复杂数据集上的高效视频生成

Aidan Clark
DeepMind
London, UK
aidanclark@google.com

Jeff Donahue
DeepMind
London, UK
jeffdonahue@google.com

Karen Simonyan
DeepMind
London, UK
simonyan@google.com

摘 要

通过多尺度学习的有力帮助，自然图像的生成模型已朝着高保真度样本发展。我们试图通过展示在复杂的 Kinetics-600 数据集上训练的大型生成对抗网络能够生成比以前的工作复杂度更高的视频样本，从而将这一成功示范带入视频建模领域。我们提出的模型，双视频判别器 GAN（DVD-GAN），通过利用其判别器的计算有效分解，扩展到更长和更高分辨率的视频。 我们评估视频合成和视频预测的相关任务，并在 Kinetics-600 的预测上实现最新的 Fréchet Inception Distance，以及 UCF-101 数据集上的综合最先进的得分，同时建立在 Kinetics-600 上合成的强基线。

1 介绍

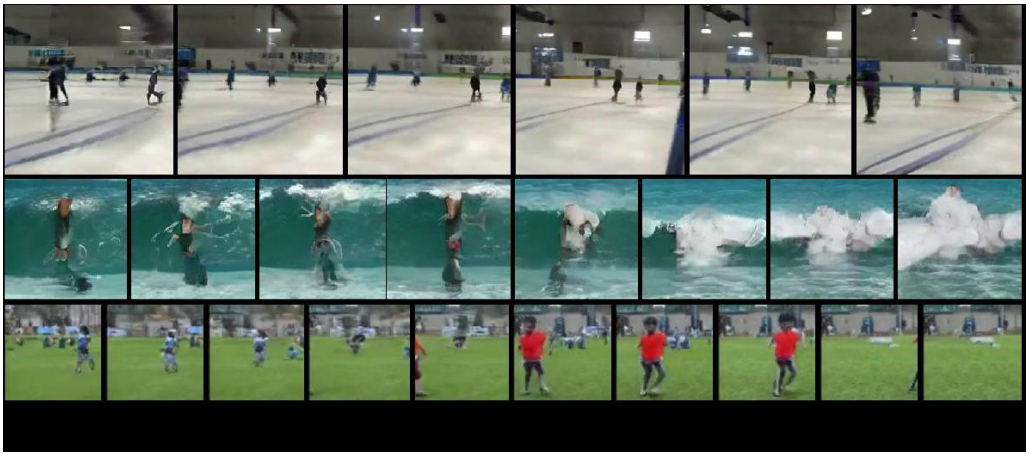

图 1：由在 Kinetics-600 上训练的 DVD-GAN 在 256×256,128×128 和 64×64 分辨率（从上到下）生成的视频中选择的帧。

当在高分辨率和多样化数据集上进行训练时，当前深度生成模型可以产生逼真的自然图像[10,25,28,34,39]。 自然视频的生成对于生成建模来说是一个明显的进一步挑战，但却受到数据复杂性和计算要求的增加的困扰。出于这个原因，许多关于视频生成的先前工作围绕着相对简单的数据集或采用可获得强时间条件信息的任务。

Preprint. Under review.

图 2：生成有趣行为的视频样本。 按照栅格扫描顺序：a）关于摄像机的变化。 b）放大物体。 c）结构良好的移动物体。 d）笔留在纸上的精细细节。 第一个样本是 128×128，其他所有样本都是 64×64。

我们专注于视频合成和视频预测（在 2.1 节中定义）的任务，旨在将生成图像模型的强大结果扩展到视频领域。我们建立在最先进的 BigGAN 架构[10]之上，引入了许多视频特定的修改，包括有效的可分离注意和判别器的时空分解。 这使我们能够在 Kinetics-600 上进行训练 - 自然视频的复杂数据集比常用数据集大一个数量级。 由此产生的模型 Dual Video Discriminator GAN（DVD-GAN）能够生成具有显著保真度的时间相干且高分辨率的视频（图 1）。

我们的贡献如下：

- 我们提出 DVD-GAN--一种可扩展的自然视频生成模型，可生成分辨率高达 256×256，长度高达 48 帧的高质量样本。
- 我们在 UCF-101 上实现了视频合成的最新技术，并在 Kinetics-600 上进行了预测。
- 我们在 Kinetics-600 上建立了类条件视频合成作为生成视频建模的新基准，并将 DVD-GAN 结果报告为强基线。

2 背景

2.1 视频合成与预测

生成视频建模是一个广泛探讨的问题，其中包括对 VAE [5,16,30,22]的工作，自回归模型[38,48,24,62]，归一化流（Flow）[29]和 GAN [33,59,42,41]。任务的精确推导在所提供的调节信号的类型方面不同。一个极端是无条件视频合成，即任务是在训练分布之后生成任何随机的视频。另一个极端是强条件模型，包括以内容传输的另一个视频为条件的生成[7,68]，每帧分割掩模[61]或姿势信息[60,58,66]。在中间地带，任务会比无条件生成更结构化，但从建模角度来看比强条件生成（通过其输入获得有关生成视频的大量信息）更具挑战性。类条件视频合成的目的是生成给定类别的视频（例如，“骑自行车”），而未来视频预测涉及给定初始帧的连续视频的生成。这些问题在几个方面有所不同，但有一个共同的要求，即需要产生真实的时间动态，在这项工作中，我们将自己局限于这两个问题。

2.2 生成对抗网络

生成性对抗网络（GANs）[19]是由判别器 D 和生成器 G 之间的极小极大游戏定义的一类生成模型。最初的目标由[19]提出，并且已经提出了许多改进，主要是针对改进训练稳定性[4,67,10,20,36]。我们使用目标[32,10]的铰链公式，通过梯度下降（ρ代表逐元素的 ReLU 函数）优化：

$$\mathcal{D}\colon \min_{\mathcal{D}} \mathop{\mathbb{E}}_{x\sim data(x)}\Big[\rho(1-\mathcal{D}(x))\Big] + \mathop{\mathbb{E}}_{z\sim p(z)}\Big[\rho(1+\mathcal{D}(\mathcal{G}(z)))\Big], \quad \mathcal{G}\colon \max_{\mathcal{G}} \mathop{\mathbb{E}}_{z\sim p(z)}\Big[\mathcal{D}(\mathcal{G}(z))\Big]$$

GAN 具有众所周知的局限性，包括生成的样本中存在有限多样性的趋势（称为模式崩溃的现象）以及由于缺乏对数据的明确似然度量而难以进行定量评估。尽管有这些缺点，GAN 已经在许多视觉领域产生了一些最高保真度的样本[25,10]。

2.3 用于视频的多判别器 GAN

高质量视频通常包含一致的对象，这些对象能够连贯地进行。先前相关工作的焦点聚集在分解上，其将对象的纹理和空间一致性与其时间动态分开建模。一种方法是将 G 分成前景和背景模型[59,47]，而另一种方法考虑 G 或 D 中的显式或隐式光流[42,37]，第三种方法是将 D 分解成与运动分开判断图像质量的子网络。 例如，除了在整个视频的切片上操作的判别器之外，MoCoGAN [54]还包含用于各个帧的单独判别器; 其他模型区分不同分辨率的帧组[65,49]或子批次[41]。 其中一些方法的好处是 D 不再以全分辨率处理整批视频。

