Hyperprior Contextual Video Compression (HyCoVC)

Hyperprior Contextual Video Compression (HyCoVC)

Anand Kumar

Institut für Informationsverarbeitung

09.08.2022

About Me

- Born and brought up in southern India
- Studying **Bachelors of Technology** in **Electrical and Electronics Engineering** at **National Institute of Technology, Tiruchirappalli, India**
- Interested in **Machine Learning** and **Computer Vision**
- Came across this Institute when I was searching for computer vision-based Institutes in Germany.

Outline

1. Introduction
2. Related Work
3. Approach & Architecture
4. Training: Loss function, Dataset and Parameters
5. Results
6. Further Improvements
7. References

Introduction

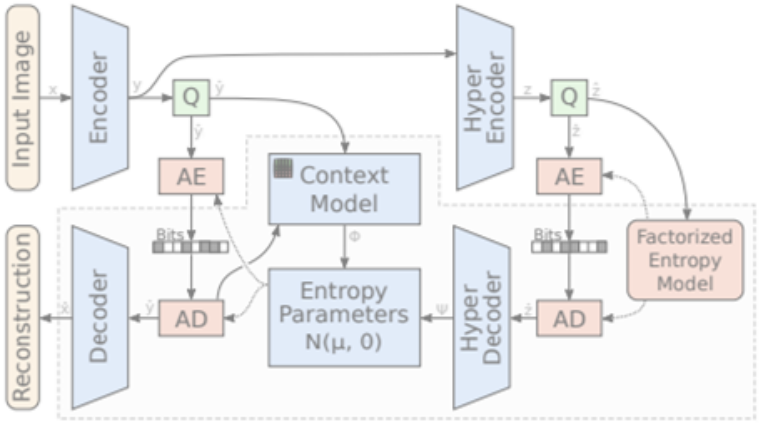
What is Hyperprior Contextual Video Compression (HyCoVC)?

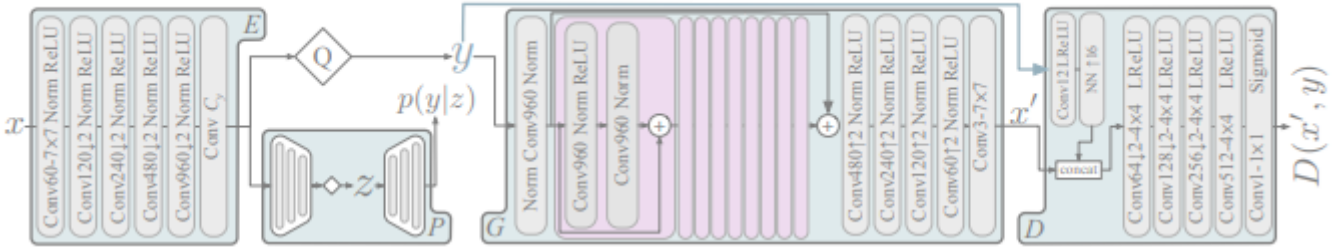
Introduction

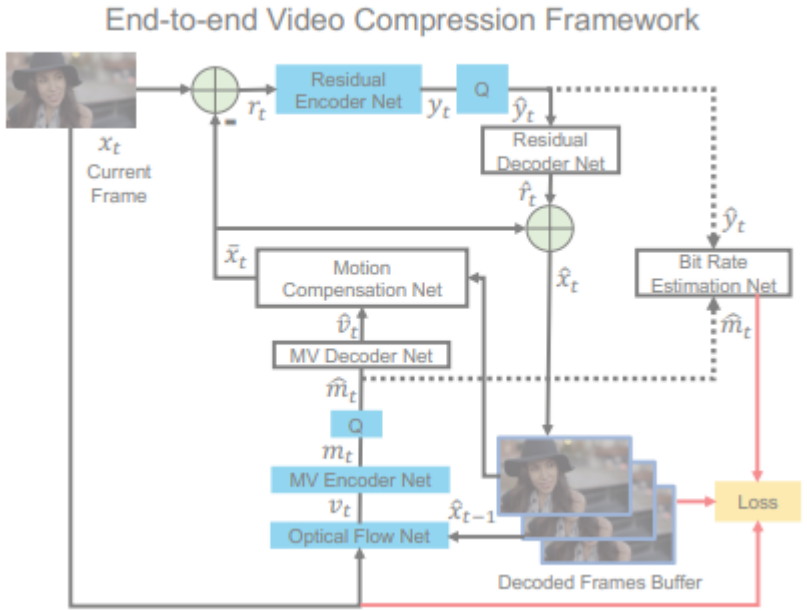
- The focus is to construct an end-to-end optimized deep video compression network using Generative Adversarial Networks (GANs)

- Hyperpriors and Contextual Layers are applied to reduce the bitrate required for lossless compression using arithmetic coding for the same quality.

