Research Paper using GANs for Video Prediction :

- From here to there: Video in between using direct 3d convolutions.
  https://arxiv.org/pdf/1905.10240.pdf

Github Code
https://github.com/tensorflow/hub/blob/master/examples/colab/tweening_conv3d.ipynb

We consider the problem of generating plausible and diverse video sequences, when we are only given a start and an end frame. In this paper, we propose instead a fully convolutional model to generate video sequences directly in the pixel domain. We first obtain a latent video representation using a stochastic fusion mechanism that learns how to incorporate information from the start and end frames. Our model learns to produce such latent representation by progressively increasing the temporal resolution, and then decode in the spatiotemporal domain using 3D convolutions. The model is trained end-to-end by minimizing an adversarial loss.

- FutureGAN: Anticipating the Future Frames of Video Sequences using Spatio-Temporal 3d Convolutions in Progressively Growing GANs
  https://arxiv.org/pdf/1810.01325.pdf

Github Code : https://github.com/TUM-LMF/FutureGAN
Encoder-decoder GAN model, FutureGAN, that predicts future frames of a video sequence conditioned on a sequence of past frames. During training, the networks solely receive the raw pixel values as an input, without relying on additional constraints or dataset specific conditions. To capture both the spatial and temporal components of a video sequence, spatio-temporal 3d convolutions are used in all encoder and decoder modules. PGGAN is used to achieve high-quality results on generating high-resolution single images

- Predicting Future Frames using Retrospective Cycle GAN
  http://openaccess.thecvf.com/content_CVPR_2019/papers/Kwon_Predicting_Future_Frames_Using_Retrospective_Cycle_GAN_CVPR_2019_paper.pdf

The key idea is to train a single generator that can predict both future and past frames while enforcing the consistency of bi-directional prediction using the retrospective cycle constraints. Moreover, we employ two discriminators not only to identify fake frames but also to distinguish fake contained image sequences from the real sequence. The latter discriminator, the sequence discriminator, plays a crucial role in predicting temporally consistent future frames.

- Learning to Generate Time-Lapse Videos Using Multi-Stage Dynamic Generative Adversarial Networks
  https://arxiv.org/pdf/1709.07592.pdf

Given the first frame, our model learns to generate long-term future frames. The first stage generates videos of realistic contents for each frame. The second stage refines the generated video from the first stage by enforcing it to be closer to real videos with regard to motion dynamics. To further encourage vivid motion in the final generated video, Gram matrix is employed to model the motion more precisely

- Deep Video Generation, Prediction and Completion of Human Action Sequences

  http://openaccess.thecvf.com/content_ECCV_2018/papers/Chunyan_Bai_Deep_Video_Generation_ECCV_2018_paper.pdf

To solve video generation from scratch, they build a two-stage framework where they first train a deep generative model that generates human pose sequences from random noise, and then train a skeleton-toimage network to synthesize human action videos given the human pose sequences generated. To solve video prediction and completion, They exploit our trained model and conduct optimization over the latent space to generate videos that best suit the given input frame constraints.

- Generating Videos with Scene Dynamics

  https://dl.acm.org/doi/pdf/10.5555/3157096.3157165
  Github Code: https://github.com/GV1028/videogan

They propose a generative adversarial network for video with a spatio-temporal convolutional architecture that untangles the scene's foreground from the background.

- Generating the Future with Adversarial Transformers

http://openaccess.thecvf.com/content_cvpr_2017/papers/Vondrick_Generating_the_Future_CVPR_2017_paper.pdf

They present a model that generates the future by transforming pixels in the past. Their approach explicitly disentangles the model's memory from the prediction, which helps the model learn desirable invariances.

- Probabilistic Video Generation using Holistic Attribute Control

https://arxiv.org/pdf/1803.08085.pdf

Github Code : https://github.com/yccyenchicheng/pytorch-VideoVAE

The proposed framework generates a video (short clip) by decoding samples sequentially drawn from a latent space distribution into full video frames. Variational Autoencoders (VAEs) are used as a means of encoding/decoding frames into/from the latent space and RNN as a way to model the dynamics in the latent space. They improve the video generation consistency through temporally-conditional sampling and quality by structuring the latent space with attribute controls; ensuring that attributes can be both inferred and conditioned on during learning/generation

- Stochastic Video Generation with a Learned Prior (SVG-LP)

  https://arxiv.org/pdf/1802.07687.pdf
  Github Code : https://github.com/edenton/svg

In this paper they introduce an unsupervised video generation model that learns a prior model of uncertainty in a given environment. Video frames are generated by drawing samples from this prior and combining them with a deterministic estimate of the future frame.