2.4 Kinetics-600

Kinetics 是一个包含 10 秒高分辨率 YouTube 剪辑的大型数据集[26,2]，最初是为人类行为识别任务而创建的。我们使用数据集的第二次迭代，Kinetics-600 [13]，其中包括 600 个类，每个类至少 600 个视频，总共约 500,000 个视频（见底部 1）。动态视频多样且不受约束，这使我们能够训练大模型而不关心在固定对象以特定方式交互的小数据集上发生的过度拟合[18,9]。在先前的工作中，一直使用的最接近的数据集（就主题和复杂性而言）是 UCF-101 [46]。我们专注于 Kinetics-600，因为它的尺寸更大（视频比 UCF-101 多近 50 倍）并且其多样性增加（600 而不是 101 类 - 更不用说增加的类内多样性）。然而，为了与现有技术进行比较，我们在 UCF-101 上进行训练，并在那里设置了一个新的最先进的初始分数。 Kinetics 包含许多来自 YouTube 的工件，包括剪辑（如图 2a 所示），标题屏幕和视觉效果。除非特别描述，否则我们选择具有步幅 2 的帧（意味着我们跳过每隔一帧）。这使我们能够生成更高复杂度的视频，而不会产生额外的计算成本。

2.5 评估指标

设计衡量生成模型质量的指标（特别是 GAN）是一个活跃的研究领域[43,48]。在这项工作中，我们报告了两个最常用的指标，即初始分数（IS）[44]和 Fréchet 初始距离（FID）[21]。这些指标的标准实例化适用于生成图像模型，并使用初始模型[51]进行图像分类或特征提取。对于视频，我们使用在 Kinetics-600 上训练的公开的 Inflated 3D Convnet（I3D）网络[12]。 因此，我们的 Fréchet 初始距离非常类似于 Fréchet 视频距离（FVD）[56]，尽管我们的实现与原始 FID 度量不同且更加一致。（见底部 2）

3 Dual Video Discriminator GAN

我们的主要贡献是 Dual Video Discriminator GAN（DVD-GAN），它能够生成高分辨率和时间一致的视频。它将大型图像生成模型（Big-GAN [10]）扩展到视频，同时引入了几种加速训练的技术。 图 3 给出了 DVD-GAN 架构的概述，详细说明见附录 A.2。 与以前的一些工作不同，我们的生成器不包含前景，背景或运动（光流）的明确先验; 相反，我们依靠高容量神经网络以数据驱动的方式来学习。DVD-GAN 包含自注意机制和 RNN，但在时间或空间上不具有自回归性。 虽然 RNN 按顺序为每个帧生成特征，但之后所有帧都由 ResNet 并行生成，共同生成每个帧中的所有像素。换句话说，

[1] 有时会修剪动态帧，因此我们无法给出确切的大小。 数据集可在[2]获得。

[2] 我们默认使用'avgpool'功能（而不是 logits），我们的 I3D 模型在 Kinetics-600（而不是 Kinetics-400）上训练，我们预先计算整个训练集的完全真实统计数据。

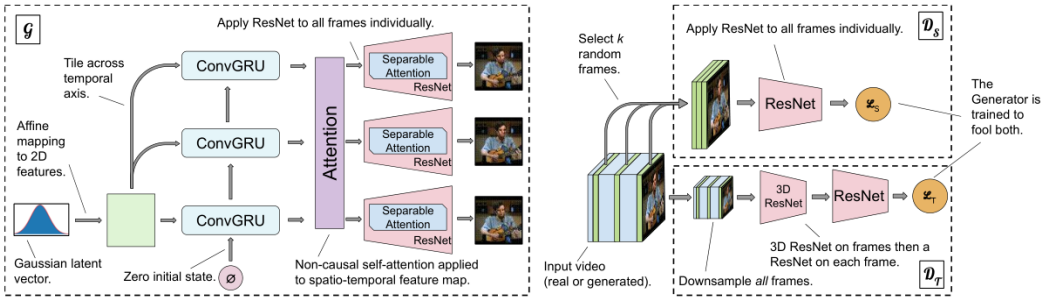

图 3：G（左）和 DS / DT（右）的简化架构图。 更多细节在 A.2 中。

每帧的像素不直接依赖于视频中的其他像素，如自回归模型的情况。

3.1 双重判别器

给定长度为 T，高度为 H 且宽度为 W 的视频，DVD-GAN 采用两个判别器进行评估：空间判别器 DS 和时间判别器 DT。 DS 通过随机采样 k 个全分辨率帧并单独处理它来评判单帧内容和结构。我们使用 k = 8 并在 4.3 节中显示出这个选择提高了性能。 DS 的最终得分是每帧得分的总和，类似于 TGANv2 [41]。

时间判别器 DT 必须向 G 提供学习信号以产生运动（不被 DS 评判的东西）。为了使模型可扩展，我们希望在不以全分辨率处理整个视频的情况下实现此目的。我们不是对批处理[41]进行二次采样或仅选择帧的子集[41,54]，而是将空间下采样功能 φ(·) 应用于整个视频并将其输出馈送到 DT。我们选择φ为 2×2 平均池化函数，但在 4.3 节讨论替代方案。这导致一种架构，其中判别器不处理整个高分辨率视频（因为 DS 仅处理 k×H×W 像素而 DT 仅处理 T×H/2×W/2），并且它们一起确保 G 学习生成高分辨率和时间上一致的视频。

DS 类似于 MoCoGAN 中的每帧判别器 DI [54]。然而，MoCoGAN 的 DT 模拟视频是全分辨率视频，而 DS 是 DVD-GAN 中高分辨率细节学习信号的唯一来源。出于这个原因，当φ不是固定函数时，DS 是必不可少的，不像在 MoCoGAN 中每帧判别器不那么重要。

3.2 可分离的自注意力机制

基于自注意力的 Transformer 模块[57]是一种流行的架构构件，其全局感受野允许在整个特征图上传播信息。然而，直接将其应用于巨大的视频特征是令人望而却步的，因为自注意力需要计算和存储大小为$(HWT)^2$的注意力权重矩阵。为了规避这种限制，我们引入了一种有效的自注意力变体，即我们称之为可分离自注意力。我们不是同时关注我们特征的所有位置，而是连续应用三个注意力层，每个注意力层接连关注在高度，宽度和时间轴上。这减小了我们需要存储在内存中的最大张量的大小，即将$(HWT)^2$成比例到$\max\{H^2WT, HW^2T, HWT^2\}$的大小。这可以看作是[14]中同时引入的分离注意力的一个特例。

4 实验和分析

我们的训练设置直接模仿 BigGAN [10]：详细说明见附录 A.3。每个 DVD-GAN 都在 TPUv3 pods 片上进行了训练[1]，使用 32 到 512 个副本与 Adam [27]优化器进行高达 300,000 个更新步骤（尽管我们通常在崩溃前的最终检查点评估模型 - 通常在 100k 和 250k 之间）。我们依靠 TF-Replicator [11]框架进行数据并行训练。模型花了 12 到 96 个小时进行训练。

