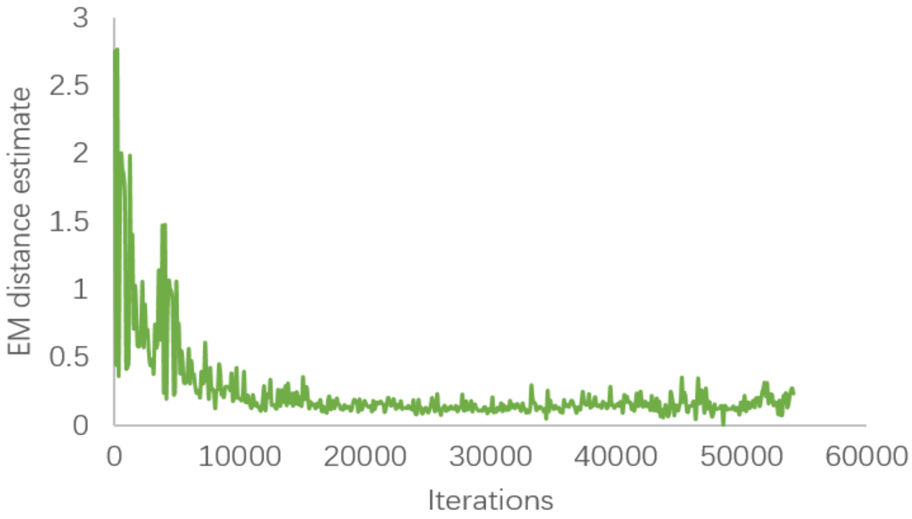

Figure9: Curveof EM distance estimate at different steps of training. The estimation of EM distance is done by a videocritic network C that is well trained. We can see that the EM distance decrease and converge with training.

Methods	Input BRISQUE	Output BRISQUE	RIQA
Ours 64 *64	45.2164	47.0168	3.98%
Ours 128*128	35.9809	36.7120	**2.03%**
Baseline 1	45.2164	50.7681	12.28%
Baseline 2	45.2164	89.2315	97.34%
Optical Flow [4]	39.3708	40.8481	3.75%
Beyond MSE [20]	46.3219	50.0637	9.24%
Video Sequences [31]	39.3708	42.8834	8.92%

Table 2: quantitative evaluation comparison among related visual prediction work. The lower RIQA indicates better frame reconstruction quality. The BRISQUE score obviously varies with scenes and resolutions. RIQA points out the decreasing proportion between input and output, hence successfully reflects the reconstruction quality.

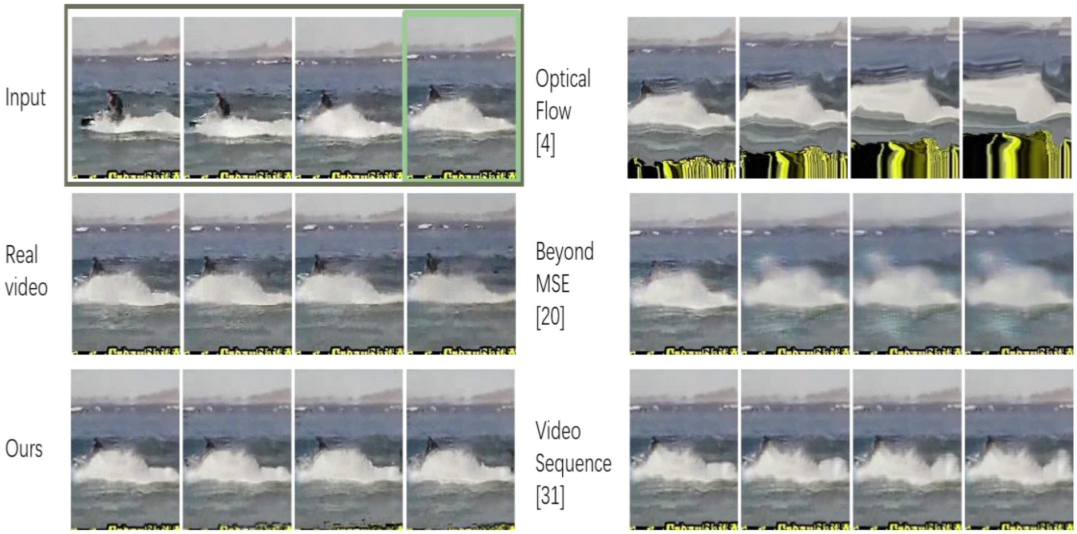

Figure 10: Perceptual Comparison among related works using UCF101 dataset. The input frames are from Skijet class. The output frames are reshaped to same size for a fair visual inspection. Notice that our framework only takes one frame as input (the green square) while the rest methods take four frames as input (the grey rectangle). Our result are sharp and relatively clear while the motions of rider and skijet are recognizable and plausible.

Settings	RIQA
affine transformation with K = 6 and P = 5	2.03%
affine transformation with K = 6 and P = 10	4.79%
convolutional transformation with K = 8 × 8 and P = 5	4.03%
convolutional transformation with K = 16 × 16 and P = 5	4.01%

Table 3: Analysis of the settings of models and hyper- parameters. *K* refers to the number of parameters forming trans- formation. *P* refers the sequence length of transformation for re- constructing one frame.

As shown in Table 2, diversity of the scenes and resolutions makes the raw BRISQUE score not comparable, but the *RIQA* tells the reconstruction quality change. We can see that our framework outperform other methods, and the poor performances of baselines suggest the architecture of our framework is reasonable. In addition, our framework and Video Sequence [31], that are based on transformation space do produce images with better qualities than [20], which reconstruct frames from scratch.

Table 3 shows the results when we change the hyper-parameters and some model se1ings, including the number of parameters K forming transformation, the sequence length of transformation P for reconstructing one frame, and the type of transformations. The results demonstrate that our framework is overall robust to those choice. It seems that affine transformation model with transformation sequence length $P = 5$ can achieve the best performance.

Qualitative inspection. Figure 10 shows the perceptual comparison between our framework and three competing methods [4, 20, 31] that also experimenting on UCF101 dataset. Our frame- work produces four frames conditioned on one frame while other methods take a sequence of four frames as inputs. The simple optical method [4] fails due to the strong assumption of constant flow speed, yet it perform relatively better in quantitative evaluation because the image get weird but still maintain sharp. Beyond MSE maintains some appearance but still struggles in deformation and blur.

Failure Case. A typical failure case in affine transformation model is that the motions between frames are plausible yet unexpected black pixels appear somewhere in the frames. We think this is caused by the empty pixels in intermediate images after applying affine transformations. In convolution model, one common failure mode is that some part of the objects lack resolution while the silhouettes remain recognizable. We believe a more powerful merge network would be a promising solution in both cases, and we leave this for future work.

Chapter 6
CONCLUSION

In this paper, we have presented a new framework to synthesize multiple videos from one single image. Specifically, our framework uses transformation generation to model the motions between frames, and reconstructs frames with those transformations in a volumetric merge network. We also present a novel evaluation metric to assess the reconstruction quality. We have demonstrated that our framework can produce plausible videos with state-of-the- art image quality on different datasets.