量的参考图像和低质量的降级图像——来提高视频质量，通过对两者中的特征进行融合，既利用了高质量图像的细节，又考虑到了降级图像的结构和内容。

使用移动最小二乘法的参考图像变形

对于大多数引导式人脸修复方法来说，参考图像和降级图像之间的姿势和表情差异会降低其性能，因为这会在重建结果中引入伪像。因此，我们使用基于移动最小二乘法（MLS）的图像变形方法对参考图像和压缩图像进行空间对齐。

这个过程涉及到根据参考图像和降级图像的面部标记找到一个变形函数，然后将这个函数应用到参考图像的所有点上，以便根据降级图像的面部标记进行扭曲。这样可以得到更好的对齐并改进恢复效果。

通过数学推导可得出变形函数的表达式如下：

$$f(v) = (v - p_*)\left(\sum_{i=1}^{N} \hat{p}_i^T w_i \hat{p}_i\right)^{-1} \sum_{j=1}^{N} w_j \hat{p}_j^T \hat{q}_j + q_*$$

将此变形函数应用于参考图像的每个点，就可以根据退化图像的面部地标对其进行扭曲。

关键帧选择和设置维护

选择最佳的参考关键帧自然是姿势和表情与退化图像相似的关键帧，而不是简单地使用前一个关键帧。

新颖的方法：随着视频的进行，我们可以保存一组有限的关键帧，以减少内存需求，然后使用最相似的关键帧作为参考来还原当前的压缩帧。

关键帧替换策略：采用最少使用频率（LFU）缓存替换策略。对于集合中的每个关键帧，记录其被选择用于重构的次数。当从视频流中接收到新的关键帧时，最少使用的关键帧将被替换。为了克服初期关键帧使用频率高的问题，引入了指数衰减规则，即当新关键帧到来时，所有关键帧的使用次数计数器将减半。

训练损失

这一部分描述了用于训练图像恢复模型的损失函数，这些损失函数包括多个组件以确保高质量的图像重建。

1. 多元损失函数：模型使用加权和的形式结合了重建损失和真实感损失来训练。
2. 重建损失：
    - 均方误差（MSE）损失：用于确保重建图像尽可能接近于高质量的未压缩真实图像（ground-truth）。MSE 损失计算方法是：

$$\ell_{MSE} = \frac{1}{CHW}\|I_R - I_{GT}\|^2$$

    - 其中 *C*, *H*, 和 *W* 分别代表图像的通道数、高度和宽度。
    - 感知损失：基于 VGG-19 特征空间，感知损失评估重建图像与真实图像在高层视觉特征上的差异。计算方法是：

$$\ell_{perc} = \sum_{l\in L} \frac{1}{C_l H_l W_l}\|\Psi_l(I_R) - \Psi_l(I_{GT})\|^2$$

    - 其中 Ψ_l 表示预训练的 VGG-19 模型中第 *l* 层的特征，*L* 是选定的层集合。
3. 真实感损失：
    - 风格损失：基于每层的 Gram 矩阵，风格损失评估重建图像与真实图像在风格上的一致性。

$$\ell_{style} = \sum_{l\in L} \frac{1}{C_l H_l W_l}\|\Psi_l(I_R)^T\Psi_l(I_R) - \Psi_l(I_{GT})^T\Psi_l(I_{GT})\|^2$$

    - 对抗损失：使用多尺度判别器，对抗损失帮助网络生成在视觉上更真实的图像。它通过评估判别器对重建图像的判别能力来更新生成器和判别器。
4. 训练稳定化：为了稳定判别器的学习，采用了谱归一化，这是通过每个卷积层后应用的，以规范化层的谱范数。
5. 总损失函数：最终的训练损失是上述各个损失的加权和，其中包括 MSE 损失、感知损失、风格损失和对抗损失：

$$\ell_{total} = \lambda_{MSE}\ell_{MSE} + \lambda_{perc}\ell_{perc} + \lambda_{style}\ell_{style} + \lambda_{adv}\ell_{adv,G}$$

6. 其中 $\lambda_{MSE}, \lambda_{perc}, \lambda_{style}, \lambda_{adv}$是各自损失的权重参数。

这些损失函数的组合旨在确保重建图像不仅在像素级别上准确，而且在感知级别和风格上也与真实图像一致，从而提升重建图像的整体质量和真实感。

四、实验结果

通过在 DFD 和 HDTF 集上的实验，作者展示了所提出方法的有效性。与其他最先进的方法相比，该方法在多个指标上取得了更好的性能，包括 LPIPS、CONTRIQUE、