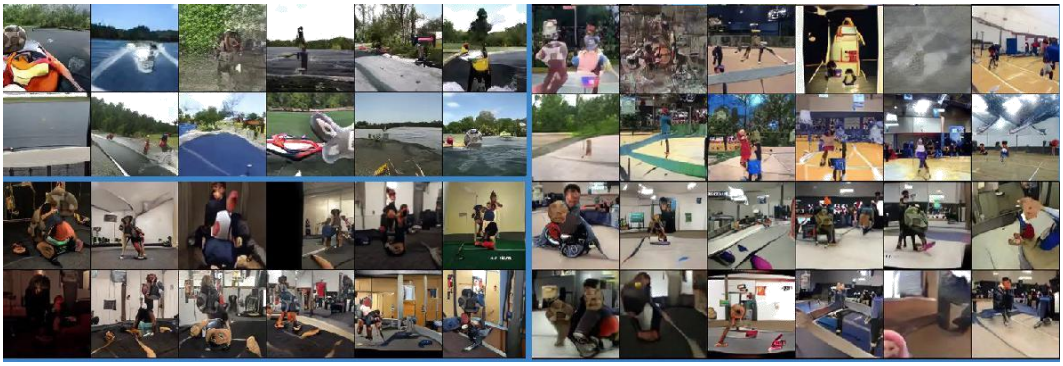

图 4：每个象限具有来自 12 个样本的初始帧，用于固定类。 从顺时针的左上方开始，类别包括：**水上摩托艇，打篮球，冰壶（运动），抓举举重。**

图 5：来自足球类的 64×64 样本的所有 48 帧（以栅格扫描顺序）。

4.1 类条件视频合成

我们的主要结果涉及类条件视频合成的问题。 我们提供 UCF-101 和 Kinetics-600 数据集的结果。随着 Kinetics-600 成为生成视频建模的新基准，我们的结果为未来的工作奠定了坚实的基础。

4.1.1 Kinetics -600 结果

表 1：用于 Kinetics-600 视频合成的 DVD-GAN 的 FID / IS。

(# Frames / Resolution)	FID (↓)	IS (↑)	
		No Truncation	With Truncation
$12/64 \times 64$	0.91	61.10	129.9
$48/64 \times 64$	12.92	97.62	219.05
$12/128 \times 128$	2.16	55.09	80.32
$48/128 \times 128$	31.50	111.19	222.07
$12/256 \times 256$	3.35	59.74	64.05

在表 1 中，我们展示了本文的主要结果：Kinetics-600 上的视频合成基准。我们考虑一系列分辨率和视频长度，并测量每个分辨率和 Fréchet 初始距离（FID）（如第 2.5 节所述）。以前没有工作可以定量比较这些结果（比较实验见 4.1.2 节和 4.2.1 节），但我们认为这些样品显示的数据集中尚未达到保真度与 Kinetics-600 一样复杂（参见附录 D.1 中每行的样本）的水平。因为所有视频都针对 I3D 网络调整大小（至 224×224），所以在不同分辨率下比较相等长度视频的指标是有意义的。IS 和 FID 在不同长度的视频中都不具有可比性，应该作为单独的指标来对待。

生成更长和更大的视频是一个更具挑战性的建模问题，这些都能用指标体现出来（特别是，比较 64×64，128×128 和 256×256 分辨率的 12 帧视频）。尽管如此，DVD-GAN 能够在所有分辨率下生成合理的视频，并且动作最长可达 4 秒（48 帧）。如附录 D.1 所示，显示出较小的视频具有

图 6：来自不同类别的 128x128 动态视频的单个生成帧。

表 2：UCF-101 上的 IS（越高越好）

Method	IS (↑)
VGAN [59]	8.31 ± .09
TGAN [42]	11.85 ± .07
MoCoGAN [54]	12.42 ± .03
ProgressiveVGAN [3]	14.56 ± .05
TGANv2 [41]	24.34 ± .35
DVD-GAN (ours)	**32.97** ± 1.7

表 3：DVD-GAN-FP 在没有跳帧的 16 帧 Kinetics-600 的视频预测中的 FVD 分数。 最后一行表示视频合成模型，生成 16 帧而不跳帧。

Method	Training Set FVD (↓)	Test Set FVD (↓)
Video Transformer [62]	-	170 ± 5
DVD-GAN-*FP*	**99.32 ± 0.55**	**103.78 ± 1.17**
DVD-GAN	32.3 ± 0.82	31.1 ± 0.56

高品质的纹理，物体组成和运动。在更高的分辨率下，生成相干对象变得更加困难（此时移动由更多数量的像素组成），但生成的场景的高级细节仍然非常连贯，并且纹理（甚至是复杂的，如溜冰场的一侧见图 1a）生成良好。值得注意的是，48 帧模型没有看到比 12 帧模型更高分辨率的帧（由于第 3.1 节中描述的 k = 8 的固定选择），但仍然学会在所有帧中生成高分辨率图像帧。

4.1.2 UCF-101 上的视频合成

我们通过在 UCF-101 [46]上测试相同的模型来进一步验证我们的结果，这是一个较小的数据集，包括 101 个类别中 13,320 个人类行为视频，这些视频以前曾被用于视频合成和预测[42,41,54]。我们的模型生成 IS 为 32.97 的样本，明显优于现有技术（定量比较见表 2，更多细节见附录 B.2）。

4.2 未来视频预测

未来视频预测是生成一系列帧的问题，这些帧直接来自一个（或多个）初始条件帧。这个和视频合成都需要 G 学习产生逼真的场景和时间动态，然而视频预测还需要 G 来分析视频并发现场景中随时间演变的元素。在本节中，我们使用在 Kinetics-400 上训练的 I3D 网络的 logits 作为特征，完全按照[56]使用 Fréchet 视频距离（FVD）。这允许与先前的工作直接比较。我们的模型 DVD-GAN-FP（帧预测）略有修改以适应评价改变后的新问题，这些变化的细节在附录 A.4 中给出。

4.2.1 条件帧 Kinetics

为了与自回归视频模型的并行工作直接比较[62]，我们考虑在 64×64 分辨率下生成 11 帧 Kinetics-600，条件为 5 帧，其中用于训练的视频不带任何跳帧。我们在表 3 中显示了所有这些情况的结果。我们的帧条件模型 DVD-GAN-FP 优于（有限的）先前关于 Kinetics 的帧条件预测的工作。标记为 DVD-GAN 的最后一行是 16 帧的 FVD

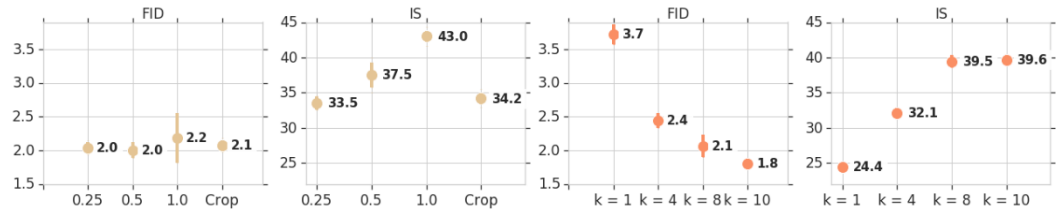

图 7：DT（左二）和 DS（右二）中的影响。 FID 与任何选择类似，而 IS 随着下采样的增加而下降。 增加 k 会带来提高，同时伴随着收益递减。

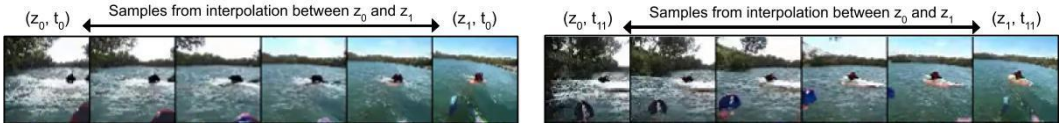

图 8：在第一（左 6）和最后（右 6）帧的共享类下的潜在插值。

视频合成样本（没有调节帧）没有任何跳帧，这是明显更好的。这意味着合成模型能够从头开始生成质量更高的视频，而不是必须推断出合理的视频延续的模型。

4.2.2 BAIR Robot Pushing

我们在 BAIR Robot Pushing [18]数据集上进一步评估了 DVD-GAN-FP，这是一个操纵一系列小物体的机器人手臂静态视频的数据集。在这里，我们的模型通过先前的工作获得了具有竞争力的 FVD，但是没有达到最先进的性能。更多细节见附录 B.1。

