Lab 4 - Conditional VAE for Video Prediction

2024 Spring
Cheng-yuan Ho

Apr 11, 2024

Outline

- Introduction
- Lab details
- Requirements
- Others

Introduction

Important Date

- Kaggle competition deadline: **2024/5/6 11:55**
- E3 upload deadline: **2024/5/7 18:00**
- **Demo: 2023/5/7**
- Format
  - Zip whole source code directory and named it in

    LAB4_{studentID}_{YOUR_NAME}.zip

    and upload to Lab 4 - Conditional VAE (code)
  - Save your report as pdf file and named it in

    LAB4_{studentID}_{YOUR_NAME}.pdf

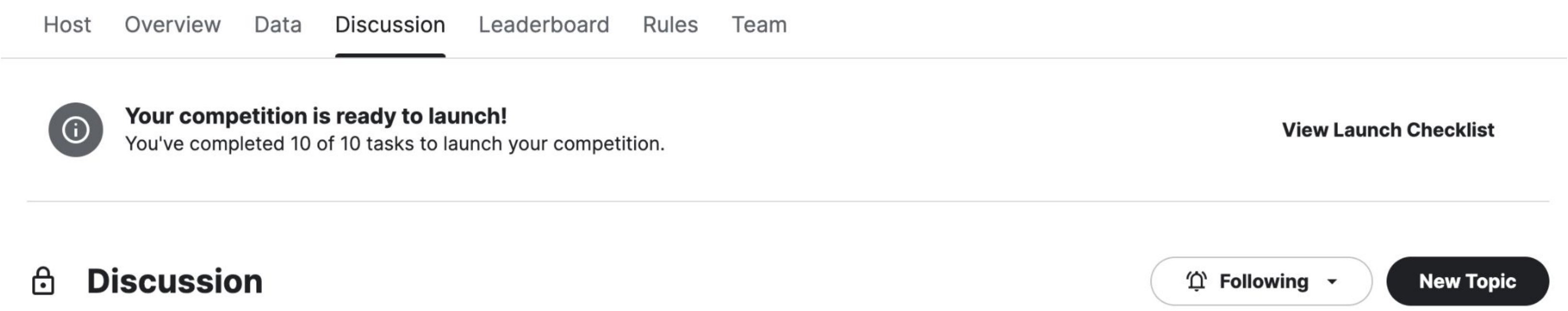
    and upload to Lab 4 - Conditional VAE (report)

Kaggle Competition

- Kaggle competition deadline: **2024/5/6 11:55**
- Team name: {your student id}_{your name}
  - **-5 points for wrong team name**
- 1 person 1 team
- 5 submission per day
- Tester.py will generate submission.csv for submission
- Scoring criteria
  1. Pass baseline: 20 points
  2. Top 30: 25 points
  3. Top 10: 30 points

General Forums

- If you got any questions, please post it in general forums
  - Other students might have the same questions
  - TAs will try their best to answer you in time



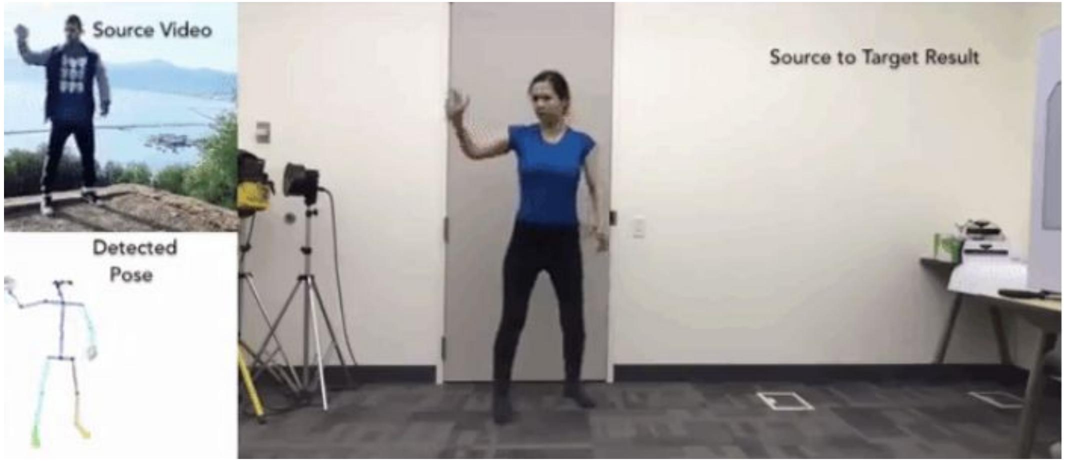
Your competition is ready to launch!
You've completed 10 of 10 tasks to launch your competition.

