Supplementary Material for Disentangling Multiple Features in Video Sequences using Gaussian Processes in Variational Autoencoders

Sarthak Bhagat[1], Shagun Uppal[1]⋆, Zhuyun Yin[2], and Nengli Lim[3]

[1] IIIT Delhi
{sarthak16189, shagun16088}@iiitd.ac.in
[2] Bioinformatics Institute, A*STAR, Singapore
yinzhuyun@gmail.com
[3] Singapore University of Technology and Design
nengli_lim@sutd.edu.sg

In this supplement, we include additional figures generated from the experiments performed in the main draft. We used a single RTX 2080 Ti GPU for training and inference in all the experiments.

1 Additional Results from Swapping Latent Channels

In this section, we provide additional figures which augment Figures 4, 6 and 7 in the main draft and that illustrate the ability of MGP-VAE to disentangle multiple factors.

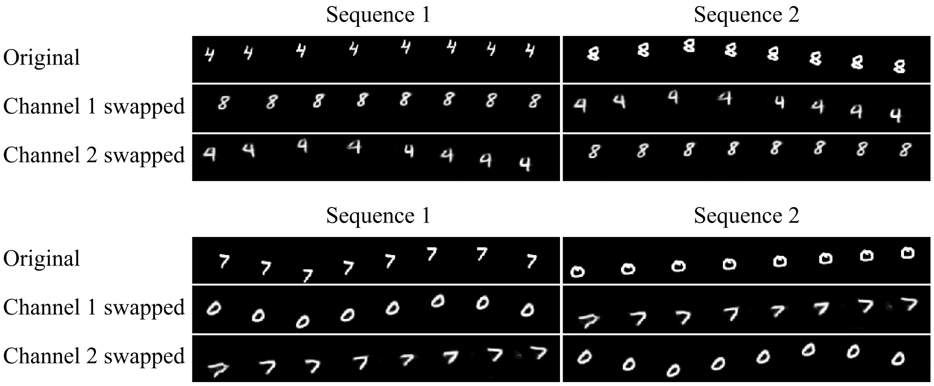


Fig. 1: Results from swapping latent channels in Moving MNIST; channel 1 (fBM($H = 0.1$)) captures digit identity; channel 2 (fBM($H = 0.9$)) captures motion.

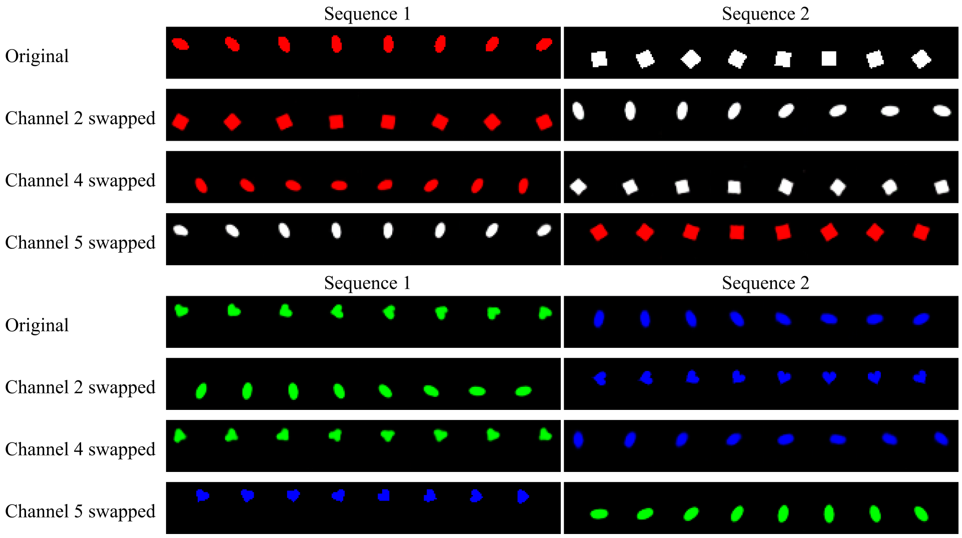


Fig. 2: Results from swapping latent channels in Coloured dSprites; channel 2 captures shape, channel 4 captures orientation / position, and channel 5 captures color.