4.3 双判别器输入

我们分析 k 的几个选择（DS 输入中每个样本的帧数）和 φ（DT 的下采样功能）。我们期望设置 φ 为恒定或 k = T 以产生最佳模型，但我们感兴趣的是最大压缩 k 和 φ 并且减少判别器输入大小（和计算量），同时仍然产生高质量的生成器。对于 φ，我们考虑：2×2 和 4×4 平均池化，恒定（没有下采样），以及 φ 采用输入视频的随机半尺寸裁剪（如[41]中所述）。结果可以在图 7 中看到。对于每次消融，我们在 Kinetics-600 的 12 帧剪辑上以 64×64 分辨率训练三个相同的 DVD-GAN 100,000 步，其具有不同的随机初始化。我们报告整个训练期间每组的平均值和标准差（通过误差线）。对于 k，我们考虑 1,2,8 和 10 帧。

4.4 截断曲线和插值

我们期望 G 从中心附近的潜伏点或分布均值（零）生成更高质量的样本。这是 Truncation Trick 背后的想法[10]。与 BigGAN 一样，我们发现 DVD-GAN 可以截断。我们还尝试在潜在空间（图 8）和类嵌入（图 9）中进行插值（另见附录 D.2）。在这两种情况下，插值都证明 G 已经学习了从潜在空间到真实视频的相对平滑的映射：对于仅记忆训练数据的网络来说，这是不可能的，或者每个类只能生成一些样本的网络。请注意，虽然沿着插值的所有潜在向量都是有效的（因此 G 应该产生合理的样本），但是在训练过程中 G 被要求在两个类之间产生一个样本。然而，G 能够在甚至非常不同的类之间进行插值。

5 结论

我们通过引入能够捕获大型视频数据集的复杂性的 GAN 来解决建模自然视频的具有挑战性的问题。我们展示了在 UCF-101 和帧条件 Kinetics-600 上，它定量地实现了新的技术水平，同时定性地生成具有高复杂性和多样性的样本视频。我们还希望强调在大型复杂视频数据集上训练生成模型的好处，例如 Kinetics-600。我们设想我们在此数据集上建立的强大基线与 DVD-GAN 将被生成建模社区用作参考点。虽然在不受约束的环境中可以始终如一地生成逼真的视频，但仍有许多工作要做，我们相信 DVD-GAN 是朝这个方向迈出的一步。

致谢

我们要感谢 Eric Noland 和 JoãoCarreira 对 Kinetics 数据集以及 Marcin Michalski 和 Karol Kurach 的帮助，帮助我们获取 Fréchet Video Distance 比较的数据和模型。我们还要感谢 Sander Dieleman，Jacob Walker 和 Tim Harley 对论文的有益讨论和反馈。

参考文献

[1] Cloud TPU. https://cloud.google.com/tpu/. Accessed: 2019.

[2] Kinetics. https://deepmind.com/research/open-source/open-source-datasets/kinetics/. Accessed: 2019.

[3] D. Acharya, Z. Huang, D. P. Paudel, and L. Van Gool. Towards high resolution video generation with progressive growing of sliced Wasserstein GANs. arXiv:1810.02419, 2018.

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. arXiv:1701.07875, 2017.

[5] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine. Stochastic variational video prediction. ICLR, 2018.

[6] N. Ballas, L. Yao, C. Pal, and A. Courville. Delving deeper into convolutional networks for learning video representations. arXiv:1511.06432, 2015.

[7] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh. Recycle-GAN: Unsupervised video retargeting. In ECCV, September 2018.

[8] S. Barratt and R. Sharma. A note on the Inception Score. arXiv:1801.01973, 2018.

[9] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In ICCV, 2005.

[10] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. ICLR, 2019.

[11] P. Buchlovsky, D. Budden, D. Grewe, C. Jones, J. Aslanides, F. Besse, A. Brock, A. Clark, S. G. Col-menarejo, A. Pope, et al. TF-Replicator: Distributed machine learning for researchers. arXiv:1902.00465, 2019.

图 9：来自两个不同类别（前两行）的相同 z 生成的两个视频的帧，以及它们在单个选定帧（底行）之间的插值。

[12] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In ICCV, 2017.

[13] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman. A short note about Kinetics-600. arXiv:1808.01340, 2018.

[14] R. Child, S. Gray, A. Radford, and I. Sutskever. Generating long sequences with sparse transformers. arXiv:1904.10509, 2019.

[15] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville. Modulating early visual processing by language. In NeurIPS, 2017.

[16] E. Denton and R. Fergus. Stochastic video generation with a learned prior. ICML, 2018.

[17] V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. In ICLR, 2017.

[18] F. Ebert, C. Finn, A. X. Lee, and S. Levine. Self-supervised visual planning with temporal skip connections. arXiv:1710.05268, 2017.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In NeurIPS, 2014.

[20] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In NeurIPS, 2017.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In NeurIPS, 2017.

[22] J.-T. Hsieh, B. Liu, D.-A. Huang, L. Fei-Fei, and J. C. Niebles. Learning to decompose and disentangle representations for video prediction. In NeurIPS. 2018.

[23] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv:1502.03167, 2015.

[24] N. Kalchbrenner, A. van den Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu. Video pixel networks. In ICML, 2017.

[25] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. arXiv:1812.04948, 2018.

[26] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The Kinetics human action video dataset. arXiv:1705.06950, 2017.

[27] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.

[28] D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In NeurIPS, 2018.

[29] M. Kumar, M. Babaeizadeh, D. Erhan, C. Finn, S. Levine, L. Dinh, and D. Kingma. VideoFlow: A flow-based generative model for video. arXiv:1903.01434, 2019.

[30] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine. Stochastic adversarial video prediction. arXiv:1804.01523, 2018.

[31] J. Lei Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. arXiv:1607.06450, 2016.

[32] J. H. Lim and J. C. Ye. Geometric GAN. arXiv:1705.02894, 2017.

[33] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. arXiv:1511.05440, 2015.

[34] J. Menick and N. Kalchbrenner. Generating high fidelity images with subscale pixel networks and multidimensional upscaling. In ICLR, 2019.

[35] T. Miyato and M. Koyama. cGANs with projection discriminator. arXiv:1802.05637, 2018.

[36] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. arXiv:1802.05957, 2018.

[37] K. Ohnishi, S. Yamamoto, Y. Ushiku, and T. Harada. Hierarchical video generation from orthogonal information: Optical flow and texture. In AAAI, 2018.

[38] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra. Video (language) modeling: a baseline for generative models of natural videos. arXiv:1412.6604, 2014.

[39] A. Razavi, A. van den Oord, and O. Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. arXiv preprint arXiv:1906.00446, 2019.

[40] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In MICCAI, 2015.

[41] M. Saito and S. Saito. TGANv2: Efficient training of large models for video generation with multiple subsampling layers. arXiv:1811.09245, 2018.

[42] M. Saito, E. Matsumoto, and S. Saito. Temporal generative adversarial nets with singular value clipping. In ICCV, 2017.

[43] M. S. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall. In NeurIPS, 2018.

[44] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In NeurIPS, 2016.

[45] A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv:1312.6120, 2013.

[46] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402, 2012.

[47] C. Spampinato, S. Palazzo, P. D'Oro, F. Murabito, D. Giordano, and M. Shah. VOS-GAN: Adversarial learning of visual-temporal dynamics for unsupervised dense prediction in videos. arXiv:1803.09092, 2018.

[48] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using LSTMs. In ICML, 2015.

[49] X. Sun, H. Xu, and K. Saenko. A two-stream variational adversarial network for video generation. arXiv:1812.01037, 2018.

