Diffusion Models for Video Generation

medium.com/@myschang/diffusion-models-for-video-generation-ef242dad6f0c

Michael X

Note: If you miss out on the first two articles of this series, please click the link below to read more.

In recent years, generative models have made significant progress in the fields of computer vision and natural language…

Note: If you miss out on the first article of this series, please click the link below to read more.

medium.com

The Diffusion model can be used to generate sequences of data, such as videos. It is based on the idea of “diffusion” through a graph structure, where the nodes represent latent variables and the edges represent dependencies between the variables. In the context of video generation, the Diffusion model can be used to generate a sequence of images that form a coherent video by predicting the next frame in a sequence given the previous frames. It can capture complex dependencies between frames in the video, allowing it to generate more realistic and coherent videos, and it can handle variable-length sequences.

Generative models for generating photo-realistic videos are at the cutting edge of what is currently possible with deep learning. While previous work has demonstrated the ability to generate short, photorealistic videos, generating longer videos that are both coherent and photorealistic is still a challenge. One difficulty is the scaling required: generating a long video requires generating many photorealistic frames, which requires a lot of memory and processing power. Additionally, generating videos with long-range coherence (where each frame can depend on other frames far back in the video) is difficult, as it requires a model that can handle dependencies over a long time period. The [23] propose the Flexible Diffusion Model (FDM), a model based on the denoising diffusion probabilistic model (DDPM) framework, which can sample any subset of video frames given any other subset of frames. The FDM can be used to explore a wide range of resource-constrained video generation schemes and can be applied to tasks such as unconditional generation, video completion, and generation of videos of different lengths. The authors also release a new autonomous driving video dataset and a new video generative model performance metric.

The FDM is based on the denoising diffusion probabilistic model (DDPM) and can be conditioned on any number of frames at any time in the past or future, and can also be marginalized to achieve resource-constrained video generation. The authors use a meta-learning training objective to encourage the FDM to learn to generate videos that are photorealistic and coherent. They demonstrate that the FDM can be used to explore a wide range of resource-constrained video generation schemes and improve performance on long-range video modeling tasks. The FDM is trained on a distribution of tasks, with latent and observed indices randomly sampled during training. The authors also introduce a new autonomous driving video dataset and a new video generative model performance metric.

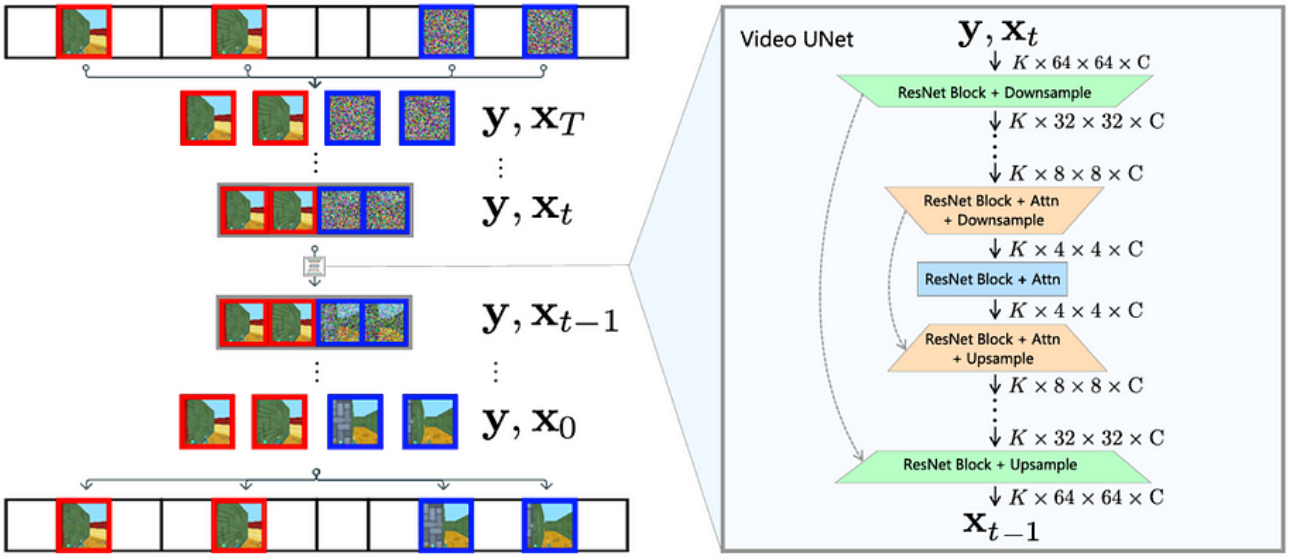


Figure30 DDPM for Video generation.