Introduction - Theme (Prior work)

Everybody Dance Now

Caroline Chan[*] Shiry Ginosar Tinghui Zhou[†] Alexei A. Efros

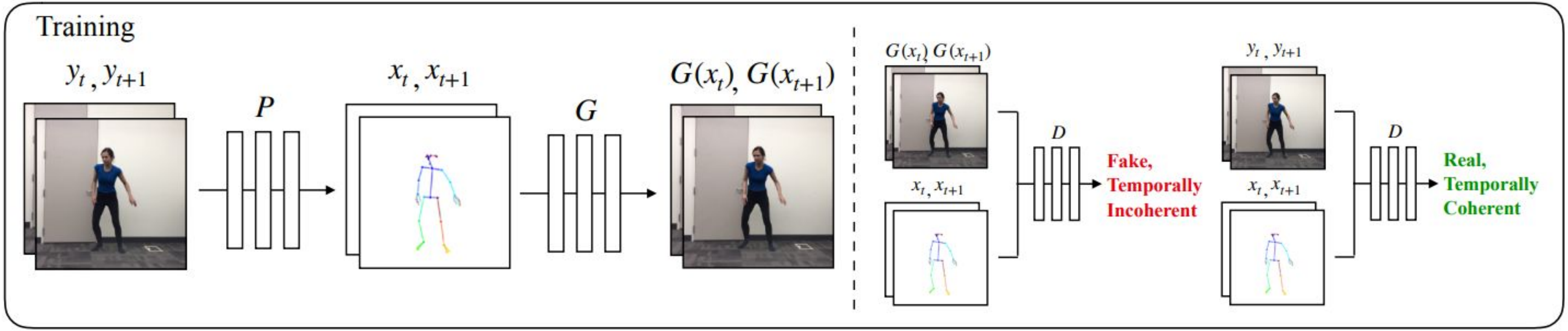
UC Berkeley



refer to https://github.com/carolineec/EverybodyDanceNow

Introduction - Theme (Prior work)

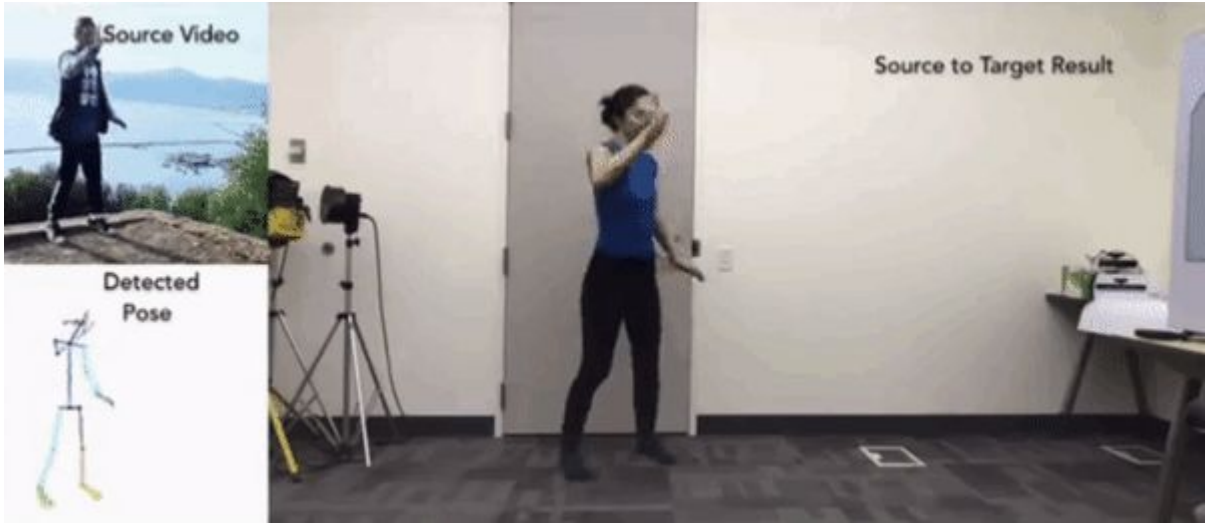
- Use pre-trained pose estimation network to generate pose images
- Predict the following video frame with GAN-based structure
- Generate the prediction by taking pose as inputs in inference time



refer to https://github.com/carolineec/EverybodyDanceNow

Introduction - Theme (Prior work)