- Stochastic Adversarial Video Prediction (savp)

  https://arxiv.org/pdf/1804.01523.pdf

  Github Code: https://alexlee-gk.github.io/video_prediction/

Combining these two methods: (a) latent variational variable models that explicitly model underlying stochasticity and (b) adversarially-trained models that aim to produce naturalistic images

- Stochastic Variational Video Prediction (sv2p)

  https://arxiv.org/pdf/1710.11252.pdf

  Github Code
  https://github.com/tensorflow/tensor2tensor/blob/master/tensor2tensor/models/video/sv2p.py

In this paper, They develop a stochastic variational video prediction (SV2P) method that predicts a different possible future for each sample of its latent variables

- Videoflow: a conditional flow-based model for stochastic video generation (videoflow)

  https://arxiv.org/pdf/1903.01434v3.pdf

Github Code : https://github.com/fatemehazimi990/Pytorch-VideoFlow

They propose multi-frame video prediction with normalizing flows, which allows for direct optimization of the data likelihood, and produces high-quality stochastic predictions. They describe an approach for modeling the latent space dynamics, and demonstrate that flow-based generative models.

Research Papers related to Audio Generation using Gans :

- Speech Loss Compensation by Generative Adversarial Networks
  http://pub.dega-akustik.de/ICA2019/data/articles/001128.pdf
  https://ieeexplore.ieee.org/document/9023132

This is based on a GAN variant called Speech Enhancement GAN (SEGAN) that operates in the time domain by producing raw audio signals directly. The SEGAN generator is constructed as an Autoencoder, where the audio is encoded by using successive convolutional layers into a vector. This is concatenated with a vector of random noise and together, they are passed to the decoder which has a mirrored structure to that of the encoder. The decoder learns to recreate an enhanced version of the audio input to the encoder. In order to not lose low-level details of the input audio, the authors use skip connections between the corresponding layers of the encoder and the decoder to allow information such as phase or alignment to pass. On the other hand, given a pair of an impaired speech and its enhanced version, the discriminator D is trained to classify if the enhanced speech is real (actually from the dataset) or fake (bad imitation of the dataset).
SEGAN: speech enhancement generative adversarial network
https://arxiv.org/abs/1703.09452
Github Code : https://github.com/santi-pdp/segan

- Audio inpainting with generative adversarial network
  https://arxiv.org/pdf/2003.07704.pdf

Github Code: https://github.com/nperraud/gan_audio_inpainting

This focuses on audio inpainting in general using GANs, having validated their results on three different datasets of different musical instruments. They base their architecture on the Wasserstein GAN

- Vision-infused deep audio inpainting
  https://arxiv.org/abs/1910.10997

Github Code : https://github.com/Hangz-nju-cuhk/Vision-Infused-Audio-Inpainter-VIAI

This paper makes use of multi-modal audio/video content. The authors make use of two independent GANs (that share a single skip connection between one of the layers of their decoder sub-networks). The first just takes the audio as input in its Mel-Spectrogram form and learns to reconstruct the spectrogram without the missing segment. The second takes as input the corrupted Mel-spectrogram, the corresponding video, motion flows extracted from the video, and the clean audio spectrogram. The components are passed to encoders whose outputs are passed to a single decoder which learns to reconstruct the clean spectrogram. Together, the outputs of both decoders are passed to a WaveNet decoder which transforms the spectrograms to time-domain audio signals.

Research Papers for Audio Prediction (Other Popular Methods) :

- Deep speech inpainting of time-frequency masks

  https://arxiv.org/abs/1910.09058

Github Code :https://github.com/bepierre/SpeechVGG
End-to-end framework for speech inpainting, the context-based retrieval of missing or severely distorted parts of time-frequency representation of speech. The framework is based on a convolutional U-Net trained via deep feature losses, obtained using speechVGG, a deep speech feature extractor pretrained on an auxiliary word classification task.

- ConcealNet: An End-to-end Neural Network for Packet Loss Concealment in Deep Speech Emotion Recognition

  https://arxiv.org/abs/2005.07777

In this paper, They present a concealment wrapper, which can be used with stacked recurrent neural cells. The concealment cell can provide a recurrent neural network (ConcealNet), that performs real-time step-wise end-to-end PLC at inference time. Additionally, extending this with an end-to-end emotion prediction neural network provides a network that performs SER from audio with lost frames, end-to-end.