Furthermore, a diffusion model for video generation proposed in [24] shows promising initial results towards generating video milestones. This model, which is an extension of the standard image diffusion architecture, allows for joint training from image and video data, which the authors find reduces the variance of minibatch gradients and speeds up optimization. To generate longer and higher resolution videos, the authors introduce a new conditional sampling technique for spatial and temporal video extension that performs better than previous methods. They present the first results on a large text-conditioned video generation task and achieve state-of-the-art results on established benchmarks for video prediction and unconditional video generation.

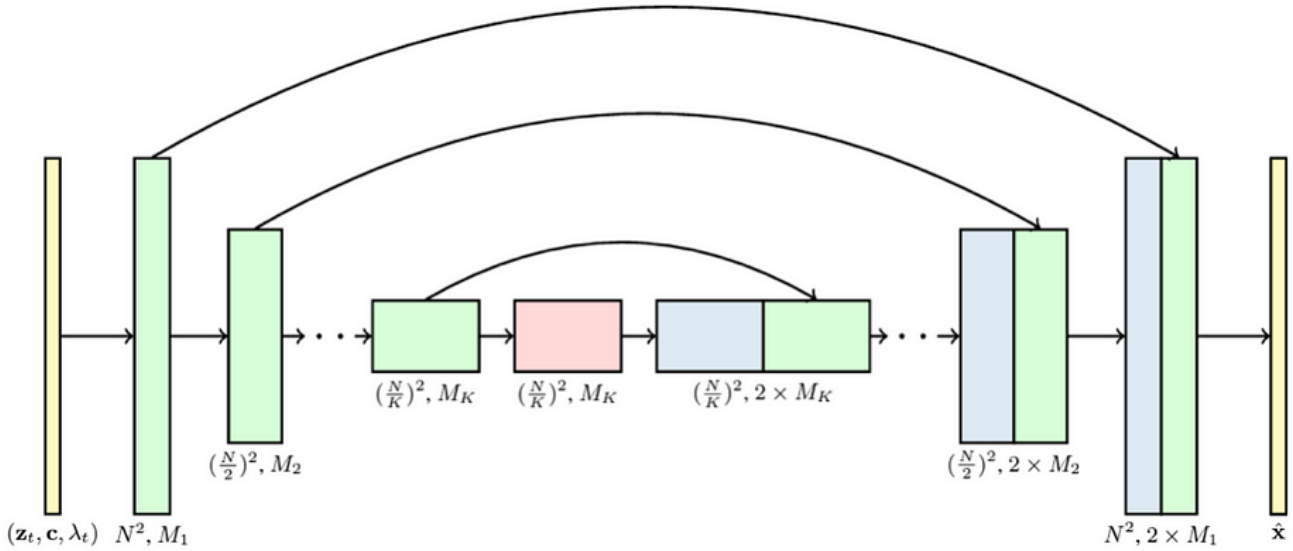
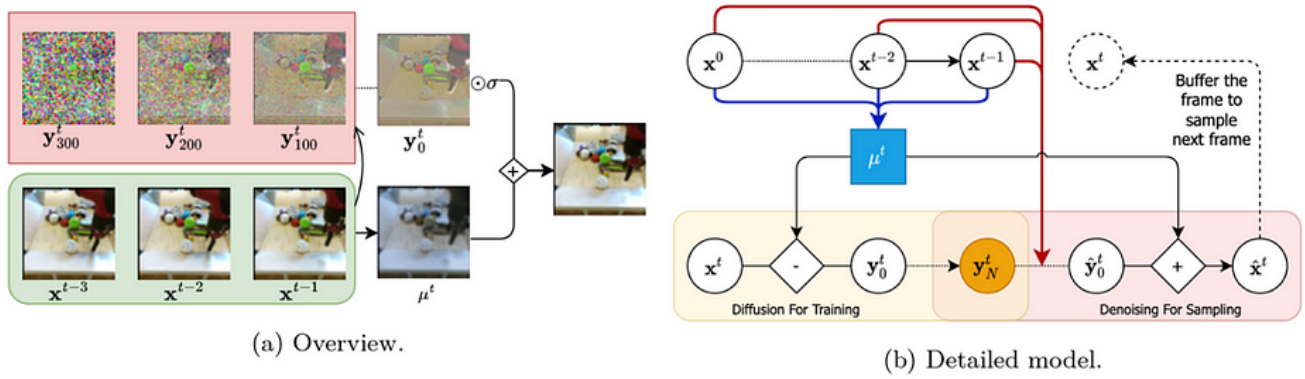


Figure 31 The Unet architecture used to generate video

The 3D U-Net architecture in the diffusion model consists of blocks that represent 4D tensors with axes labeled as frames, height, width, and channels. These blocks are processed in a space-time factorized manner, as described in a previous section. The input to the architecture is a noisy video, conditioning data, and the log SNR. The downsampling/upsampling blocks adjust the spatial input resolution (height and width) by a factor of 2 through each of the K blocks. The channel counts are specified using channel multipliers, and the upsampling pass has concatenation skip connections to the downsampling pass.

