Chapter 4
EXPERIMENT

In this section, we experiment our framework on 3 video datasets: Moving MNIST, 2D shape and UCF101. For evaluations, we perform qualitative inspection and novel quantitative assessment *RIQA* to measure the objective quality of the imaginary video.

Baselines and Competing Methods

Current work about this task is quiet limited. To find out whether our framework outperforms those methods that do not involve our crucial components, we develop two simple but reasonable base- lines for this task. For the first one, **Baseline 1**, the transformation generator and volumetric merge network in our original framework are replaced by a generator network that directly outputs fia1en pixels. For the second one, **Baseline 2**, the whole adversarial training procedure including critic network is removed, and the network is trained minimizing l_2 loss function. Those two baselines can also be considered as a form of ablation experiments. We also consider several latest works as competing methods as shown in 1.

Model	Input	Output
* Ours*	image	5 frames(multiple)
Visual Dynamic [5]	image	1 frame(multiple)
Scene Dynamic [33]	image	32 frames
Dynamic Filter [5]	4 frames	1 frame
Beyond MSE [20] *	4 frames	1 frame
Video Sequences [31] *	4 frames	4 frames

Table 1: Task setting comparison of related work. Multiple output means that the method build a probabilistic model and can sample different results. indicates the methods can also experiment on natural scenes like in UCF101 dataset.

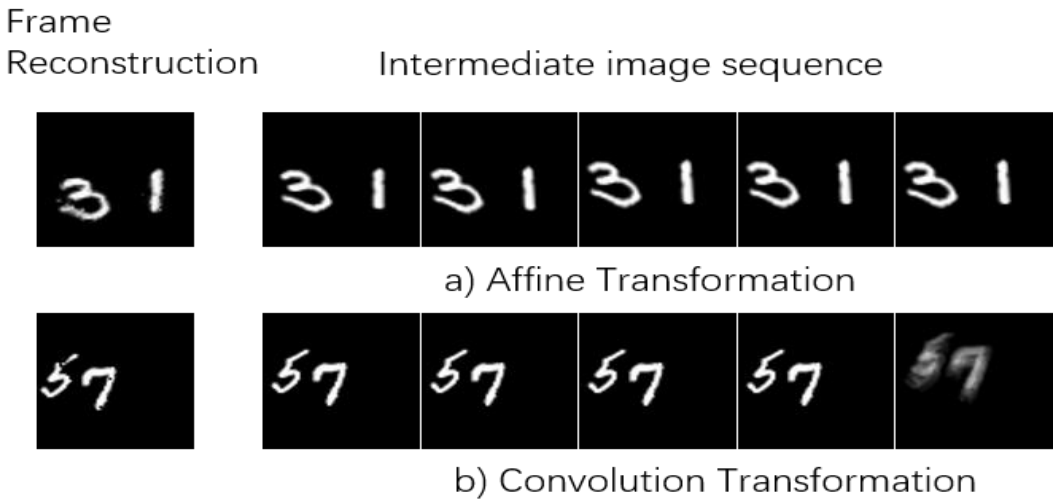

Figure 6: Intermediate image sequences visualization. Different transformation models result in different intermediate image sequences I_T. Each intermediate image represent one mode of simple transformations. A sequence of these intermediate image can form more complex motion.

Moving MNIST Dataset

Dataset: We first experiment on a synthetic grey video dataset: moving MNIST dataset [26]. It consists of videos where two MNIST digits move in random directions with constant speed inside a 64*64 frame. The 64,000 training video clips and 320 testing clips are generated on-the-fly. Each video clip consists of 5 frames. Taking the first frame as input, the goal is to synthesize multiple imaginary 5-frames videos.

Setup: There is barely no pre-processing in our work except for normalizing all videos to be in range 0,1. We experiment on two transformation models. For convolutional transformation we set kernel size as 9*9, and the transformation

sequence length P is set as 5 for both models. We generate 4 transformation sequences $\{\Phi 1, \Phi 2, \Phi 3, \Phi 4\}$ corresponding to 4 consecutive frames $\{ f1, f2, f3, f4 \}$ at once.

Result: Figure 5(a) illustrates the qualitative performance in moving MNIST dataset. As we can see, frames are sharp and clear while the shape information of digits is well preserved as we expect. The difference images show that the generated transformations successfully model one motion mode so that the synthesized imaginary video has plausible consecutive motion. Figure 6 shows reconstructed frames and the corresponding intermediate image sequences in different transformation models.

Synthetic 2D Shapes Dataset

Dataset: We experiment our framework using a synthetic RGB video dataset: Synthetic 2D Shapes Dataset [41]. There are only three types of objects in this dataset moving horizontally, vertically or diagonally with random velocity in $0, 5$. All three objects are simple 2D shapes: circles, squares, and triangles. The original dataset only contains image pairs that have 2 consecutive frames. We extrapolation it to convert image pairs into video clips that have 5 frames. There are 20,000 clips for training and 500 for testing just like se1ings in [41]. We aim at synthesizing multiple imaginary videos each containing five consecutive frames.

Setup: The input image size is set as 64 64 so that we can inherit the network architecture and se1ings in section 4.2. The transformations applied to each color channel are set to be identical for the consistent of RGB channels.

Result: Figure 5 (b) illustrates the qualitative performance in 2D shape dataset. Appearance information including color and shape is reconstructed at a satisfying level, and the motion is plausible and non-trivial. Multiple sampling results are shown in Figure 7. It is clear that sampling different z s lead to different imaginary