1. Motivation

Region-of-interest (ROI) are important regions of an image or video. When compressing data ROIs should be more accurate than non-ROIs. We introduce two models for ROIs based neural video coding, integrated in a scale-space flow architecture. One model is an implicit model that is fed with a binary ROI mask. The other model is integrated with latent scaling to control the quantization bit widths, conditioned on the ROI mask. We show that our methods: (1) outperform baselines in terms of rate-distortion performance in the ROI, (2) can be trained with synthetic ROI masks with little to no degradation in performance and (3) generalize to different datasets at inference time.

4. Setup

5.1 Overall performance

We test all models trained and tested on DAVIS and show the results in rate-distortion plots (PSNR vs rate per pixel). SSF has better PSNR on non-ROI regions. All other models learn to spend more bits on ROI regions. LS ROI SSF has best tradeoff for ROI regions and is therefore further investigated.

5.2 Synthetic ROI masks

Training LS ROI SSF with synthetically generated perlin noise masks results in almost similar performance as when trained with true annotations. When no true annotations are available, these masks are worthwhile.

5.3 Generalization

To test the generalization of our best performing model LS ROI SSF, we train the model on DAVIS and test its performance on Cityscapes. We benchmark against a model trained and tested on Cityscapes. We find that our model has higher ROI performance than SSF.

5.4 Swapping instances

We use the DAVIS and Cityscapes dataset. Corresponding semantic maps are binarized where a selection of classes is chosen to be ROI. Synthetic masks: Generate Perlin noise blobs evolving continuously over time. Has no correlation with video but is easy to obtain. We compare our methods against:

1) SSF
2) ROI-aware loss
3) OBC SSF