Chapter 2
RELATED WORK

Although the works of future video synthesis from single image are rather li1le, our task shares common techniques with two related tasks: video prediction [27] and image reconstruction [32], where researchers have made impressive progress.

Reconstruction in pixel space. Early works of visual prediction focus on modeling and estimation in pixel space [38] [28] [37]. These methods reconstruct images by calculating pixel values directly. With recent resurgence of deep networks, researchers tend to replace standard machine learning models with deep networks. In particular, [14] proposes a video pixel network and estimates the discrete joint distribution of the raw pixel values. [26] uses LSTM network to learn representations of video and predict future frames from it. A key issue in pixel-level prediction is the criterion metrics. A recent work [20] argues that standard mean squared error (MSE) criterion may fail with the inherently blurry predictions. They replace MSE in pixel space with a MSE on image gradients, leveraging prior domain knowledge, and further improves using a multi-scale architecture with adversarial training.

Mid-level tracking and matching. To overcome the challenge of high dimensionality and ambiguity in pixel space, the prediction framework of mid-level elements gradually becomes popular. [19] explores a variation on optical flow that computes paths in the source images and copies pixel gradients along them to the interpolated images. [35] combines the effectiveness of mid-level visual elements with temporal modeling for video prediction. [24] defines a recurrent network architecture inspired from language modeling, predicting the frames in a discrete space of patch clusters.

Existing pixels utilization. A insightful idea of improving the quality of prediction image is to utilize existing pixels. [18] synthesizes video frames by flowing pixel values from existing ones through voxel flow. [41] outputs the difference image, and produces the future frame by sum up the difference image and raw frame. [5] and [7] share a similar methods with us of applying filters to raw frames to predict new frames, and they provide the validation of gradients flow through filters.

Generation model evolutions. Traditional works treat visual prediction as a regression problem. They often formulate prediction tasks with machine learning techniques to optimize the correspondence estimation [11, 16, 17]. With the development of deep networks, community of visual prediction has begun to produce impressive results by training variants of neural network structures to produce novel images and videos [9, 39, 40]. The probabilistic models become popular again. More recently, generative adversarial networks (GANs) [8] and variational autoencoders [15] have been used to model and sample from distributions of natural images and videos [6, 23, 42].

To the best of our knowledge there are no existing model that can produce multiple videos given one single image. Perhaps the most similar works to our task are [34, 41], where both works aim to build a probabilistic model of future given an image. But [41] only outputs one frame and [34] just produce optical flows.

Also, note a concurrent work that learns to predict in transformation space is [31], where the authors predict the new frames by predicting the following transformations. But their task is to generate frames from sequence of frames while ours is to synthesize imaginary videos given a single image. In addition, our work differs in that there methods are close to a regression problem as to predict precise future frames, but our task requires a probabilistic view and aims at generating multiple videos.