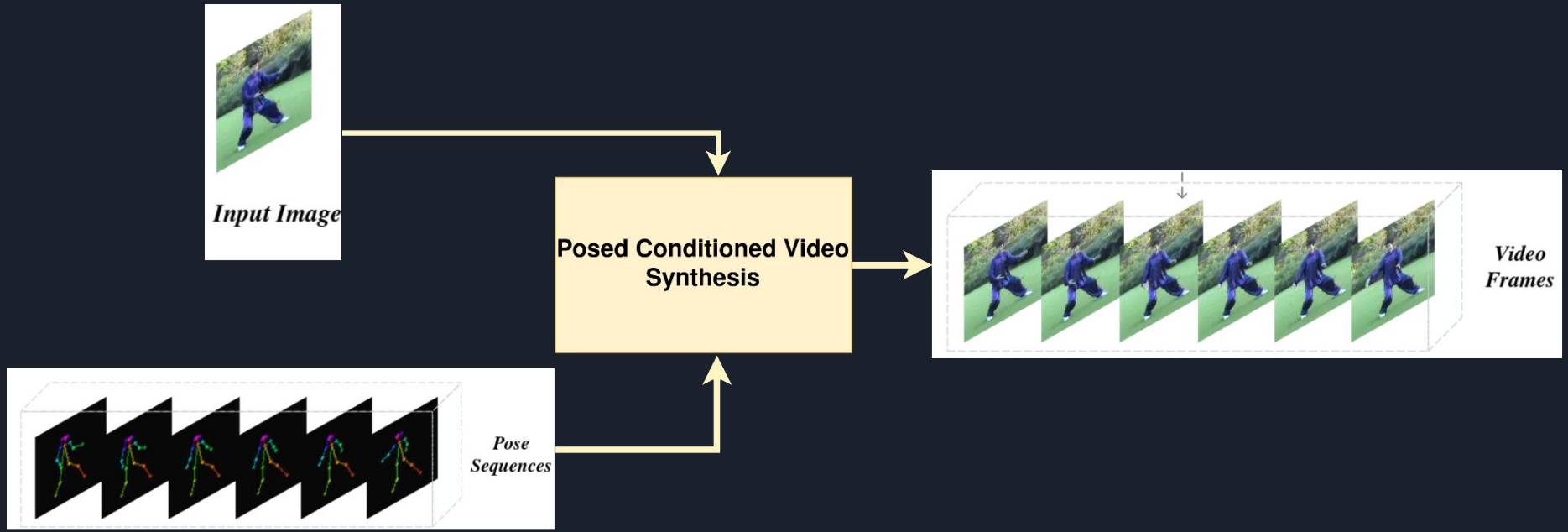
Person Video Synthesis via Denoising Diffusion Model

Project Pitch by Manu Pillai, Mukund Dhar, Prudvi Kamtam and Adeel Yousaf

Introduction

- Person Video Synthesis aims to generate video frames of a human given an image and a set of poses.



Background

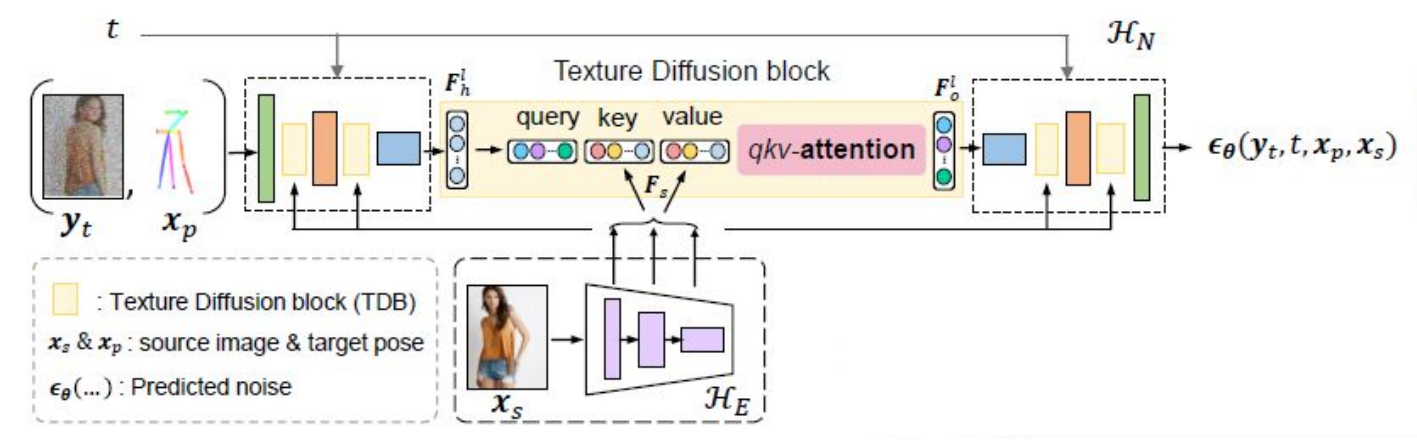
Person Image Synthesis via Denoising Diffusion Model [1]

