23.11.20~23.11.26周报

一：本周工作

- 读论文PVDM
- 学习基础课程

二：下周计划

- 继续读Difussion相关的论文
- 学习基础课程

三：结论

- PVDM采用latent space来帮助diffusion模型更好地视频生成

四：详细工作

4.1 概述

图像生成模型在近些年取得了极大的突破，但与之相近的视频生成模型却进展缓慢。主要的原因还是在于，视频的高维度和时空连续性等问题限制了生成产品的质量。现行的各种方案，即便取得了较为不错的效果，也仍旧受限于计算和内存的低效率。而近期在图像领域所出现的latent diffusion模型，就很好的解决了这些问题，而本文首先将这一方式引入了图像生成领域，并取得了较之其他模型更好的效果。

4.2 相关工作

视频生成方面

视频生成模型也是一个长期存在着的科研目标，其目前的实现思路主要有三种：

- 采用GAN进行生成，但GAN经常要面对模式崩溃(mode collapse)等问题，并且这一方式也很难扩展到复杂且巨大的视频数据集上。
- Transformers也是一个常用的生成模型，其生成方式较之GAN更为优秀，但在生成长视频时也会需要昂贵的计算和存储消耗

- 最后时采用diffusion模型进行生成，可以得到最为优秀的输出，但也会受到计算和存储低效率的限制。本文就是采用低维的隐藏空间来解决diffusion的这一瓶颈。

diffusion model

difusion在多个领域都可以取得较好的效果，在视觉领域也比GAN效果更加，也可以更好应对零次训练的情况。但因为要在高维的数据空间中进行迭代，diffusion也面对着严重的计算低效问题。为了解决这一问题，低维的latent space方式被提出。我们也是通过这一方法来创建一个视频的diffusion 生成模型。

diffusion models for videos

在图片领域实现突破之后，视频领域的工作也得到了许多人重视。但尽管有着较为惊艳的输出，但通向高质量长视频的道路依旧被计算和内存的低效所阻碍。本文来解决了这一限制。

Triplane representation

一些最近的研究表明，高维的3D模型可以被高效的用2维的triplane 隐藏表示，并不改变encoding质量。之前的研究探索了这一方式在GAN中的应用，而本文则探索了其在diffusion中的应用。

