Future Video Synthesis with Object Motion Prediction
Supplementary Material

Yue Wu, Rongrong Gao, Jaesik Park, Qifeng Chen

1 Experiments on other datasets

We also conduct experiments beyond driving scenes on the BAIR robot pushing dataset [1] and the Penn Action dataset [3]. The BAIR dataset consists of videos about a robot arm pushing multiple objects. The Penn dataset has videos with various non-rigid human actions.

1.1 BAIR robot pushing dataset

For the BAIR dataset, we manually label five video sequences, consisting of 156 frames, to decompose the scene into a robot arm and objects on the flat. The robot arm is treated as a rigid object and motion prediction is approximated by a 2D affine transformation. The background is predicted using an optical flow prediction network. We use four sequences for training and one sequence for testing. Some visualization results are shown in Figure 1. Our method can preserve the shape of the robot arm quite well.

1.2 Penn Action dataset

For the Penn dataset, which is a human motion dataset, we roughly decompose a human body into multiple nearly rigid parts such as feet and torso. The motion of each part is approximated as 2D affine transformation. Since the Penn Action dataset only provides the keypoint coordinates, we use [2] to obtain joint segment masks. Some visualization results are shown in Figure 2. The results turn out not good. The shape of the human body is preserved, but the motion prediction is incorrect. Predicting human joints requires a more complex model to simulate the relations between different parts of human bodies.

2 Pixel-wise prediction and constraint.

We tried a similar idea: predicting the optical flow of the whole image while regularizing the flow for each object so that the flow values for each object should be similar.

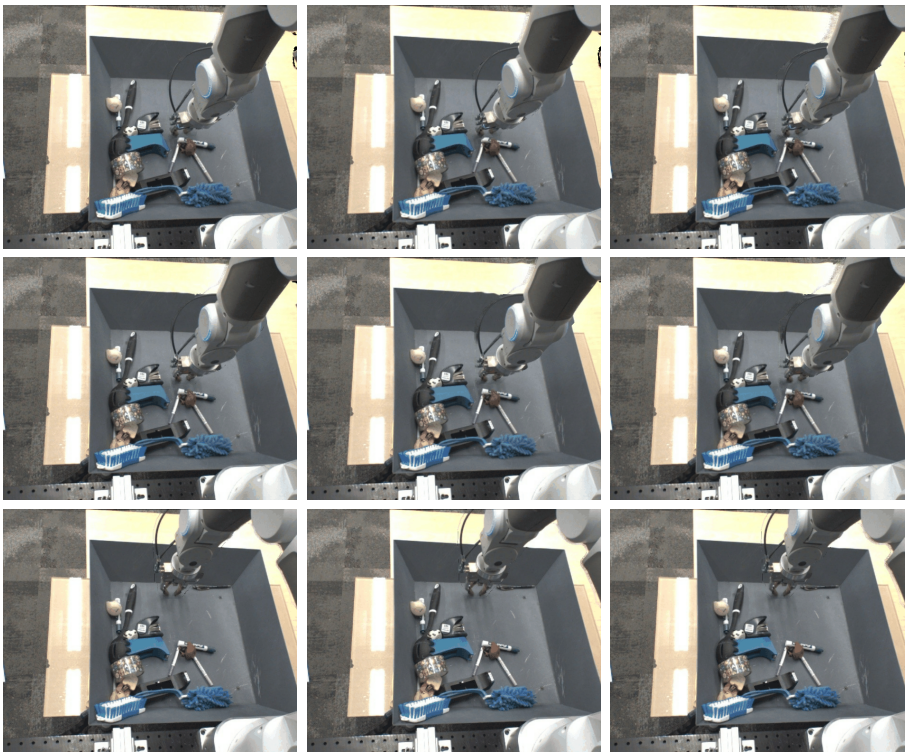

Figure 1: Results of predicting the frames $t+1$, $t+3$, and $t+5$ on the BAIR robot pushing dataset

However, the visual results were not good yet: the objects deform in inconsistent directions in the long term.

3 Advantages and limitations.

By forecasting object trajectories, our method can better preserve the high visual quality of objects in the future with less undesirable distortion. On the other hand, our approach is limited in highly non-rigid scenarios, such as predicting the future of waterfall.

References

[1] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *NeurIPS*, 2016. 1

[2] Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun. Cross-domain complementary learning using pose for multi-person part segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 1

$t+1$ $t+3$ $t+5$

Figure 2: Results of predicting the frames $t+1$, $t+3$, and $t+5$ on the Penn Action dataset

$t+1$ $t+3$ $t+5$

Figure 3: Results of predicting the frames $t+1$, $t+3$, and $t+5$ on the Cityscapes dataset using pixel-wise constraint loss.

[3] Weiyu Zhang, Menglong Zhu, and Konstantinos Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding". In *ICCV*, 2013. 1