[50] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In ICML, 2011.

[51] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception architecture for computer vision. In CVPR, 2016.

[52] S. Tokui, K. Oono, S. Hido, and J. Clayton. Chainer: a next-generation open source framework for deep learning. In Workshop on Systems for ML and Open Source Software at NeurIPS, 2015.

[53] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In ICCV, 2015.

[54] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz. MoCoGAN: Decomposing motion and content for video generation. In CVPR, 2018.

[55] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. arXiv:1607.08022, 2016.

[56] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards accurate generative models of video: A new metric & challenges. arXiv:1812.01717, 2018.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In NeurIPS, 2017.

[58] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. Learning to generate long-term future via hierarchical prediction. In ICML, 2017.

[59] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In NeurIPS, 2016.

[60] J. Walker, K. Marino, A. Gupta, and M. Hebert. The pose knows: Video forecasting by generating pose futures. In ICCV, 2017.

[61] T. Wang, M. Liu, J. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. arXiv:1808.06601, 2018.

[62] D. Weissenborn, O. Täckström, and J. Uszkoreit. Scaling autoregressive video models. arXiv:1906.02634, 2019.

[63] Y. Wu and K. He. Group normalization. In ECCV, 2018.

[64] Y. Wu, S. Zhang, Y. Zhang, Y. Bengio, and R. R. Salakhutdinov. On multiplicative integration with recurrent neural networks. In NeurIPS, 2016.

[65] Y. Xie, E. Franz, M. Chu, and N. Thuerey. tempoGAN: A temporally coherent, volumetric GAN for super-resolution fluid flow. TOG, 2018.

[66] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin. Pose guided human video generation. In ECCV, 2018.

[67] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. arXiv:1805.08318, 2018.

[68] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg. Dance dance generation: Motion transfer for internet videos. arXiv:1904.00129, 2019.

A 实验方法论

A.1 数据集处理

对于所有数据集，我们独立地随机混合每个模型副本的训练集。 BAIR Robot Pushing 数据集的实验以 64×64 的原始分辨率进行，对于 UCF-101，我们以（下采样）128×128 分辨率进行操作。这是通过双线性调整大小来完成的，使得视频的最小尺寸被映射到 128 个像素（保持纵横比）。从这里我们沿着另一个维度采取随机的 128-像素裁剪。我们使用相同的程序为 Kinetics-600 构建不同分辨率的数据集。所有三个数据集都包含我们生成的具有更多帧的视频，因此我们从调整大小的输出中获取连续帧的随机序列。

A.2 架构描述

我们的模型采用了 BigGAN [10]的许多架构选择，包括我们用于描述网络宽度的命名法，网络宽度由通道乘法器 ch 的乘积和网络中每层的常数确定。 G 的分层常数是[8, 8, 8, 4, 2] 对于 64 × 64 视频和[16, 16, 8, 4, 2, 1]对于 128 × 128 视频（与 BigGAN 相同）。第 i 层的宽度由 ch 和第 i 个常数的乘积给出，并且在 G 中的残差网络之前的所有层使用初始层的乘数，并且我们将其和 ch 的乘积称为ch_0。用于 D 的 ch 总是 128，其中对于 G，在较小的模型中为 128，对于大于 64×64 分辨率的模型为 64，它或许会生成 48 帧。 DT 和 DS 的相应 ch 列表是[2, 4, 8, 16, 16] 对于 64×64 分辨率和[1, 2, 4, 8, 16, 16]对于 128×128 分辨率。

G 的输入由高斯潜在噪声$z \sim \mathcal{N}(0, I)$和所需类 y 的学习嵌入 e(y)组成。两个输入都是 120 维向量。 G 首先计算$[z; e(y)]$到$[4, 4, ch_0]$形状的张量的仿射变换，这被视作每个 T 时间步的递归神经网络的输入。在我们的大多数实验中，我们在 G 中使用乘法卷积门控递归单元(Multiplicative Convolutional Gated Recurrent Unit)[6,64,50]，其输入x_t和先前输出h_{t-1}的更新规则由下式给出：

$$
\begin{aligned}
g &= (W_{gh} \star_1 h_{t-1} + b_{gh}) \odot (W_{gx} \star_1 x_t + b_{gx}) \\
r &= \sigma(W_r \star_3 [g; x_t] + b_r) \\
u &= \sigma(W_u \star_3 [g; x_t] + b_u) \\
c &= \rho(W_c \star_3 [x_t; r \odot g] + b_c) \\
h_t &= u \odot g + (1 - u) \odot c
\end{aligned}
$$

在这些方程中，σ 和 ρ 分别表示逐元素的 sigmoid 和 ReLU 函数，$\star_n$运算符表示一个具有核的大小为 n ×n 的卷积运算，运算符$\odot$是一个逐元素

乘法。括号用于表示要素串联。我们发现，循环网络架构的选择对样本质量的影响不可忽略。对于我们的大多数实验，我们生成的帧少于 48 帧，并且在该设置中，乘法卷积门控递归单元表现最佳。但是，对于 48 帧的视频，标准的 ConvGRU（其中$g = h_{t-1}$）表现更好。我们相信，G 的 RNN 中的进一步架构改进应该能够在相同的架构下实现所有长度的高质量视频建模，并且我们将其留待未来的工作。

RNN 的结果是形状特征$[T, 4, 4, ch_0]$作为非因果自注意力区块的输入[57]。它有一个注意力头，不可分离（如 3.2 节所述），因为它可以在少量特征上运行。最后，每个帧由关注点输出的相应时间片产生，该残差网络几乎与 BigGAN 的网络相同;虽然遵循 BigGAN-deep，我们将整个块的数量加倍并传递整个条件向量$[z; e(y)]$到每个块（没有分层分割）。残差网络单独应用于每个帧（即，在前向通过之前将时间轴折叠到批量轴中），但是在计算批量标准化统计时，我们不会减少时间维度。这可以防止网络利用批量标准化层在时间步之间传递信息。残差网络在很大程度上独立地对每个帧进行升级，除了自注意力层，其遵循 BigGAN，被放置在最终残余块之前。由于张量大，我们使用 3.2 节中描述的可分离自注意块。

空间判别器 DS 的功能几乎与 BigGAN 的判别器相同。计算每个均匀采样的 k 帧的分数（我们默认为 k = 8），DS 输出是每帧分数的总和。时间判别器 DT 具有类似的架构，但是使用 2×2 平均池化下采样功能预处理实际或生成的视频。此外，DT 的前两个残余块是 3-D，其中每个卷积被 3-D 卷积替换，内核大小为 3×3×3。其余的架构遵循 BigGAN [10]。

A.3 训练细节

从 DVD-GAN 采样是非常有效的，因为生成器架构的核心是前馈卷积网络：在单个 TPU 核心上，可以在不到 150ms 的时间内对两个 64×64 - 48 帧视频进行采样。对于 G [21]的每次更新，双判别器 D 被更新两次，并且我们对所有权重层（由第一奇异值近似）和权重的正交初始化[45]使用谱归一化[67]。使用 G 的权重的指数移动平均值进行采样，其在 20,000 个训练步骤之后开始以衰减率 γ = 0.9999 累积衰减。使用 Adam [27]优化模型，批量大小为 512，G 和 D 的学习率分别为$1 \cdot 10^{-4}$和$2 \cdot 10^{-4}$。 D [35]中的类调节是基于投影的，而 G 依赖于类条件批量标准化[23,15,17]：相当于标准的批量标准化，没有学习的放缩和偏移，接着是元素仿射变换，其中每个参数是噪声向量和类调节的功能。