- Perceptually similar and visually video frames are obtained, and the network operates at a broad range of bitrates.





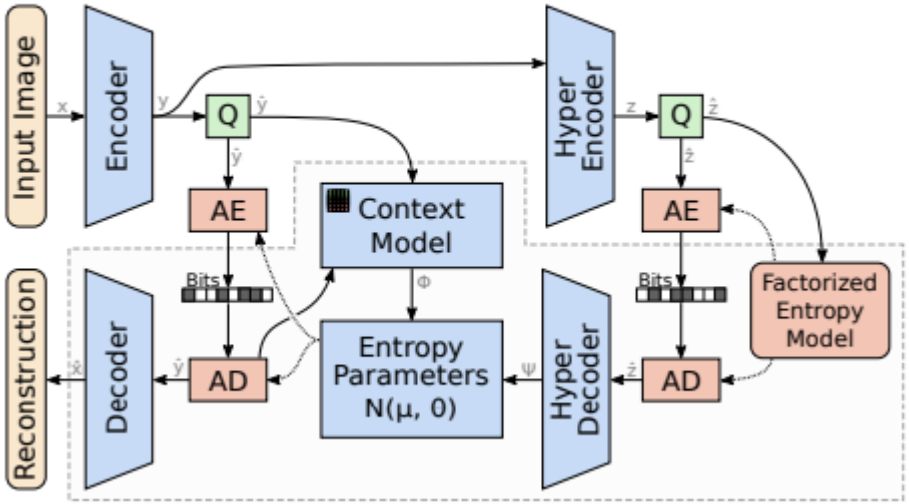


Related Works

Image and Video compression networks

Joint Autoregressive and Hierarchical Priors for Learned Image Compression

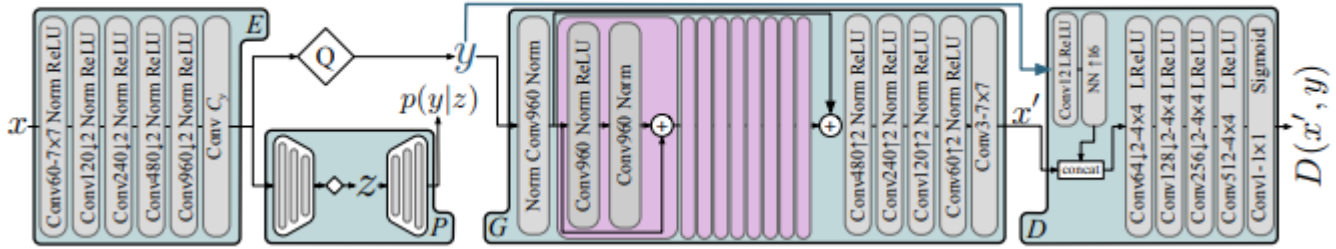
- Usage of Hyperpriors to improve quality of reconstructed image
- Hyperprior network is combined with context layer for a probability models for the latents



Minnen, David, Johannes Ballé, and George D. Toderici. "Joint autoregressive and hierarchical priors for learned image compression." *Advances in neural information processing systems* 31 (2018).

High-Fidelity Generative Image Compression

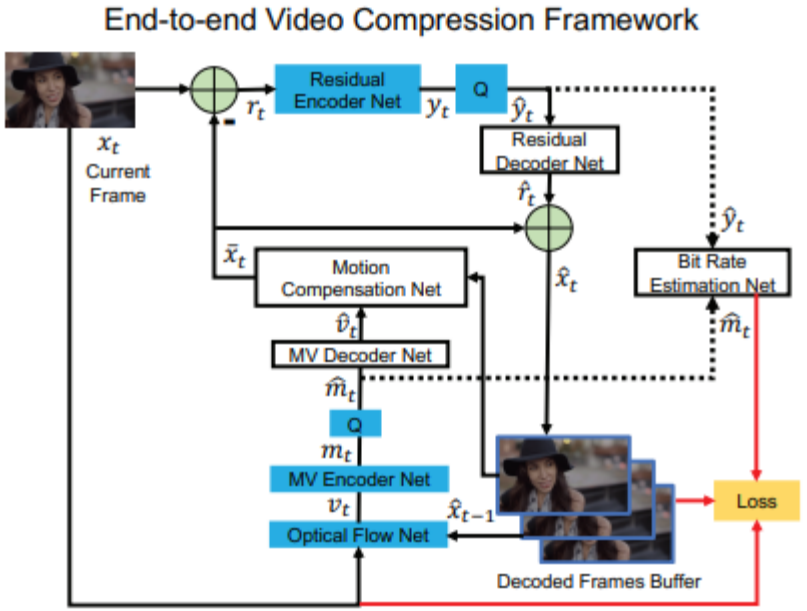
- Usage of Hyperpriors to improve quality of reconstructed image
- Improvement of loss function using Learned Perceptual Image Patch Similarity (LPIPS) to generated more visually pleasing images
- Conditional Discriminator