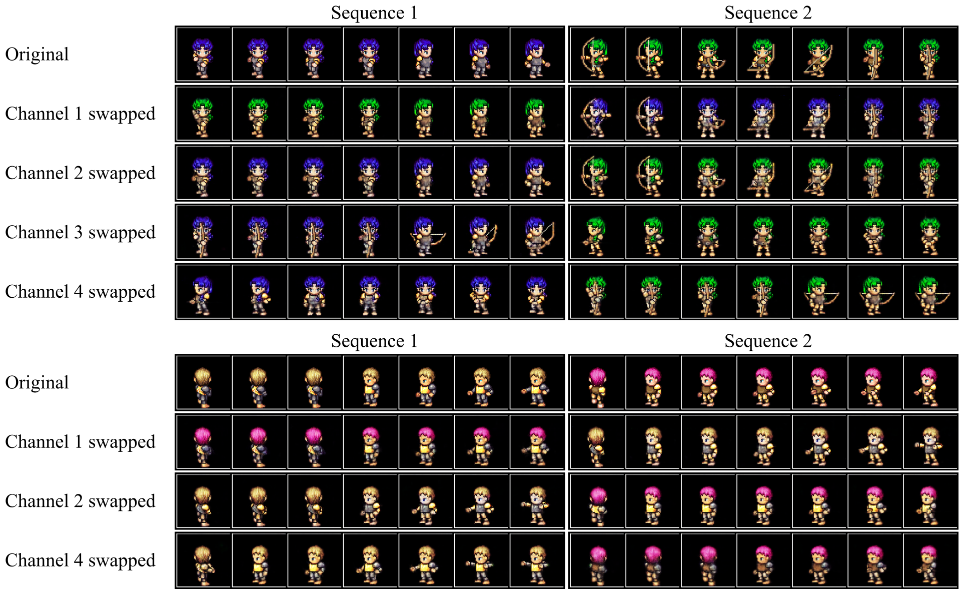


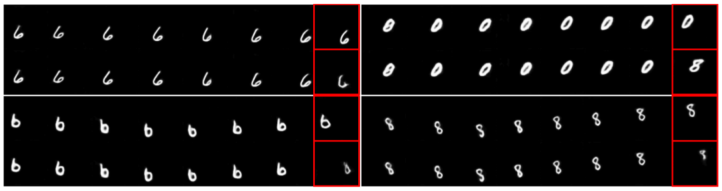
Fig. 3: Results from swapping latent channels in Sprites; channel 1 captures hair, channel 2 captures armour color, channel 3 captures weapon and channel 4 captures pose.

2 Video Prediction

2.1 Qualitative Evaluation of the Geodesic Loss Function

The figures in this section augment Figure 8 in the main draft and highlight the qualitative improvements from using the geodesic loss function in the video prediction task.

Without geodesic loss



With geodesic loss

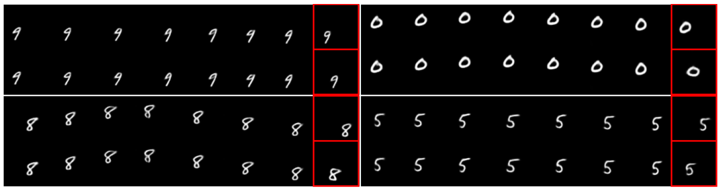
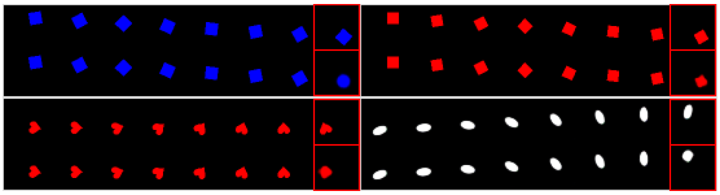


Fig. 4: Comparison between predictions with and without using the geodesic loss function for Moving MNIST.

Without geodesic loss



With geodesic loss

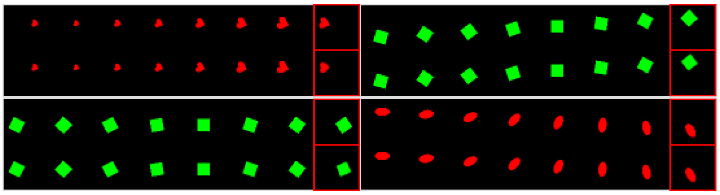
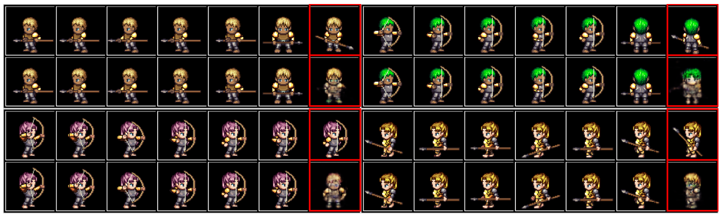


Fig. 5: Comparison between predictions with and without using the geodesic loss function for Coloured dSprites.