A.4 视频预测的架构扩展

为了提供未来视频预测问题的结果，我们描述了对 DVD-GAN 的简单修改，以便于增加调节能力。扩展模型图如图 10 所示。

给定 C 个条件帧$\{f_1, ..., f_C\}$，我们修改的 DVD-GAN-FP 通过与 DS 相同的深度残差网络分别通过每个帧。每个条件框架的结果特征在通道维度中连接，并且 1×1 卷积将通道数目减少到 512。此网络的输出的空间维度与 G 的 RNN 中的循环状态的形状完全匹配，因此我们通过此 ResNet 的输出为初始状态。DT 和 DS 都在条件帧和 G 的输出的串联上运行，这意味着判别器不接收任何额外详细信息以认为第一个 C 帧是特殊。为了进一步方便从帧条件中保留小细节，我们在处理调节框架和 G 的中间特征的残余块的中间输出之间添加 U-Net [40]样式跳过连接。准确地说，来自调节帧网络的第 i 个块与 G 的残差网络的第 i 个块连接，并且所得到的特征通过单个 3×3 卷积传递以将通道数量压缩回正常

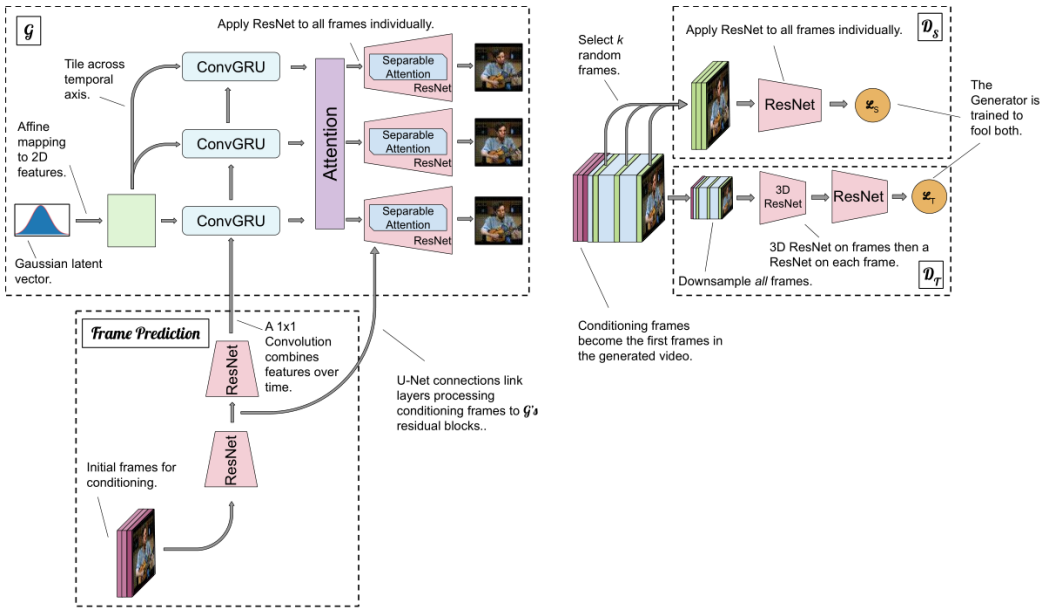

图 10：描述帧条件模型变化的架构图。

在继续下一个区块之前的大小。 最后，我们的视频预测变体不会对任何类信息进行条件限制，从而允许我们直接与现有技术进行比较。

B 进一步的实验

B.1 BAIR Robot Pushing

我们在单级 BAIR Robot Pushing Dataset [18]上测试未来的视频预测，这是一个围绕一组变化的物体移动的机器人手臂的静态视频数据集。 为了与[56]中报告的先前结果（对于 SAVP 模型[30]）和[62]直接比较，我们考虑在单个起始帧上生成 15 帧。 结果报告在表 4 中。我们发现 DVD-GAN-FP 与之前的对抗方法相比具有竞争力，但比自回归模型的并行结果更差[62]。

表 4：BAIR 机器人推动的 FVD（越低越好）。视频转变的结果来自[62]，所有其他结果如[56]中所述。

Method	FVD
SV2P [5]	262.5
SAVP [30]	116.4
Video Transformer [62]	**94** ± **2**
DVD-GAN-FP (ours)	127.1

B.2 UCF-101

UCF-101 [46]是一个包含 101 个类别的 13,320 个人类视频的数据集，这些视频以前用于视频合成和预测[42,41,54]。 我们报告使用 C3D 网络计算的初始分数（IS）[53]，用于与先前工作的定量比较（见底部 3）。我们的模型产生 IS 为 32.97 的样本，显著优于现有技术（见表 2）。 UCF-101 上的 DVD-GAN 架构与用于 Kinetics 的模型相同，并且受到来自 UCF-101 的 16 帧 128×128 剪辑片段的训练。

然而，值得一提的是，我们的得分改进至少部分是由于记忆训练数据。在图 11 中，我们展示了来自我们最好的 UCF-101 模型的插值样本。 如图 8 所示，我们对 2 个潜伏点（左侧和最右侧列）进行采样，并显示线性样本

[3] 我们使用 Chainer [52]实现 C3D 的初始分数，代码在： https://github.com/pfnet-research/tgan.

图 11：UCF-101 样本之间的第一帧插值。每行是一个单独的插值。与图 8 和附录 D.2 对比。

在每行的潜在空间中插值。这里我们展示 4 个这样的插值（每个视频的第一帧）。与从一个样本平滑过渡到另一个样本的图 8 不同，我们看到高度不同的样本之间的潜在空间突然跳跃，并且每组样本之间的视频内分集很少。可以进一步看出，一些生成的样本与来自训练集的样本高度相关。

我们将此显示为初始分数度量标准的失败，UCF-101 上类条件视频合成的常见报告值，但也表明 UCF-101 不是一个复杂或多样化的数据集，以促进有趣的视频生成。每个类都相对较小，并且重用来自共享底层视频的剪辑意味着类内多样性可以限制为每个类的少数几个视频。这表明需要更大，更多样化和具有挑战性的数据集用于生成视频建模，我们相信 Kinetics-600 为此任务提供了更好的基准。

C 杂项实验

在这里，我们详细介绍了我们试验过的一些修改或杂项结果，但未产生结论性结果。

- 我们尝试了几种标准化的变体，它们不需要计算一批数据的统计数据。组标准化[63]表现最佳，几乎与批量标准化相当（但差于）。 我们进一步尝试了层标准化[31]，实例标准化[55]，没有标准化，但发现这些明显落后于批量标准化。

- 我们发现删除 G 中的最终批量标准化，这发生在 ResNet 之后和最终卷积之前，导致了学习中的灾难性失败。有趣的是，只是删除 G 残差区块内的批量标准化层仍然导致良好（虽然稍差）的生成模型。 特别地，在残余块中没有批量标准化的变体通常实现明显更高的 IS（对于 64×64 帧样本高达 110.05 - 正常的两倍）。 但是这些模型的 FID 分数差得多（上述模型为 1.22） - 并且产生质量较差的视频样本。

- DVD-GAN 的早期变体包含批量标准化，其在所有批次元素的所有帧上标准化。 这为 G 提供了一个跨时间传递信息的额外渠道。它利用了这一点，结果是一个需要批量统计以产生良好样本的模型。我们发现，在时间步长上标准化的版本也可以独立工作，而且不依赖于统计数据。