4.3 技术实现

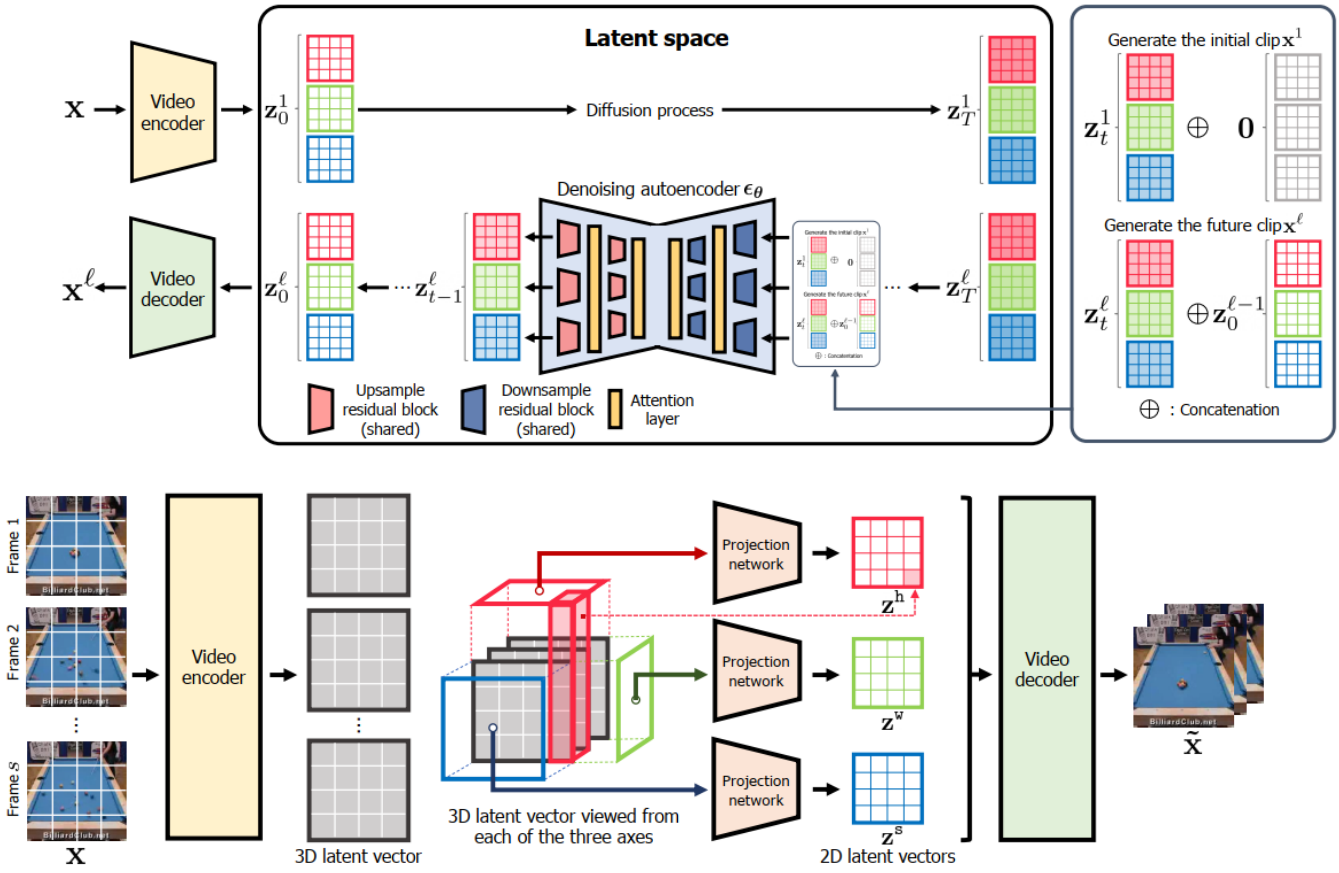

Figure 2. Detailed illustration of our autoencoder architecture in PVDM framework ((a) in Figure 1).

4.4 实验结果

采用两个典型的数据集进行训练：UCF–101 和 Sky–Timelapse

采用两种评分方式：FVD 和 Inception score(IS)

Method	UCF-101		SkyTimelapse	
	$FVD_{16}\downarrow$	$FVD_{128}\downarrow$	$FVD_{16}\downarrow$	$FVD_{128}\downarrow$
VideoGPT [65]	2880.6	N/A	222.7	N/A
MoCoGAN [57]	2886.8	3679.0	206.6	575.9
+ StyleGAN2 [28]	1821.4	2311.3	85.88	272.8
MoCoGAN-HD [55]	1729.6	2606.5	164.1	878.1
DIGAN [67]	1630.2	2293.7	83.11	196.7
StyleGAN-V [47]	1431.0	1773.4	79.52	197.0
PVDM-S (ours); 100/20-s	457.4	902.2	71.46	159.9
PVDM-L (ours); 200/200-s	398.9	**639.7**	61.70	137.2
PVDM-L (ours); 400/400-s	**343.6**	648.4	**55.41**	**125.2**

FVD的得分越低越好，PVDM的得分低于所有对照组

Method	IS ↑
MoCoGAN [57]	12.42±0.07
ProgressiveVGAN [1]	14.56±0.05
LDVD-GAN [23]	22.91±0.19
VideoGPT [65]	24.69±0.30
TGANv2 [43]	28.87±0.67
StyleGAN-V* [47]	23.94±0.73
DIGAN [67]	29.71±0.53
VDM* [21]	57.00±0.62
TATS [12]	57.63±0.24
PVDM-L (ours)	**74.40±1.25**

IS得分越高越好，PVDM高于所有对照组。

	Train	Inference (time/memory)	
Length →	16	16	128
TATS [12]	0	84.8/18.7	434/19.2
VideoGPT [65]	0	139/15.2	N/A
VDM [21]; 100/20-s	0	113/11.1	N/A
PVDM-L (ours); 200/200-s	2	20.4/5.22	166/5.22
PVDM-L (ours); 400/400-s	2	40.9/5.22	328/5.22
PVDM-S (ours); 100/20-s	**7**	**7.88/4.33**	**31.3/4.33**

与其他模型相比，时间及内存开销也较低。