- Inference the output in frame by frames
- Generate the video output by concatenate a sequence of output images

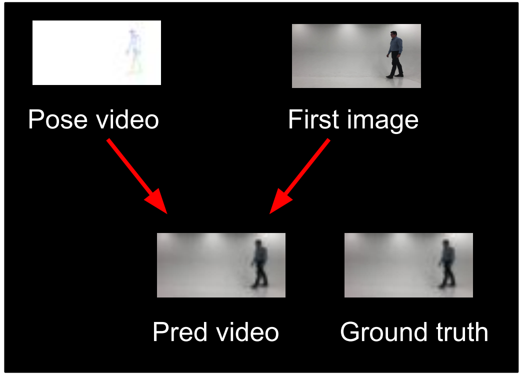


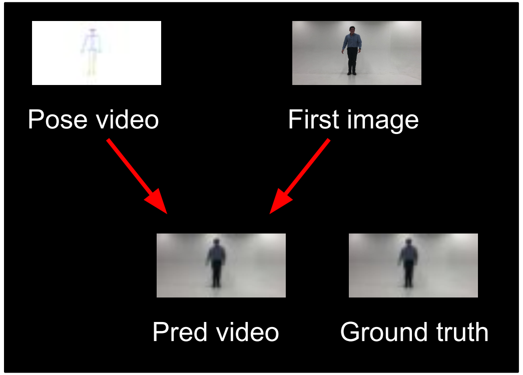
refer to https://github.com/carolineec/EverybodyDanceNow