- 基于 BigGAN 的残差块的模型训练得更快（在挂钟时间内）但在指标方面更慢，并且努力达到基于 BigGAN 残差块的模型的准确性。

D 生成样本

通过静止帧准确地传送复杂的生成视频是比较难的。如果提供，我们建议读者通过提供的链接查看生成的视频。我们通过行/列号引用这些批次中的视频，其中第 0 行和第 0 列中的视频位于左上角。

D.1 生成样本

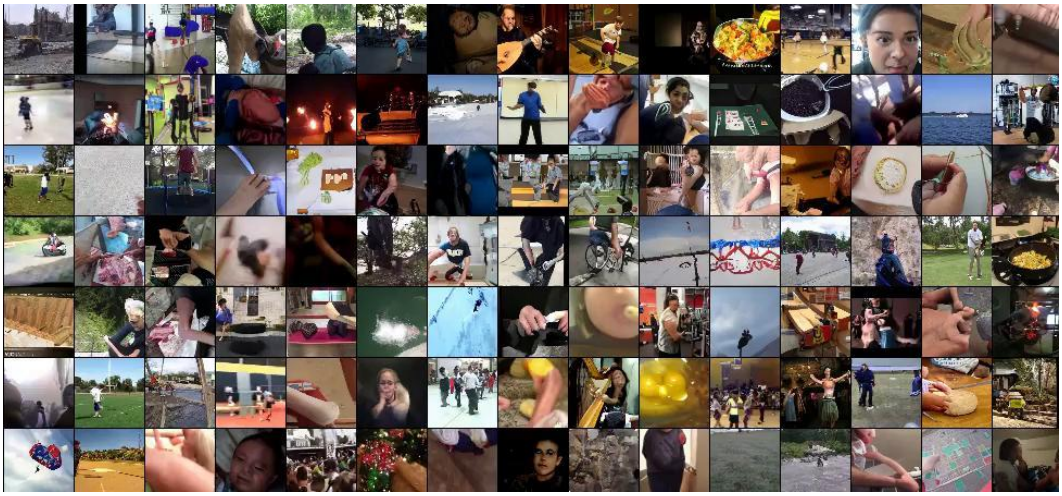

图 12：来自 DVD-GAN 的随机批次样本的第一帧在 12 帧 64×64 Kinetics-600 上训练。 全部样本在：https://drive.google.com/open?id=1YJtaQgVDnt_r35xKghelgd4V8Po-Ueaz。

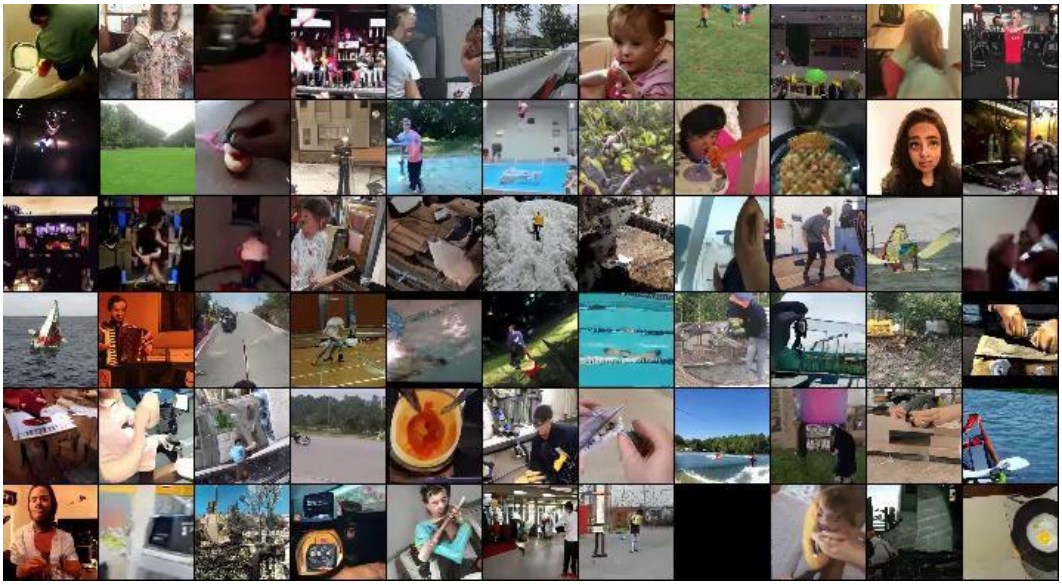

图 13：来自 DVD-GAN 的随机批次样本的第一帧在 48 帧 64×64 Kinetics-600 上训练。 全部样本在：https://drive.google.com/open?id=18pcN8W1AH-lVbGMCrR5VzllXsqbEchL0。第 1 行第 1 列中的视频展示了 DVD-GAN 能够记住“屏幕外”生成细节的能力，而第 1 行第 5 列中的视频显示了潜水员与潜水员之间复杂的因果关系以及水花的飞溅。

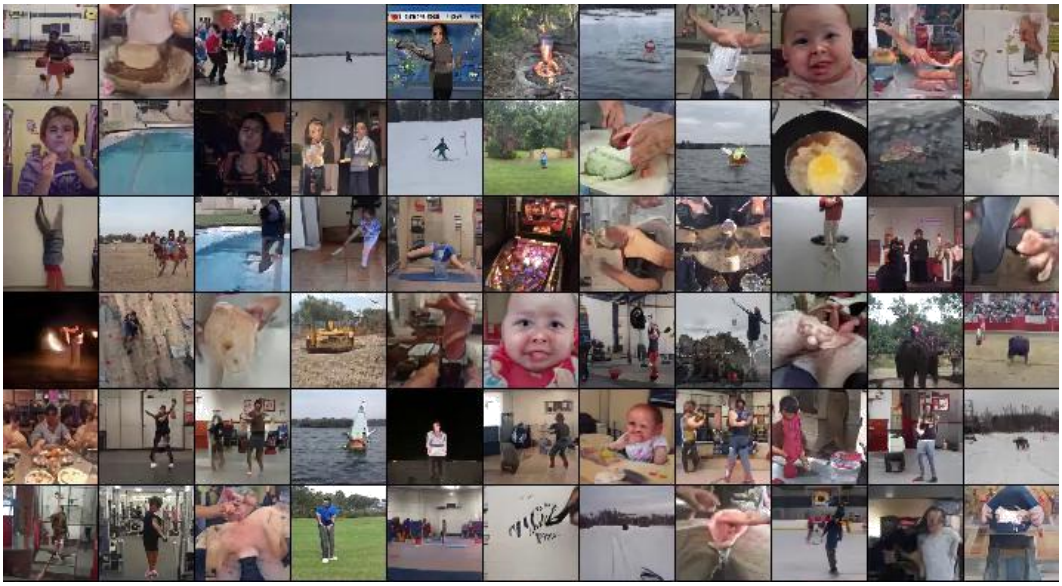

图 14：来自 DVD-GAN 的随机批次截断样本的第一帧在 48 帧 64×64 Kinetics-600 上训练。 全部样本在：https://drive.google.com/open?id=1XbqdFD70JrWSk0sW4v31QCQexpjxlALB。这些样本取自σ = 0.44 的高斯分布。

