videos with the same input image. Motions in those videos are notably dissimilar. Figure 8 gives an perception comparison among our framework and two baselines. The three methods are trained in same iteration. Obviously, generation from scratch as Baseline 2 needs much longer training time and $l2$ loss criterion as Baseline 1 not only make the result lacking of diversity, but also leads to blur because of intrinsic ambiguity of image.

UCF 101 Dataset

Dataset and setup: The former datasets are both synthetic datasets. For natural scene, we experiment on UCF101 dataset [25]. The dataset contains 13,320 videos with an average length of 6.2 seconds belonging to 101 different action categories. The original dataset are labeled for action recognition, but we do not employ those labels and instead we use the dataset in an unsupervised way. Videos with an average length of 6.2 seconds are cut into clips that each consists of five frames. We prepare 15,680 video clips for each category as training samples and 1,000 unseen image as testing samples. The video frames are reshaped to 128 128 and 64 64 for different resolution experiments. The convolutional kernel size is set to 16 and 9 accordingly.

Result: Figure 5 (c)(d) illustrate the qualitative performance in surfing class of different resolutions. Obviously our framework pro- duce fairly sharp frames. It successfully escapes from appearance deformation of surfer and wave. The difference images suggest that our framework can model plausible waving and surfing motions. The dynamic results seem rather realistic, so we strongly recommend a quick look at the small gif demo in supplementary material. Figure 9 shows the convergence curve of EM distance. We can see the curve decrease with training and converge to a small constant.

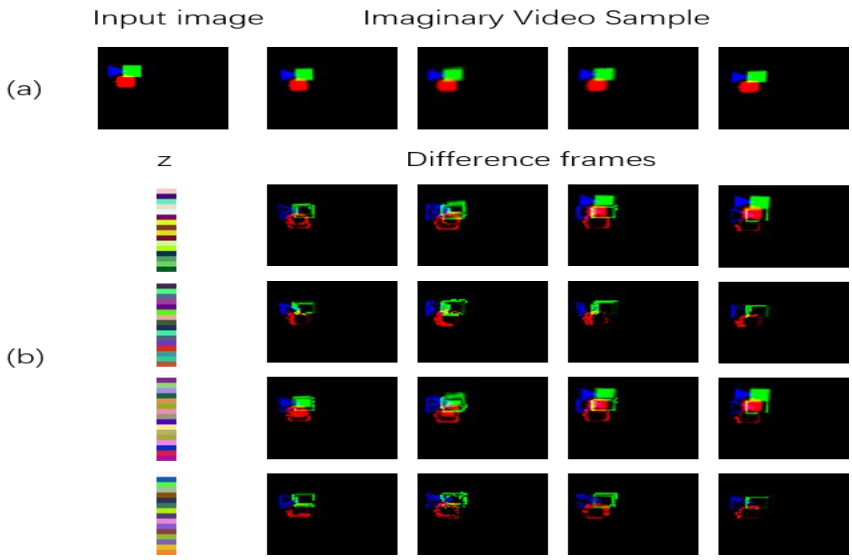

Figure 7: Diverse video imagination: multiple imaginary videos from same input image. (a) denotes the input image and one imaginary video sample as reference. The first column of (b) indicates different input zs, the rest columns shows the difference frames of imaginary video samples minus the reference. Each row of (b) illustrates a unique imaginary video and its unique z.

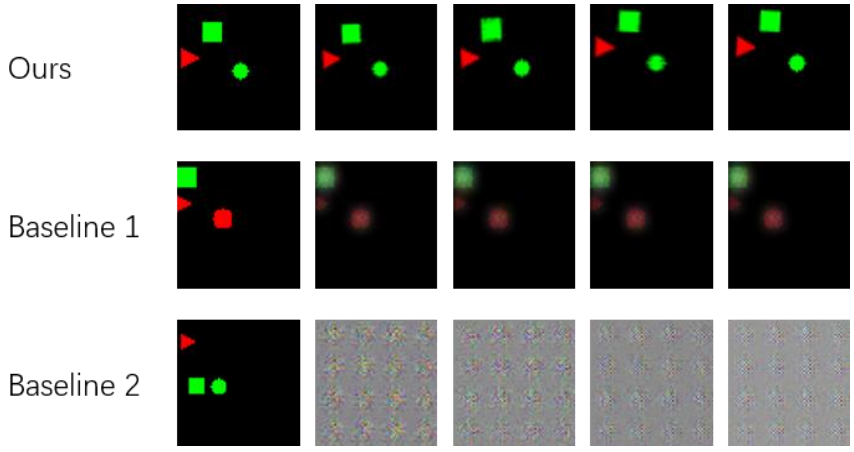

Figure 8: Synthesis result of custom Baselines With same fixed training iterations, our framework produce obviously be1er result. l_2 loss in baseline 1 brings blur. Baseline 2 that reconstruct pixel from noise needs much longer training time and cannot produce recognizable frames. The absolute of the constant is meaningless because the scale of EM distance varies with architecture of critic network.