Lab details

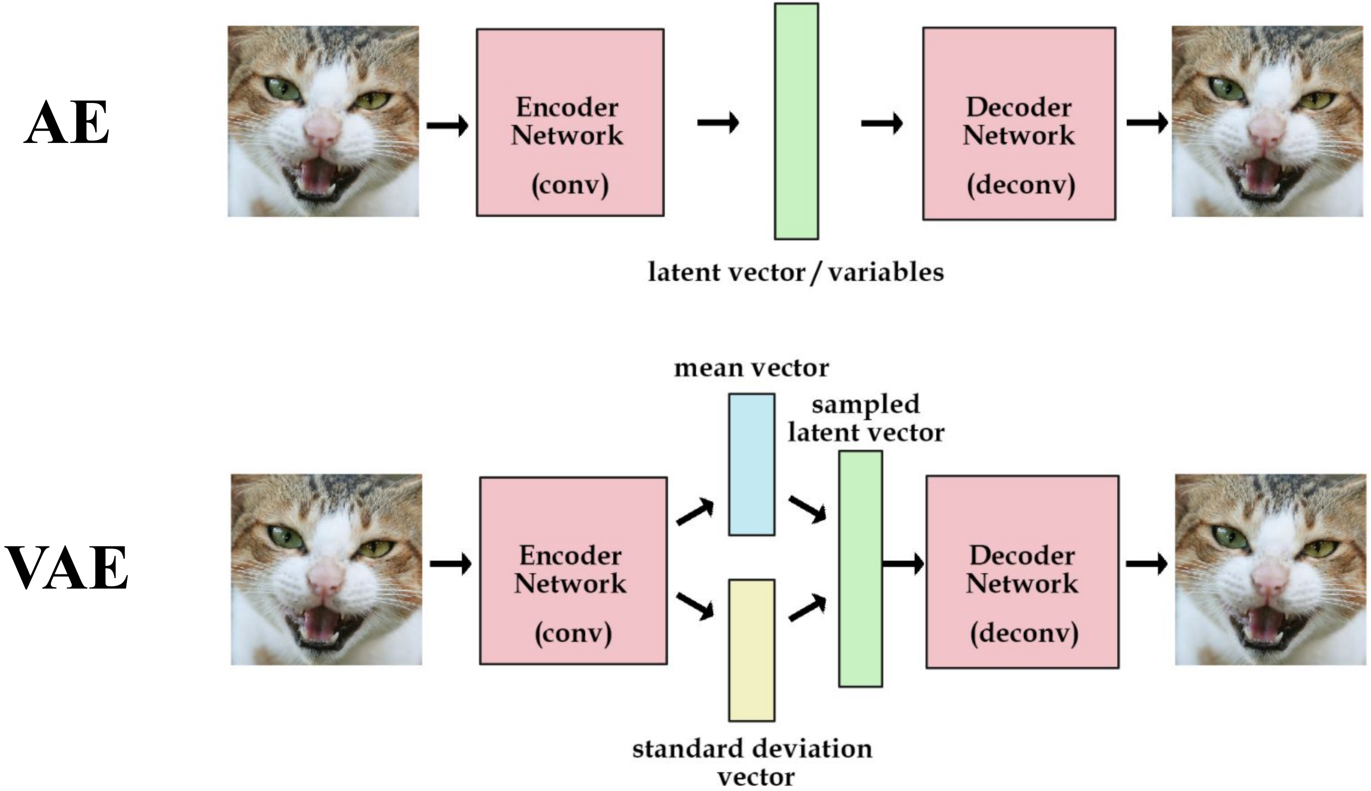
Lab Objective

- In this Lab, you need to implement video prediction by VAE method
  - Pose Video sequence
  - One Initial image
  - To generate predicted video sequence

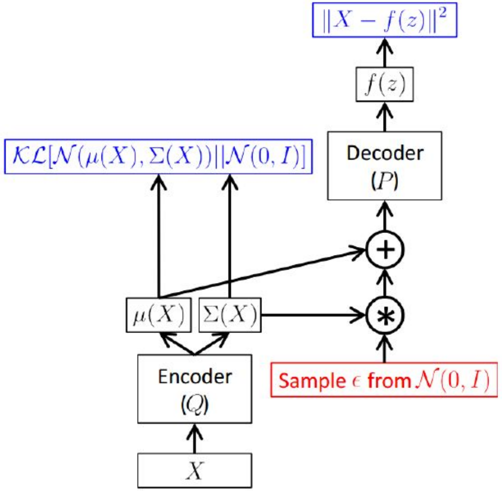




VAE recap


AE
Encoder Network (conv)
latent vector / variables
Decoder Network (deconv)
VAE
Encoder Network (conv)
mean vector
standard deviation vector
sampled latent vector
Decoder Network (deconv)

VAE recap - Reparameterization tricks



$\|X - f(z)\|^2$
$f(z)$
$\mathcal{KL}[\mathcal{N}(\mu(X),\Sigma(X))||\mathcal{N}(0,I)]$
Decoder
(P)
$+$
$\mu(X)$
$\Sigma(X)$
$*$
Encoder
(Q)
Sample ϵ from $\mathcal{N}(0,I)$
X

$$\underbrace{E_{\boldsymbol{Z}\sim q(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta}')} \log p(\boldsymbol{X}|\boldsymbol{Z};\boldsymbol{\theta})}_{\text{Re-parameterization for end-to-end training}} \quad -\mathrm{KL}(q(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\theta}')||p(\boldsymbol{Z}))$$

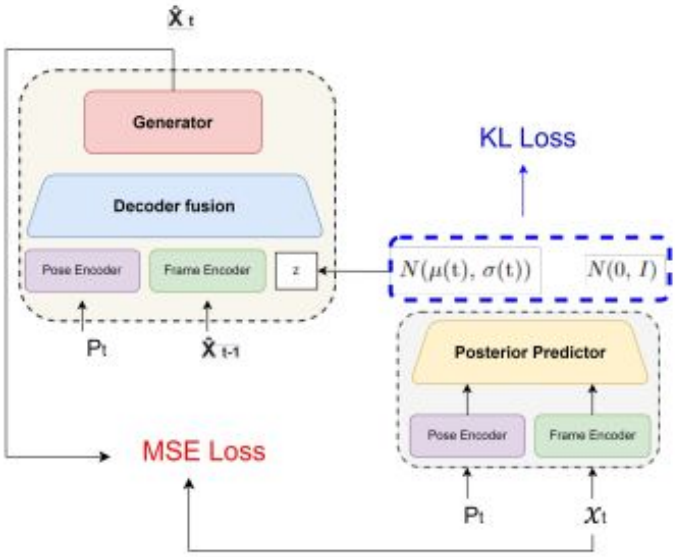
$$\mathcal{L}(\mathrm{X},\mathrm{q},\theta) = E_{Z\sim q(Z|X;\phi)} \log p(X|Z;\theta) - KL(q(Z|X;\phi)||p(Z))$$

where $q(Z|X;\phi)$ is considered as encoder and $p(X|Z;\theta)$ as decoder.