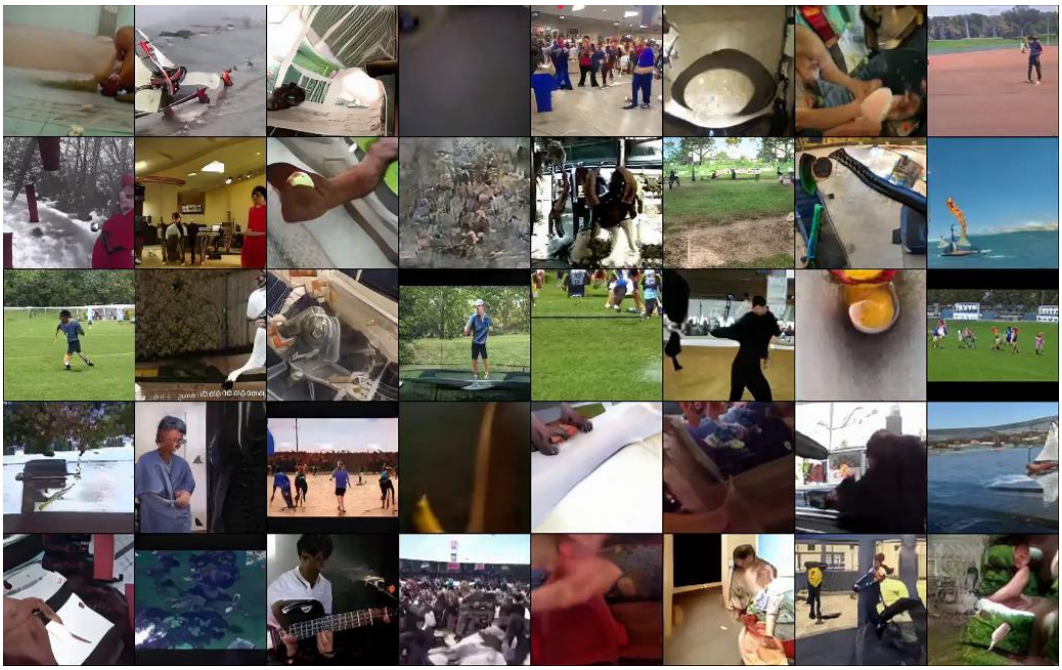

图 15：来自 DVD-GAN 的随机批次样本的第一帧在 12 帧 128×128 Kinetics-600 上训练。 全部样本在：https://drive.google.com/open?id=15MkrAkP3B9U4n12CgWUdlzr-YR0EDRSk

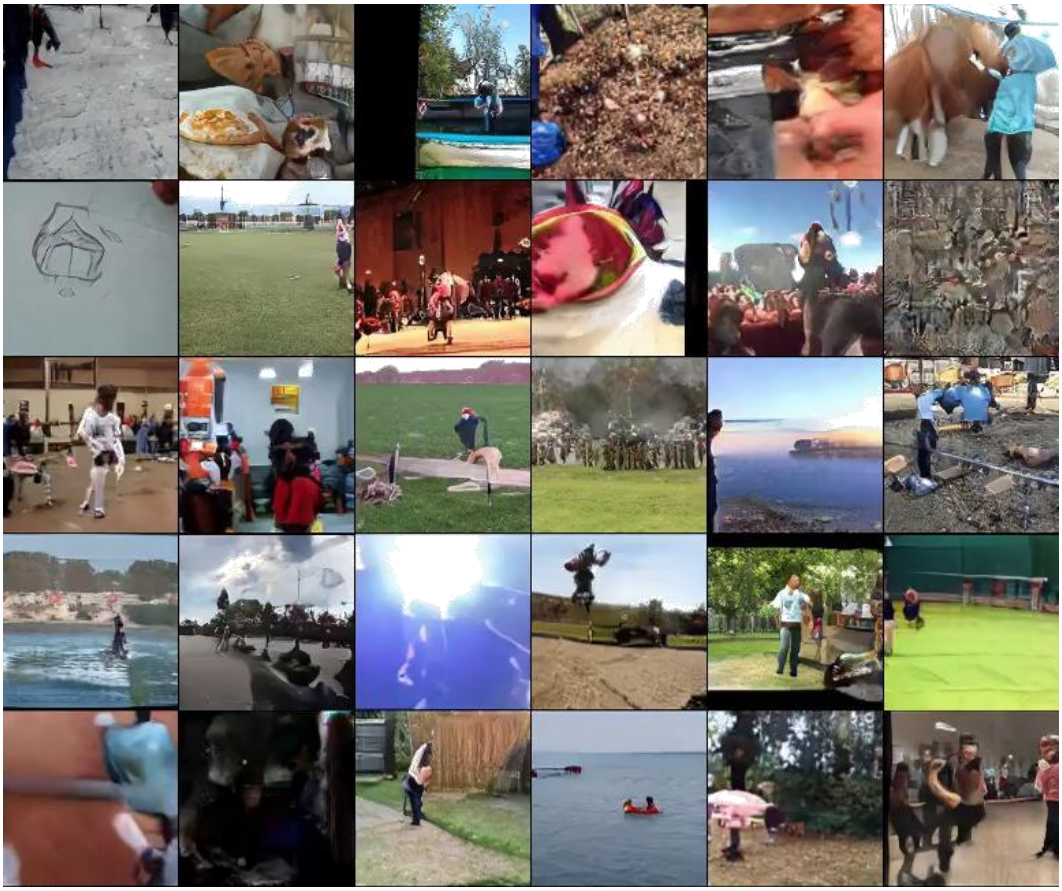

图 16：来自 DVD-GAN 的随机批次样本的第一帧在 48 帧 128×128 Kinetics-600 上训练。 全部样本在：https://drive.google.com/open?id=19kXShENC-7KC-VjkIR3GixcdLVgSGSW5。第 1 行第 5 列中的样本是偶尔由 DVD-GAN 生成的简并样本的定型实例。

图 17：来自 DVD-GAN 的随机批次样本的第一帧在 12 帧 256×256 Kinetics-600 上训练。 全部样本在：https://drive.google.com/open?id=1wagcMpBAnIfYSEgnOoAbEJoqmHTnrpcr.

D.2 插值样本

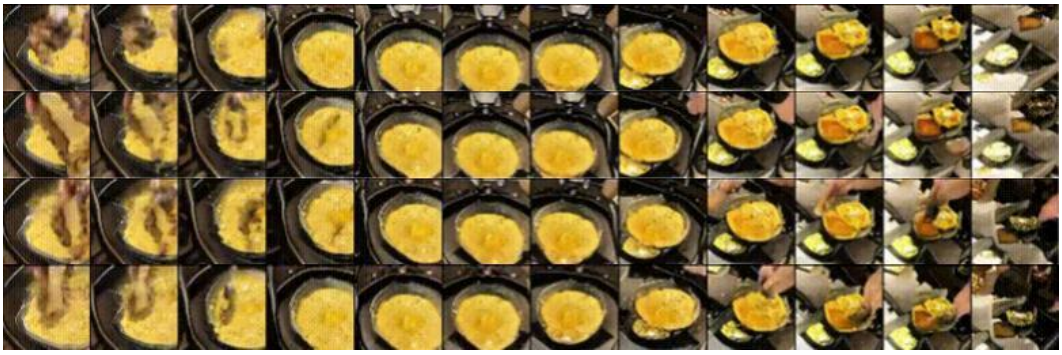

图 18：示例类内插值。 每列是一个单独的视频（垂直轴是时间维度）。左侧和最右侧列是随机采样的潜在向量，并在共享类下生成。其间的列表示在两个随机样本之间的线性插值下在相同类下生成的视频。 请注意此处显示的所有四个时间步的视频之间的平滑过渡。

图 19：类内插值的另一个例子。

图 20：类插值的示例。和以前一样，每列都是单个视频的一系列时间步长。在这里，我们采样一个潜在的向量，左侧和最右侧的列表示在两个不同的类（游泳和火舞）下生成潜在的视频。其间的列表示通过类嵌入的插值生成的相同潜在的视频。 尽管 DVD-GAN 在内插类中没有对数据进行过训练，但它仍能产生合理的样本。