- A context encoder for audio inpainting

https://arxiv.org/pdf/1810.12138.pdf

Github Code : https://github.com/andimarafioti/audioContextEncoder

They propose a DNN structure that is provided with the signal surrounding the gap in the form of time-frequency (TF) coefficients. Two DNNs with either complex-valued TF coefficient output or magnitude TF coefficient output were studied by separately training them on inpainting two types of audio signals (music and musical instruments) having 64-ms long gaps. (works especially for music instruments)

- Adversarial audio synthesis (Wave Gan) (The Article Shared on the group)

  https://arxiv.org/pdf/1802.04208.pdf

Github Code: https://github.com/chrisdonahue/wavegan

WaveGAN, an attempt at applying GANs to unsupervised synthesis of raw-waveform audio.

- Audio super-resolution using neural nets

  https://arxiv.org/pdf/1708.00853.pdf

Github Code : https://github.com/kuleshov/audio-super-res

They introduce a new audio processing technique that increases the sampling rate of signals such as speech or music using deep convolutional neural networks. Their model is trained on pairs of low and high-quality audio examples; at test-time, it predicts missing samples within a low-resolution signal in an interpolation process similar to image super-resolution

- Wave Glow: A flow-based generative network for speech synthesis. *CoRR*, abs/1811.00002, 2018

https://arxiv.org/abs/1811.00002

https://github.com/NVIDIA/waveglow

In this paper we propose WaveGlow: a flow-based network capable of generating high quality speech from mel spectrograms. WaveGlow is implemented using only a single network, trained using only a single cost function: maximizing the likelihood of the training data, which makes the training procedure simple and stable.

- Adversarial generation of time-frequency features with application in audio synthesis

https://arxiv.org/abs/1902.04072

https://github.com/tifgan/stftGAN

They demonstrate the potential of deliberate generative TF modeling by training a generative adversarial network (GAN) on short-time Fourier features (STFT) . They show that by applying their guidelines, their TF-based network was able to outperform a state-of-the-art GAN generating waveforms directly, despite the similar architecture in the two networks.

- "I have vxxx bxx connexxxn!": Facing Packet Loss in Deep Speech Emotion Recognition

  https://arxiv.org/pdf/2005.07757.pdf

Mini- Survey on Packet Loss Generation :

- On Deep Speech Packet Loss Concealment: A Mini-Survey
  https://arxiv.org/pdf/2005.07794.pdf

Packet-loss is a common problem in data transmission, using Voice over IP. Review of classical methods, deep learning and generative models like Generative Adversarial Networks and Autoencoders for attempting to solve packet-loss using deep learning, by generating replacements for lost packets. In this mini-survey, They review all the literature we found to date, that attempt to solve the packet-loss in speech using deep learning methods. Additionally, They briefly review how the problem of packet-loss in a realistic setting is modelled, and how to evaluate Packet Loss Concealment techniques. Moreover, They review a few modern deep learning techniques in related domains that have shown promising results. These techniques shed light on future potentially better solutions for PLC and additional challenges that need to be considered simultaneously with packet-loss

Research Paper for Video Prediction (Other Methods) :

- Folded Recurrent Neural Networks for Future Video Prediction -

  https://arxiv.org/pdf/1712.00311.pdf

- Video Pixel Networks -

  https://arxiv.org/pdf/1610.00527.pdf

- Unsupervised Learning of Video Representations using LSTMs

https://arxiv.org/pdf/1502.04681.pdf

- Video Frame Interpolation via Adaptive Separable Convolution

  https://arxiv.org/pdf/1708.01692.pdf

- Unsupervised Learning for Physical Interaction through Video Prediction

  https://arxiv.org/abs/1605.07157

- Deep Multi-Scale Video Prediction beyond Mean square error

  https://arxiv.org/pdf/1511.05440.pdf

- MoCoGAN: Decomposing Motion and Content for Video Generation

  https://arxiv.org/pdf/1707.04993.pdf

  https://github.com/sergeytulyakov/mocogan

Research Papers related to Video Generation of Facial Images -

- Hierarchical Cross-Modal Talking Face Generation with Dynamic Pixel-Wise Loss https://arxiv.org/abs/1905.03820
- Every Smile is Unique: Landmark-Guided Diverse Smile Generation https://arxiv.org/pdf/1802.01873.pdf
- Audio-driven Talking Face Video Generation with Learning-based Personalized Head Pose https://arxiv.org/abs/2002.10137