CONTRIQUE-FR 和 VMAF-NEG。此外，主观实验结果也表明，所提出的方法在人类评估者中获得了更高的真实性评分。

定量结果

表 I 是在 DFD 数据集上对拟议方法和其他最新 CRF 42 方法的定量比较：

Method	PSNR ↑	SSIM ↑	LPIPS ↓	BRISQUE ↓	CONTRIQUE ↓	CONTRIQUE-FR ↓	VMAF ↑	VMAF-NEG ↑
GWAINet [23]	22.25	0.608	0.129	**24.18**	50.16	20.79	44.65	36.60
HiFaceGAN [25]	29.38	0.828	0.075	28.41	48.75	18.67	47.77	45.11
PSFR-GAN [27]	**29.68**	0.833	0.057	29.07	46.87	16.46	48.55	46.22
GFP-GAN [31]	27.51	0.822	0.081	34.17	50.84	23.01	**57.55**	48.51
GPEN [26]	27.61	0.813	0.075	28.67	49.42	21.36	55.86	49.26
DFDNet [29]	27.03	0.827	0.065	32.38	46.84	16.04	55.15	48.95
ASFFNet [7]	28.29	**0.834**	0.062	29.67	46.27	17.48	51.74	46.84
Ours	26.19	0.779	**0.037**	27.41	**44.95**	**13.16**	56.87	**54.20**

Best and second best results are in bold and underlined, respectively. ↑= higher values are better, ↓= lower values are better.

1. 在最具指示性的全参考感知度量 LPIPS 方面，以及在 CONTRIQUE、CONTRIQUE-FR 和 VMAF-NEG 方面，所提出的方法都取得了最佳性能。

2. PSFR-GAN 在信号指标 PSNR 和 SSIM 方面表现更好，而 GWAINet 则在 BRISQUE 方面取得了最佳结果。

3. 不过，人工检测显示，GWAINet 生成的图像包含过多的高频伪影，因此作者在其他实验中没有考虑这种方法。

4. 从定性结果中可以看出，GFP-GAN 可获得最佳 VMAF 值，这可能是因为它倾向于使色彩饱和并增加对比度，但代价是失去逼真度。这种趋势与图像增强方法的应用类似，而图像增强会提高 VMAF 分数。 为了支持这一理论，可以注意到 VMAF-NEG 分数与 VMAF-NEG 分数之间的巨大差异，相比之下，VMAF-NEG 分数不受图像增强技术的影响。

5. 作者的方法同时获得了第二好的 VMAF 值和最好的 VMAG-NEG 值，证明了它有能力在保持逼真度的同时获得很好的整体视频质量。

6. 此外，VMAF 和 VMAF-NEG 分数表明，作者的视频结果具有时间一致性，不会出现过多的运动抖动、闪烁或蚊虫噪音。

表 II 是作者提出的方法与其他最新方法在 HDTF 数据集上对 CRF 42 的定量比较

Method	PSNR ↑	SSIM ↑	LPIPS ↓	BRISQUE ↓	CONTRIQUE ↓	CONTRIQUE-FR ↓	VMAF ↑	VMAF-NEG ↑
HiFaceGAN [25]	**30.70**	**0.864**	0.047	31.71	41.50	10.69	44.50	42.40
PSFR-GAN [27]	30.31	0.853	0.046	**30.01**	44.99	13.57	40.18	38.40
GFP-GAN [31]	28.19	0.846	0.064	35.02	46.80	13.95	**51.06**	41.27
GPEN [26]	27.72	0.817	0.061	30.62	46.15	13.31	47.52	40.57
DFDNet [29]	28.16	0.847	0.050	32.10	47.50	12.04	49.48	43.90
ASFFNet [7]	27.88	0.835	0.058	32.18	45.39	12.90	40.15	35.72
Ours	30.39	0.862	**0.028**	33.34	**37.35**	**7.76**	47.82	**45.07**

Best and second best results are in bold and underlined, respectively. ↑= higher values are better, ↓= lower values are better.

表 II 所报告的第二个实验中，作者在 HDTF 数据集上比较了所提出的方法和基线方法。

1. 模型尚未在该数据集上进行过训练，因此我们无法评估其泛化能力。

2. 在 LPIPS、CONTRIQUE、CONTRIQUE-FR 和 VMAF-NEG 方面，所提出的方法优于其他方法。

3. 人工检查结果表明，这可能是由于几种竞争方法倾向于增加（或相反，隐藏）皮肤瑕疵或过度增强嘴唇和眉毛的颜色。

定性结果

DFD 数据集定性结果展示如下图所示：

1. 作者的方法在生成逼真细致的结果方面优于所有基线方法。模型利用了参考关键帧，在不损失逼真度的情况下重现了经过这种强压缩后丢失的高频细节。

2. 值得注意的是，参考图像（即输入列中左下方的图像）往往与降级图像不太相似，但作者提出的方法仍能利用它。例如，在最后一行中，参考图像的眼睛是睁开的，而压缩图像的眼睛是闭着的，尽管如此，作者的模型还是正确地描绘出了恢复后的闭眼画面。