Output should be seem as log variance rather than simply variance

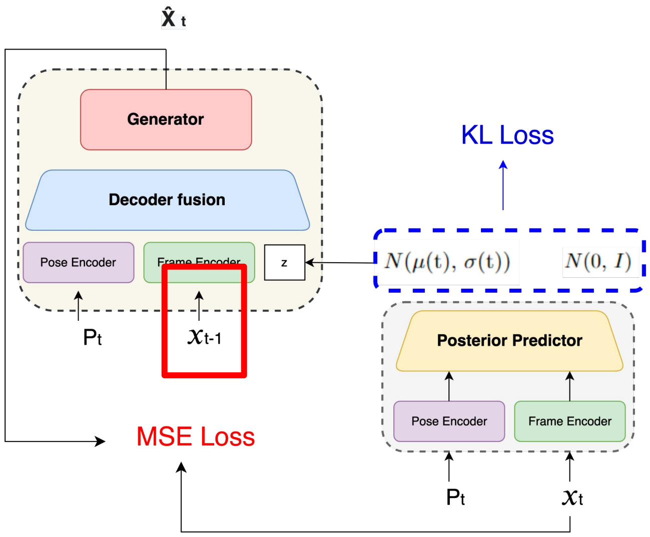
LAB Description - Training

- Generate posterior by **posterior predictor**
- Take the following info as input to Decoder Fusion
  - Pose image
  - Previous frame
  - Sample Z from posterior predictor
- Generate the final output by **Generator**
- **Loss = MSE-term + KL-term**



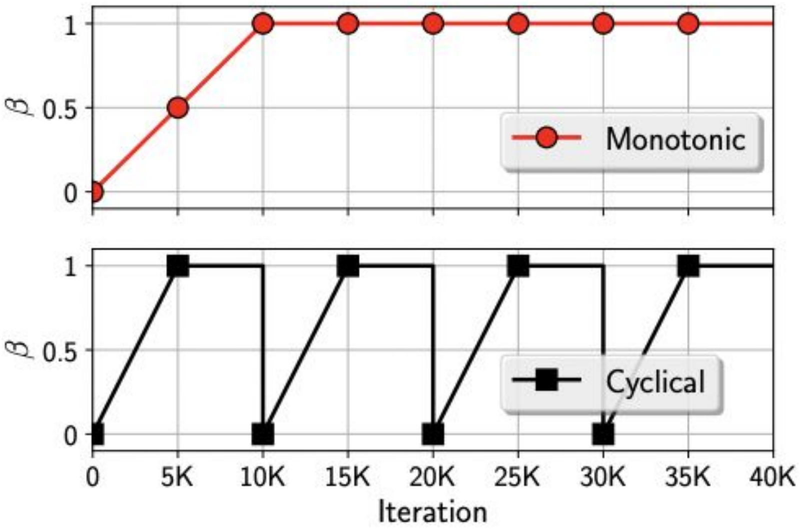
LAB Description - Teacher forcing

- Take ground truth frame as input rather than last generated frame
- Teacher forcing ratio is set to 0 ~ 1 and
    - When to use teacher forcing depends on your design



LAB Description - KL annealing

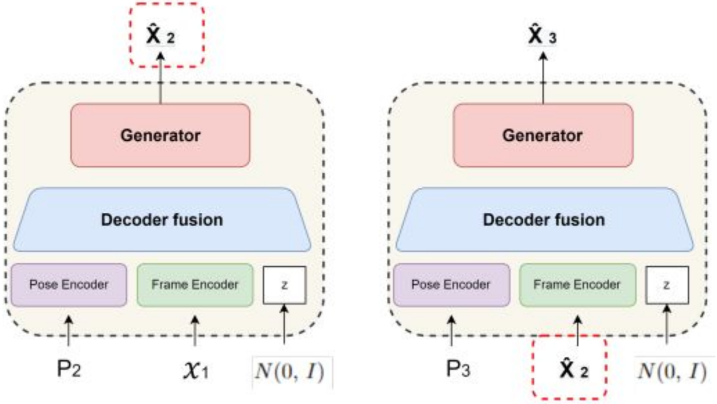
- To stable training
  - Loss = MSE-term + KL-term * α



H. Fu, et al., "Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing," NAACL 2019

LAB Description -Inference

- Generate frame by taking
- Pose image
- last generated frame
- Z sample from prior distribution
  - Prior distribution can be set by yourself
  - **N(0, I) is recommended**



Requirements

Requirements

- Training details implementation
  - Training protocol
  - Teacher forcing strategy. Teacher forcing ratio **range 0 ~ 1**
  - KL annealing strategy
  - Other training strategy (training trick)
- Plot diagram
  - Plot the loss curve in different kind of setting while training
  - Plot PSNR per frame while validation your output result
  - Plot teacher forcing ratio while training