- The first diffusion-based approach for pose-guided person image synthesis task which can work under challenging pose transformations while preserving appearance, texture and global shape characteristics.

- Effectively models the complex interplay between appearance and pose information using proposed “texture diffusion module”.

- Introduces disentangled classifier-free guidance to tightly align the output image style and pose with the source image appearance and target pose

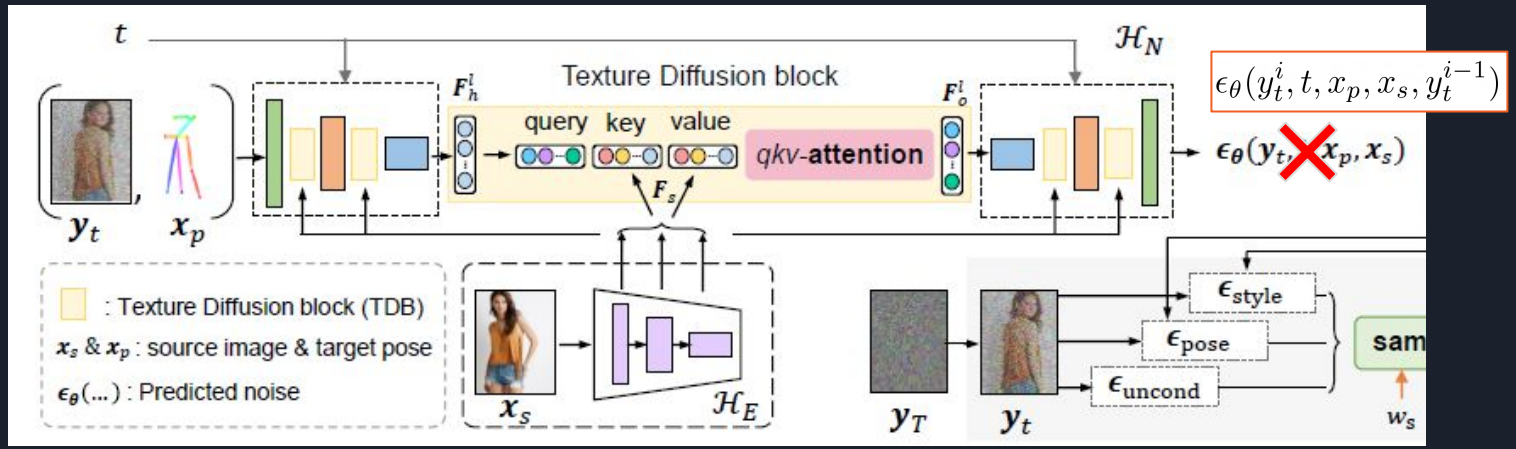
Person Image Synthesis via Denoising Diffusion Model [1]



Why use skeletons for human video generation?

- The existing GAN-based approaches attempt to directly transfer the style of the source image to a given target pose, which requires the architecture to model complex transformation of pose.
- Pose/skeletons provides a direct supervision for the structure of the generated image and aids in easier learning of the model.

Person **Video** Synthesis via Denoising Diffusion Model (High-level Design)



Person **Video** Synthesis via Denoising Diffusion Model (High-level Design)

Experiments:

- Apply PIDM frame by frame given a sequence of poses generated from a video.
- Introducing temporal consistency within Texture diffusion block.
- Having previous frames as an input to source encoder module.

Datasets:

- Human 3.6M
- Deep Fashion

References

[1]. Bhunia, Ankan Kumar, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. "Person Image Synthesis via Denoising Diffusion Model." arXiv preprint arXiv:2211.12500 (2022).

[2]. Yang, Ceyuan, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. "Pose guided human video generation." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 201-216. 2018.

Thank You