ROI based Neural Video Compression

Yura Perugachi-Diaz,
Guillaume Sautière,
Davide Abati,
Yang Yang,
Amirhossein Habibian &
Taco Cohen

Qualcomm AI research

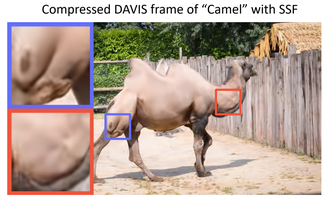
VU

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.

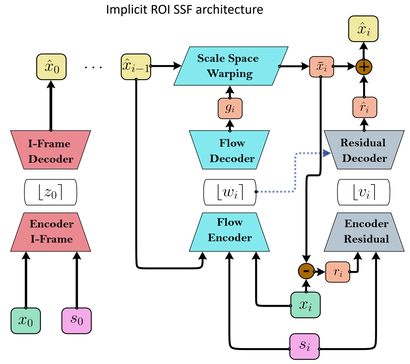
We introduce two models for ROIs based neural video coding, integrated in a scale-space flow architecture. One model is an implicit model that is fed with a binary ROI mask. The other model is integrated with latent scaling to control the quantization bin widths, conditioned on the ROI mask. We show 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.

Compressed DAVIS frame of "Camel" with SSF



2. Implicit ROI SSF

Implicit ROI SSF architecture



To make SSF ROI-aware, we feed ROI mask S_i as input to each of the three hyperpriors. Feeding the mask along with the video frame, encourages the model to focus on important aspects of the user. To optimize we adjust the distortion loss $\mathcal{L}_D$ of the SSF by modifying the mean squared error

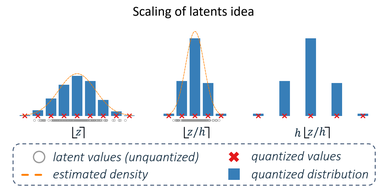
3. Latent-scaling ROI SSF

which is defined with the binary ROI mask:

$$\mathcal{L}_{D,i} = \frac{1}{HWC}\sum_{j=1}^{H}\sum_{k=1}^{W}\sum_{l=1}^{C}\left(s_i \odot \boldsymbol{\varepsilon}_i + \frac{1}{\gamma}\cdot(1-s_i)\odot \boldsymbol{\varepsilon}_i\right)_{jkl},$$

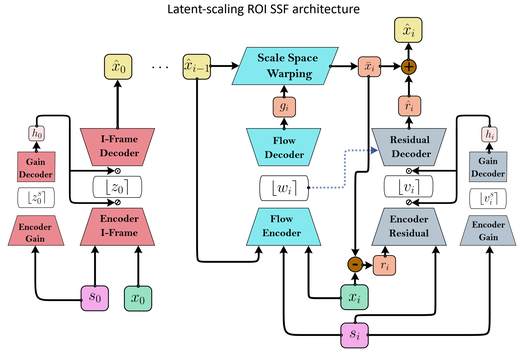
Where $\boldsymbol{\varepsilon_i} = (\boldsymbol{x_i} - \hat{\boldsymbol{x}}_i)^2$. We use the regular rate-loss for computation of the estimated cross entropy. $\mathcal{H}(\cdot)$.

Scaling of latents idea



Latent-scaling (LS) ROI SSF uses two extra hyperpriors that controls spatial bit allocation of the I-Frame and P-frame residual hyperprior. The key idea is that ROI-based information controls the scale of latents and the quantization grid can be explicitly adjusted.

Latent-scaling ROI SSF architecture

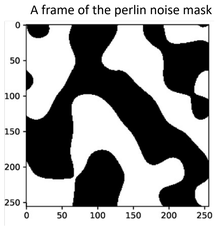


Therefore our model can learn that ROIs require finer quantization than non-ROI regions. For optimization we use the distortion loss of the implicit model. We adjust the rate-loss to include all latent variables of the extra hyperpriors:

$$\mathcal{L}_{LS,R} = \mathcal{H}(z_0^s) + \mathcal{H}(z_0) + \sum_{i=1}^{T}[\mathcal{H}(v_i^s) + \mathcal{H}(v_i) + \mathcal{H}(w_i)]$$

4. Setup

A frame of the perlin noise mask



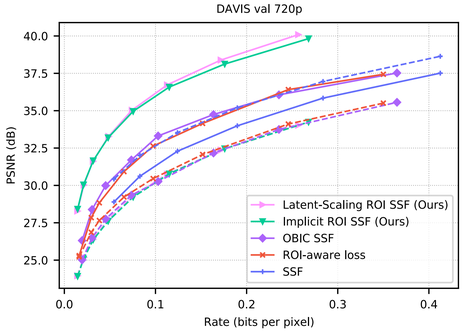
We use the DAVIS and Cityscapes dataset. Corresponding semantic maps are binarized where a selecion of classes is chosen to be ROI. Synthetic masks: Generate Perlin noise blobs evolving continuously over time. Has no correlation with video but is easy to obtain.

We compare our methods against:

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

5.1 Overall performance

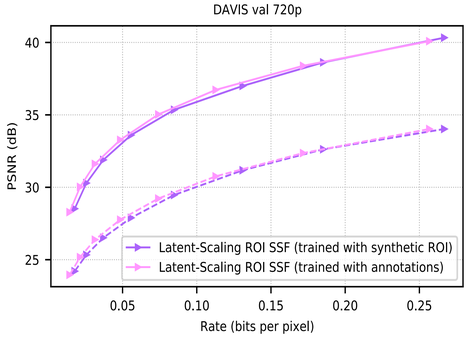
Overall performance
DAVIS val 720p



We test all models trained and tested on DAVIS and show the results in rate-distortion plots (PSNR versus bits 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

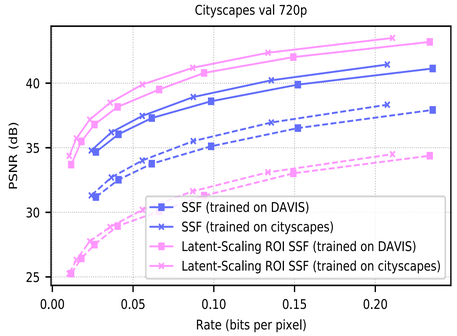
Trained with synthetic and true annotations
DAVIS val 720p



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

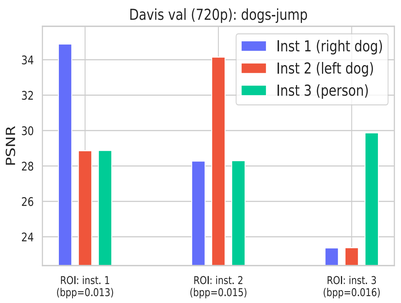
Generalization LS ROI SSF tested on Cityscapes
Cityscapes val 720p



To test the generalization of our best performing model LS ROI SSF, we train the model on DAVIS and test its performance on Citycapes. 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

Bar plot PSNR result of swapped instance



With a trained LS ROI SSF we swap instances during inference time. We find that ROI is compressed with higher PSNR. This indicates that we can control the sharpness of the preferred instance.