Mentzer, Fabian, et al. "High-fidelity generative image compression." *Advances in Neural Information Processing Systems* 33 (2020): 11913-11924.

DVC: An End-to-end Deep Video Compression Framework

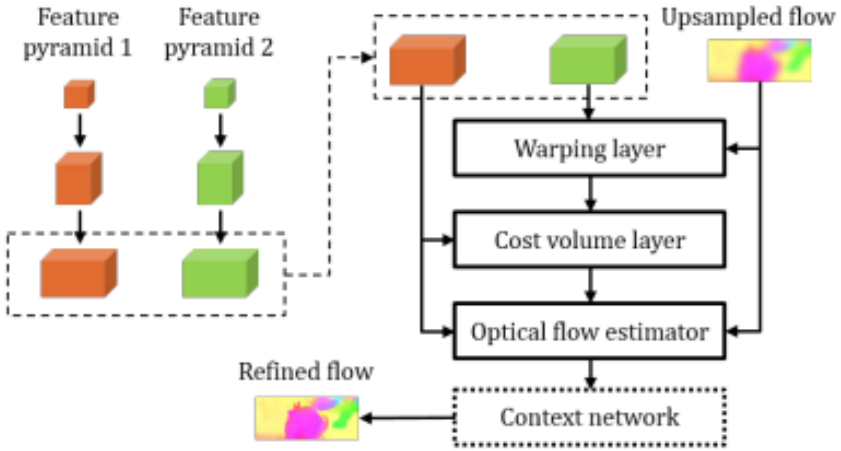
- Motion estimation using SpyNet and motion compensation using decoded frames
- Possibility of Bidirectional Prediction



Lu, Guo, et al. "Dvc: An end-to-end deep video compression framework." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume

- Motion Estimation network better than SPYNet
- Uses cost volume in the loss function to better predict motion flow fields.



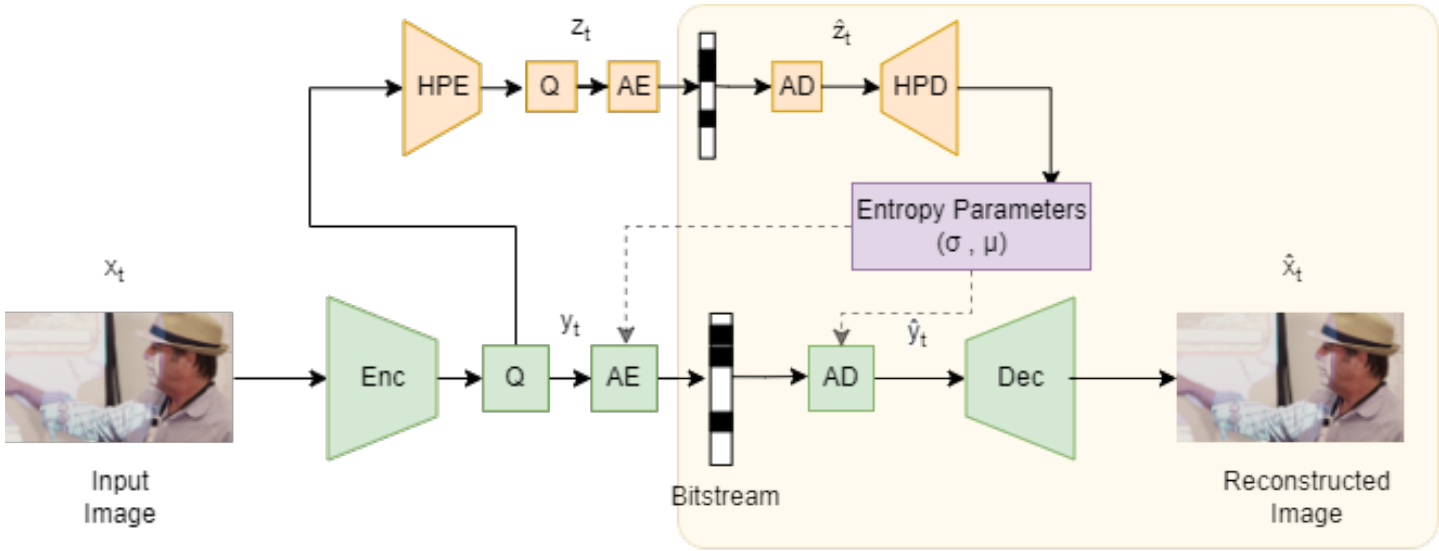
Sun, Deqing, et al. "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Approach & Architecture