Requirements

- Analysis
  - Compare the loss curve difference in different setting and make your assumption
- Make your validation result into gif files (This should be shown in Demo)
- Derivation of conditional VAE (see the detail in spec)

Others

Testing time

- **5 videos** should be generated
  - Each video sequence contains
    - **One initial frame**
    - **630 pose images**
  - You need to take given datas to generate **the following 629 future frames**

Provided files

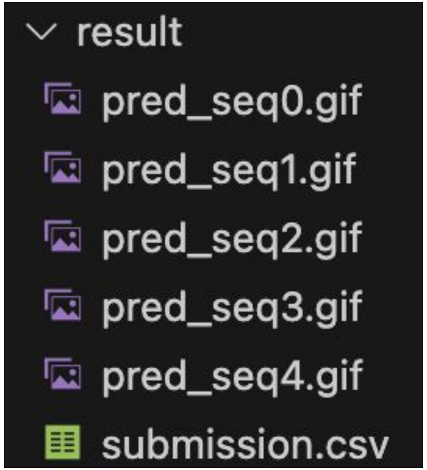
- Trainer.py
- Tester.py
- dataloader.py
- requirements.txt (**python3.9 is recommended**)
- modules/

RUN Test

- Testing file has been done, simply type the following command

```
python Tester.py --DR {YOUR_DATASET_PATH}
--save_root {PATH_TO_SAVE_YOUR_CHECKPOINT}
--ckpt_path {PATH_TO_YOUR_CHECKPOINT}
```

- If success, it will output 5 gif files and submission.csv
- submission.csv is used for submission to kaggle



```
⌄ result
  pred_seq0.gif
  pred_seq1.gif
  pred_seq2.gif
  pred_seq3.gif
  pred_seq4.gif
  submission.csv
```

Dataset

a. Training dataset

  i. train_img: 23410 png files

  ii. train_label: 23410 png files

b. Valadition dataset

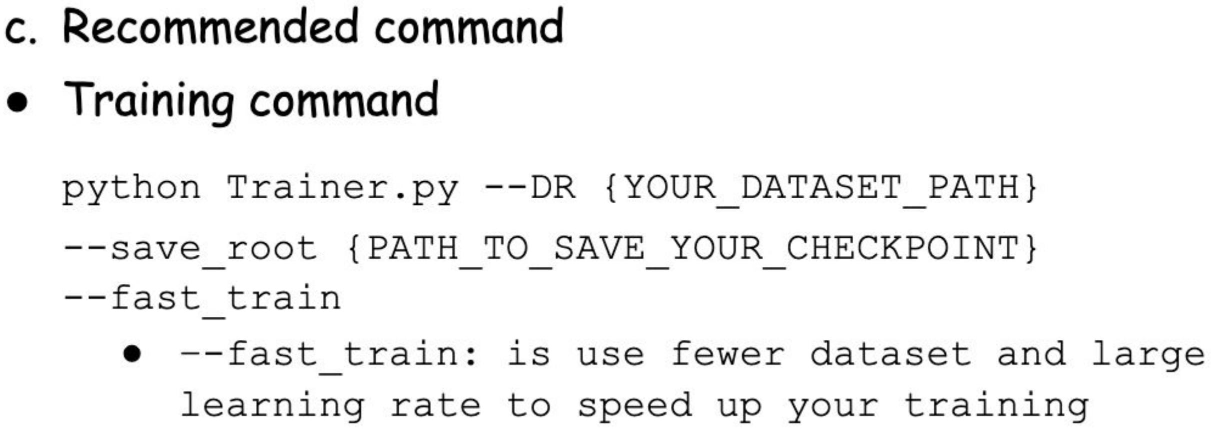
  i. val_img: 630 png files

  ii. val_label: 630 png files

c. Testing dataset

  i. 5 video sequences are given. Each video sequence contains one first frame and 630 label frames.

Recommend commands



c. Recommended command

- Training command

```
python Trainer.py --DR {YOUR_DATASET_PATH}
--save_root {PATH_TO_SAVE_YOUR_CHECKPOINT}
--fast_train
```

  - --fast_train: is use fewer dataset and large learning rate to speed up your training

Reference

[1] C. Chan, et al., "Everybody Dance Now," ICCV, 2019

[2] E. Denton, et al., "Stochastic Video Generation with a Learned Prior," ICML, 2018

[3] H. Fu, et al., "Cyclical Annealing Schedule: A Simple Approach to Mitigating KL Vanishing," NAACL 2019