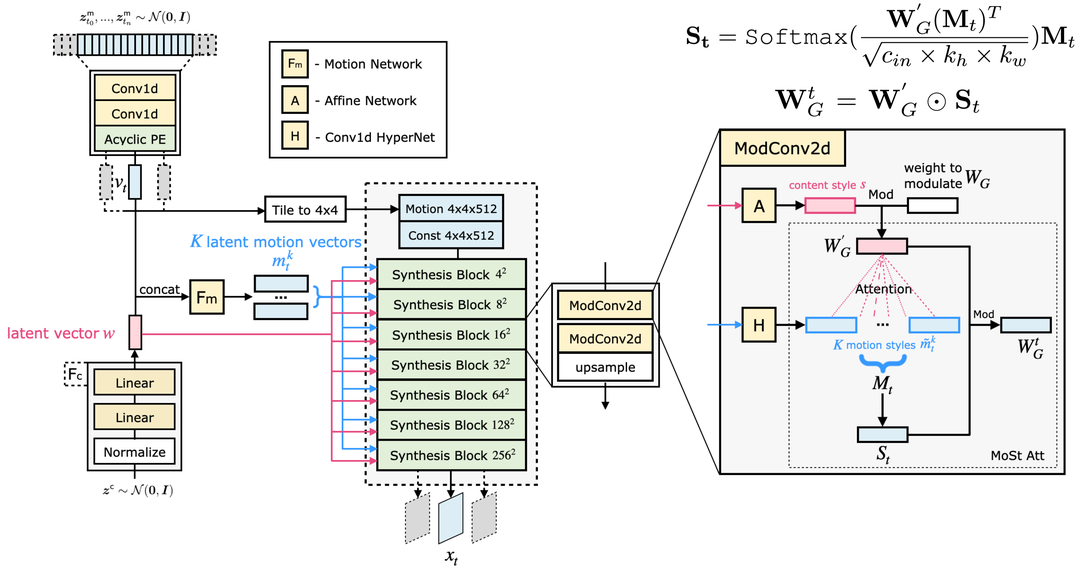
MoStGAN: Video Generation with Temporal Motion Styles

Introduction

Video generation remains a challenging task due to spatiotemporal complexity and the requirement of synthesizing diverse motions with temporal consistency. Previous works attempt to generate videos in arbitrary lengths but they struggle to synthesize detailed and diverse motions with temporal coherence. In this work, we argue that a single time-agnostic latent vector of a style-based generator is insufficient to model various and temporally-consistent motions. Hence, we introduce additional time-dependent motion styles to model diverse motion patterns. In addition, a **Mo**tion **St**yle **Att**ention modulation mechanism, dubbed as MoStAtt, is proposed to augment frames with vivid dynamics for each layer, which assigns attention scores for each motion style w.r.t deconvolution filter weights in the target synthesis layer and softly attends different motion styles for weight modulation. Experimental results show our model achieves state-of-the-art performance on four unconditional video synthesis benchmarks trained with only 3 frames per clip and produces better qualitative results with respect to dynamic motions.

Method

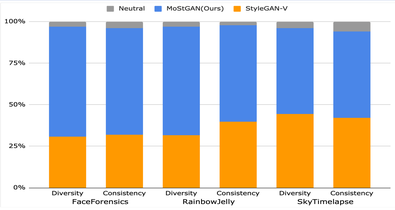


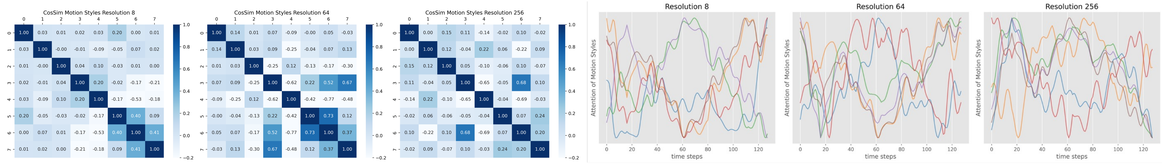
$$\mathbf{S_t} = \texttt{Softmax}(\frac{\mathbf{W}_G^{'}(\mathbf{M}_t)^T}{\sqrt{c_{in} \times k_h \times k_w}})\mathbf{M}_t$$

$$\mathbf{W}_G^t = \mathbf{W}_G^{'} \odot \mathbf{S}_t$$

Experiments

Method	FaceForensics 256^2		SkyTimelapse 256^2		RainbowJelly 256^2		CelebV-HQ 256^2	
	FVD_{16}	FVD_{128}	FVD_{16}	FVD_{128}	FVD_{16}	FVD_{128}	FVD_{16}	FVD_{128}
MoCoGAN-HD	111.8	653.0	164.1	878.1	579.1	628.2	212.4	753.1
VideoGPT	185.9	N/A	222.7	N/A	136.0	N/A	177.8	N/A
DIGAN	62.5	1824.7	83.1	196.7	436.6	369.0	72.9	163.2
StyleGAN-V	47.4	89.3	79.5	197.0	195.4	262.5	68.0	158.6
MoStGAN (ours)	**39.7**	**72.6**	**65.3**	**162.4**	**70.1**	**74.3**	**56.1**	**132.1**





Qualitative Results