Chapter 1
INTRODUCTION

Given a static image, humans can think of various scenes of what will happen next using their imagination. For example, considering the ballerina in Figure 1, one can easily picture the scene of the dancer jumping higher or landing softly. In this work, we clarify the task as intimating human capability of ***Video Imagination***: synthesizing imaginary videos from single static image. This requires synthesized videos to be diverse and plausible.

Compared to related tasks, e.g. video anticipation and prediction, there are more challenges for video imagination. Video imagination means to produce real high-dimension pixel values unlike low-dimension vectors in semantic anticipation. In addition, videos that are not identity to each other can all be reasonable, like imaginary video 1 and imaginary video 2 in Figure1. So there is no precise ground truth as in common video prediction task. This intrinsic ambiguity makes regular criterion like MSE fails in evaluating whether the synthesized video is plausible.

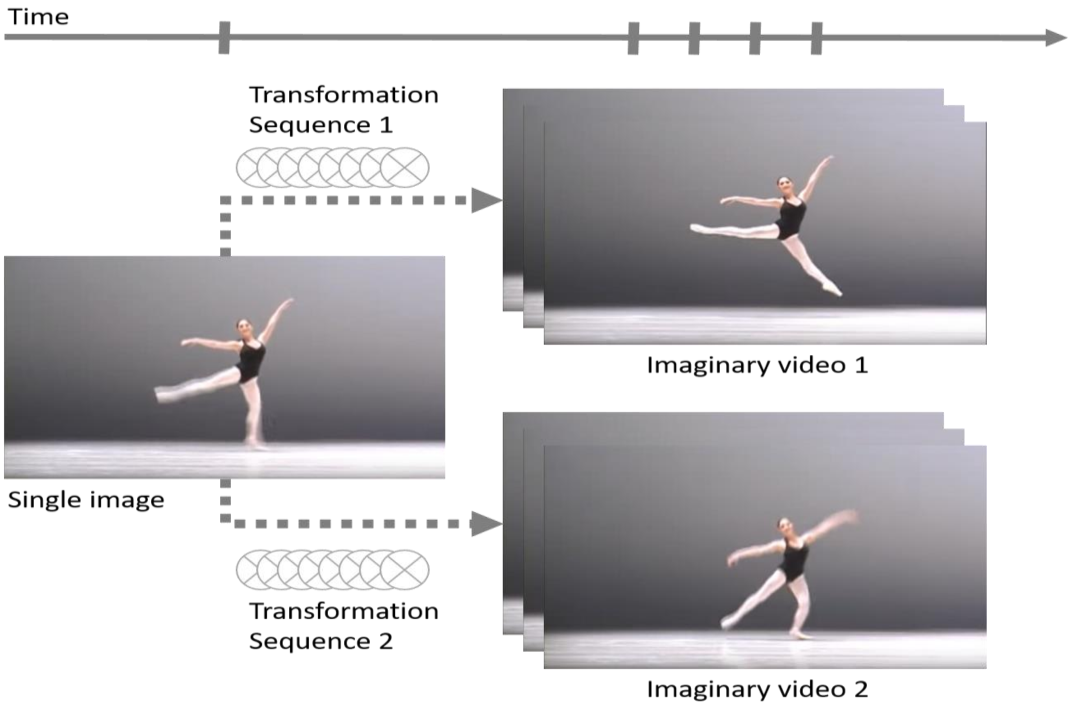

Figure1: Synthesizing multiple imaginary videos from one single image. Given an image of a dancing ballerina, the videos of the dancer jumping higher or landing softly are both plausible imaginary videos. Those videos can be synthesized through applying a sequence of transformations.

In this work, we present an end-to-end unsupervised framework with transformation generation for video imagination. Our key intuition is that we can model in transformation space instead of pixel space. Since scenes in frames are usually consistent, we assume that the major motions between frames can be modeled by transformations. If we reconstruct frame based on the original image and corresponding transformations, both scene dynamic and invariant appearance can be preserved well. In addition, we draw inspiration from image generation works [23] that use adversarial training. We believe an elaborate critic network that understands both spatial and temporal dependency would serve as reasonable criterion.

Based on the intuition and inspiration above, we design our framework focusing on model distributions in transformation space implicitly, and train it in adversarial way. In this framework, we generate transformation conditioned on the given image. Then we reconstruct each frame by applying the generated transformation to the given image. Latent variable is also introduced to enable diverse sampling. Casting this into an adversarial architecture, we train our framework in a fully end-to-end fashion.

The latent variable enables diverse imagination through sampling different transformations corresponding to different imaginary videos. Furthermore, there is nearly infinite resource for this unsupervised training. No label is needed, so every video clip can serve as a training sample.

For evaluation, since there is no general evaluation metrics for this task, we employ image quality assessment method to evaluate the quality of reconstructed frames and present a relative image quality assessment (*RIQA*) to eliminate the scene difference. In experiments, we evaluate our idea on three datasets, including two artificial video datasets with simple motions and one natural scene video dataset with complex motions. The synthesized 4-frames video results show that our framework can produce diverse sharp videos with plausible motions. The quantitative evaluation results suggest that our framework outperforms others including those methods that are given more prior information, and the qualitative comparison also shows the advance of our synthesized videos.