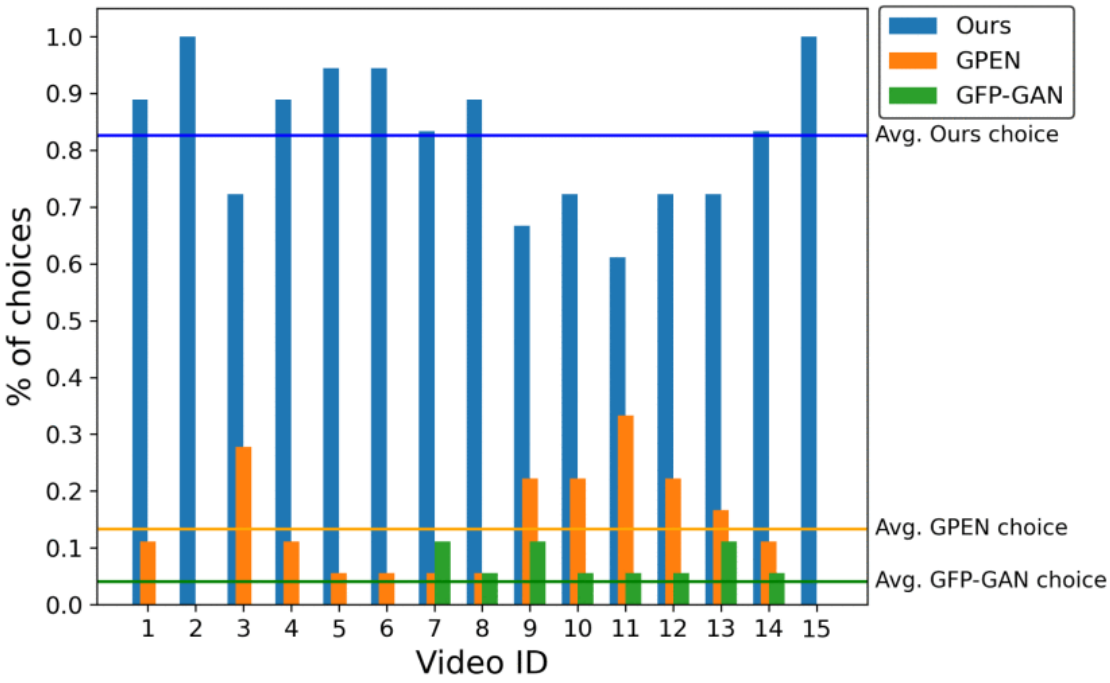
HDTF 数据集的定性结果如下图所示：

1. 作者的方法生成的图像还是最为细致逼真。
2. 总的来说，作者的方法能最稳定地生成与地面实况相似的令人满意的结果。

主观实验

使用 AVrate Voyager 工具进行了基于三选一强制选择（3-AFC）方法的主观测试

与 GPEN 和 GFP-GAN 相比，作者提出的 GAN 方法引入的高频细节和色彩偏移更少，因此更受青睐；与使用质量指标对帧进行单独评估相比，作者提出的 GAN 方法在视频序列中引入的高频细节和色彩偏移更明显。



推理时间

模型的推理时间已经与基线进行过比较，相关的实验是在 NVIDIA RTX 2080 Ti GPU 上执行的。这个模型实现了每秒大约 45 帧的性能，证明了其具备实时推理的能力，适合用于视频会议。

这些内容表明，论文不仅关注模型的恢复质量，也重视其在实际操作中的表现，特别是在速度和效率方面，以满足实际应用的需要。

消融研究

探讨了模型中各个组件的重要性以及它们对最终性能的影响。

作者系统地测试了模型中的各个组件，如 ASFF 块、SFT 模块和不同的损失函数，通过逐一移除这些组件来评估它们对整体性能的贡献。实验表明，使用多尺度特性是架构的最重要组成部分，其次是 ASFF 和 SFT 块。

五、结论

1. 本文提出的基于 GAN 的方法和关键帧选择系统在改善视频会议视频的视觉质量方面表现出色，特别是在增强面部外观方面。

2. 该系统的一个关键要素是更新一组先前的 I 帧并利用它们来改善视觉质量的策略。

3. 所提出的方法在感知度量上优于竞争方法，并且在人类评估中得到了更高的真实性评分。

附录

作者提供了详细的定量结果分析，支持了所提出方法的有效性，并讨论了不同评估指标与实际感知质量的相关性。

个人评价

这篇论文针对视频会议中的一个实际问题——压缩伪影，提出了一个创新的解决方案。通过利用 GAN 和关键帧的概念，作者不仅提高了视频的视觉质量，而且还保持了实时处理的能力。论文的实验设计严谨，结果令人信服。此外，作者还进行了详尽的消融研究，以证明所提出方法的各个组成部分的有效性。总体而言，这是一篇在视频处理领域具有重要贡献的研究工作。