While models based on sequential variational autoencoders tend to be better at generating multi-modal predictions that accurately reflect data dynamics, sequential extensions of generative adversarial networks perform better at handling high-resolution content without blur. Diffusion probabilistic models have recently made progress in image generation and have perceptual qualities comparable to GANs without the optimization challenges of adversarial training. In the [25], authors extend diffusion probabilistic models for video generation and achieve a new state-of-the-art in video generation that produces sharp frames at higher resolutions. Their approach uses a deterministic convolutional RNN to predict the next frame and then corrects this prediction using an additive residual generated by a conditional denoising diffusion process. By modeling residuals from the predicted next frame instead of directly predicting the next frame, [25] achieve better results. Authors method is also better at probabilistic forecasting than modern GAN and VAE baselines.



(a) Overview. (b) Detailed model.

Figure 32 Overview of [25]

The approach predicts the next frame of a video using two convolutional recurrent neural networks (RNNs). The first RNN predicts the most likely next frame, and the second RNN generates a context vector for a denoising process. The denoising process is trained to model the residual between the actual next frame and the predicted frame, given the temporal context. During generation, the generated residual is added to the predicted next frame to produce the actual next frame.

Diffusion models have shown remarkable success in several generative tasks, but have not been extensively explored in the video domain. The [26] present Random-Mask Video Diffusion (RaMViD), which extends image diffusion models to videos using 3D convolutions, and introduces a new conditioning technique during training. By varying the mask conditioned on, the model is able to perform video prediction, infilling, and upsampling. Due to the simple conditioning scheme, the same architecture can be utilized as used for unconditional training, which allows the model to be trained in a conditional and unconditional fashion at the same time. The authors evaluate RaMViD on two benchmark datasets for video prediction, on which state-of-the-art results are achieved, and one for video generation.

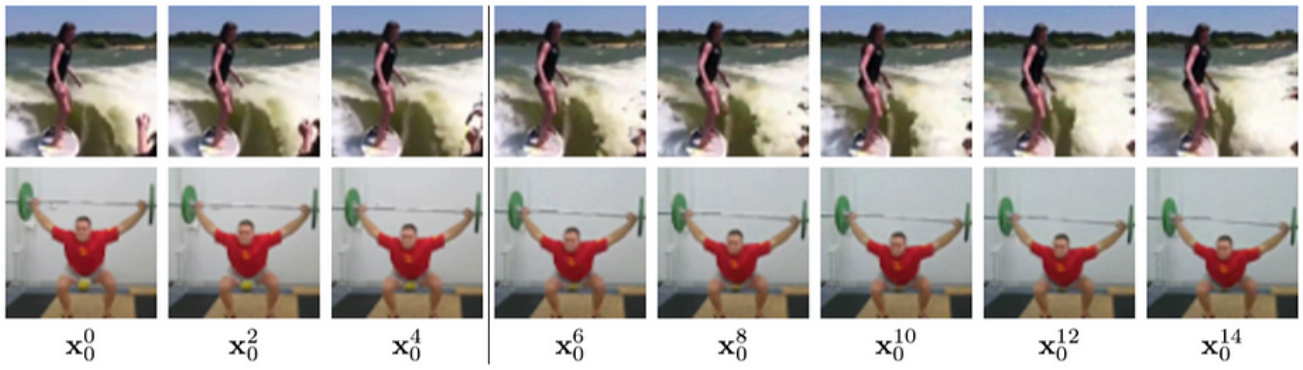


Figure 33 Prediction of 11 frames given the first 5 frames on Kinetics-600 with RaMViD

(TO BE CONTINUED)

Reference：

[22] MedSegDiff: Medical Image Segmentation with Diffusion Probabilistic Model

[23] Flexible Diffusion Modeling of Long Videos

[24] Video Diffusion Models

[25] Diffusion Probabilistic Modeling for Video Generation

[26] Diffusion Models for Video Prediction and Infilling