Without geodesic loss



With geodesic loss

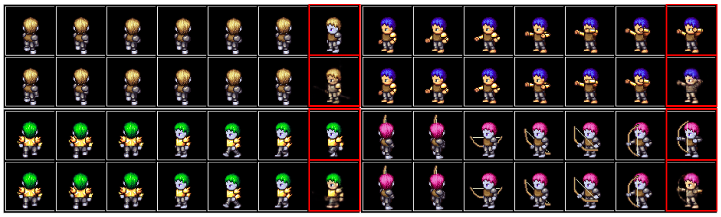
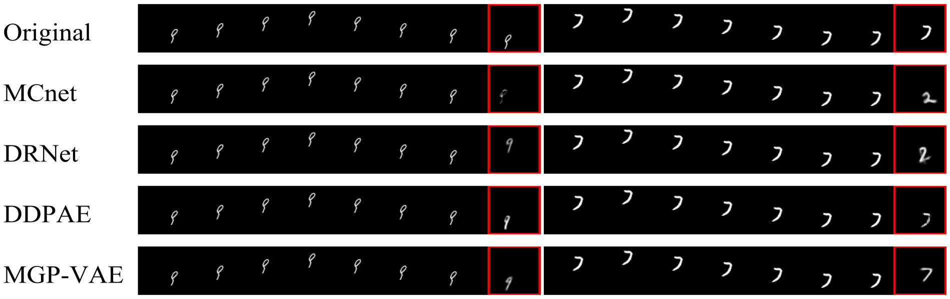


Fig. 6: Comparison between predictions with and without using the geodesic loss function for Sprites.

2.2 Comparison of MGP-VAE with the Baselines

Here, we show qualitative results of MGP-VAE versus the various baselines for the prediction experiment run in Section 4 of the main draft. The figures below depict the original video sequences along with the last frames predicted by the different models, marked in red. In general, MGP-VAE's predicted frames match the originals more closely than the baselines.

Moving MNIST



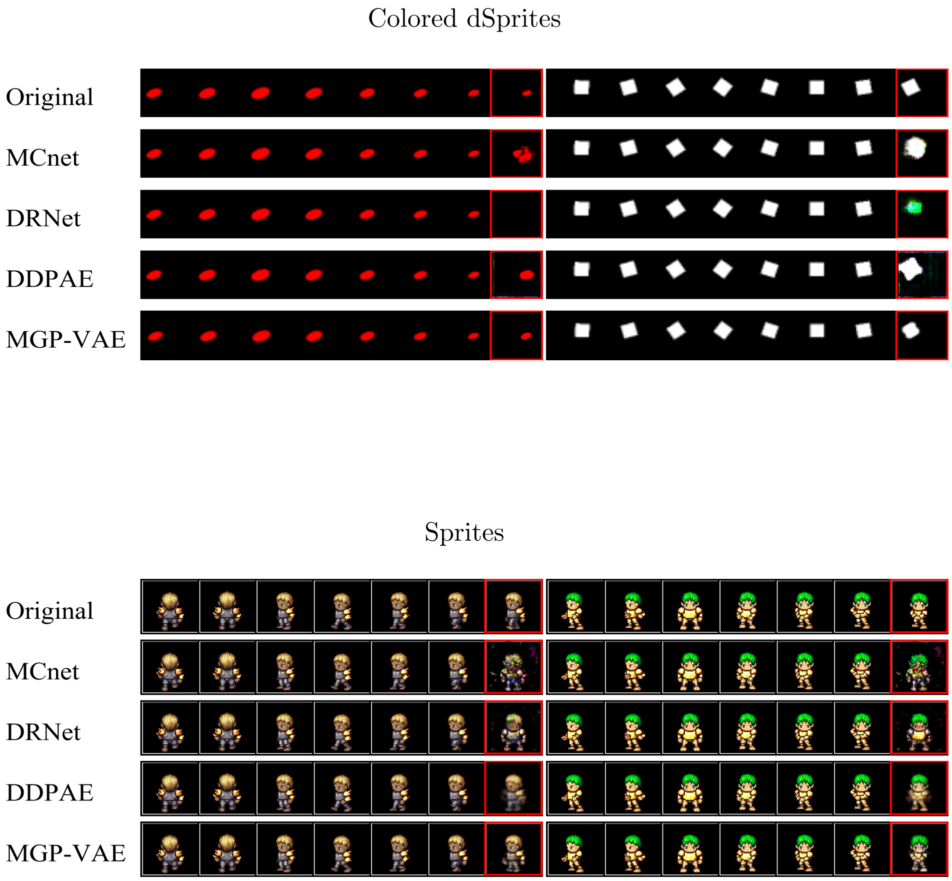


Fig. 7: Qualitative results of MGP-VAE and the baselines in the video prediction task. Predicted frames are marked in red, and the first row depicts the original video sequence.