Architecture of Hyperprior Contextual Video Compression (HyCoVC)

Baseline Network using HyperPriors

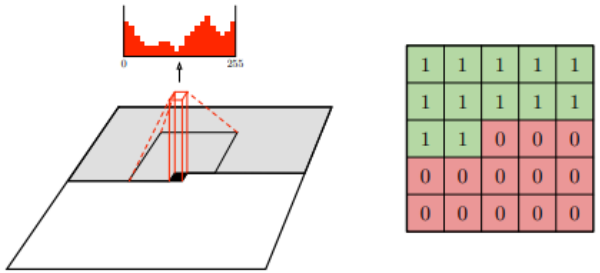
- Usage of Hyperpriors to improve quality of reconstructed image.
- The entire network is trained and the region covered by the yellow box is used for decoding.
- The entropy parameters of the latents are predicted using hyperlatents

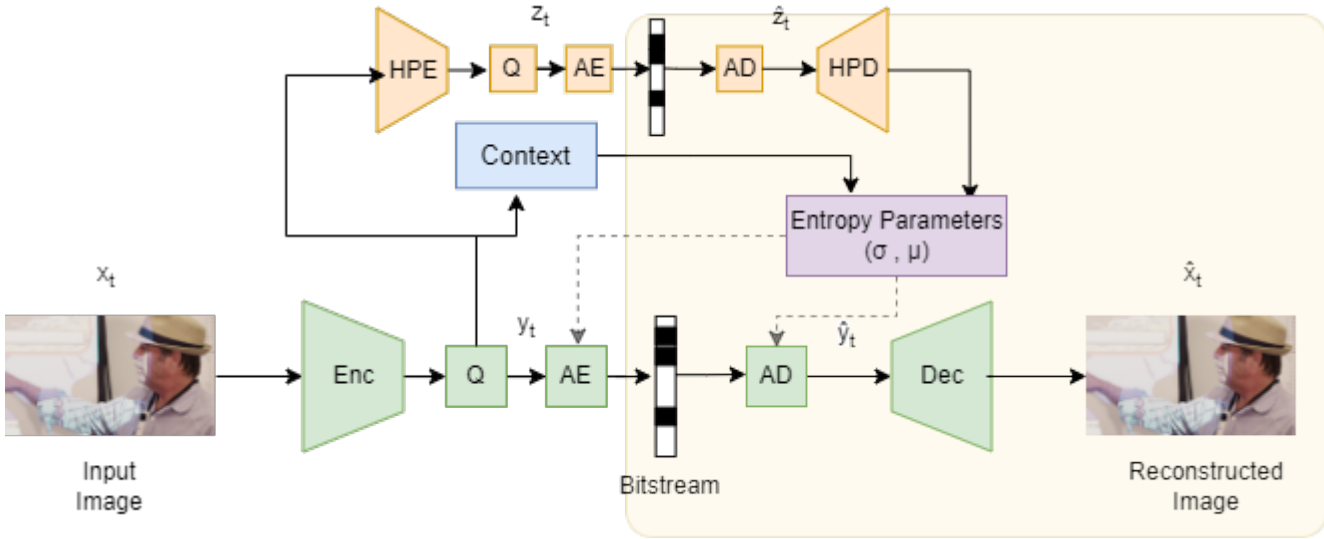


Context Network using Context Model

- Usage of context model to improve quality of compression and get better bitrates.
- The entropy parameters of the latents are predicted using hyperlatents and context model.
- The context model consists of masked convolution layer so that it predicts based on decoded pixels.

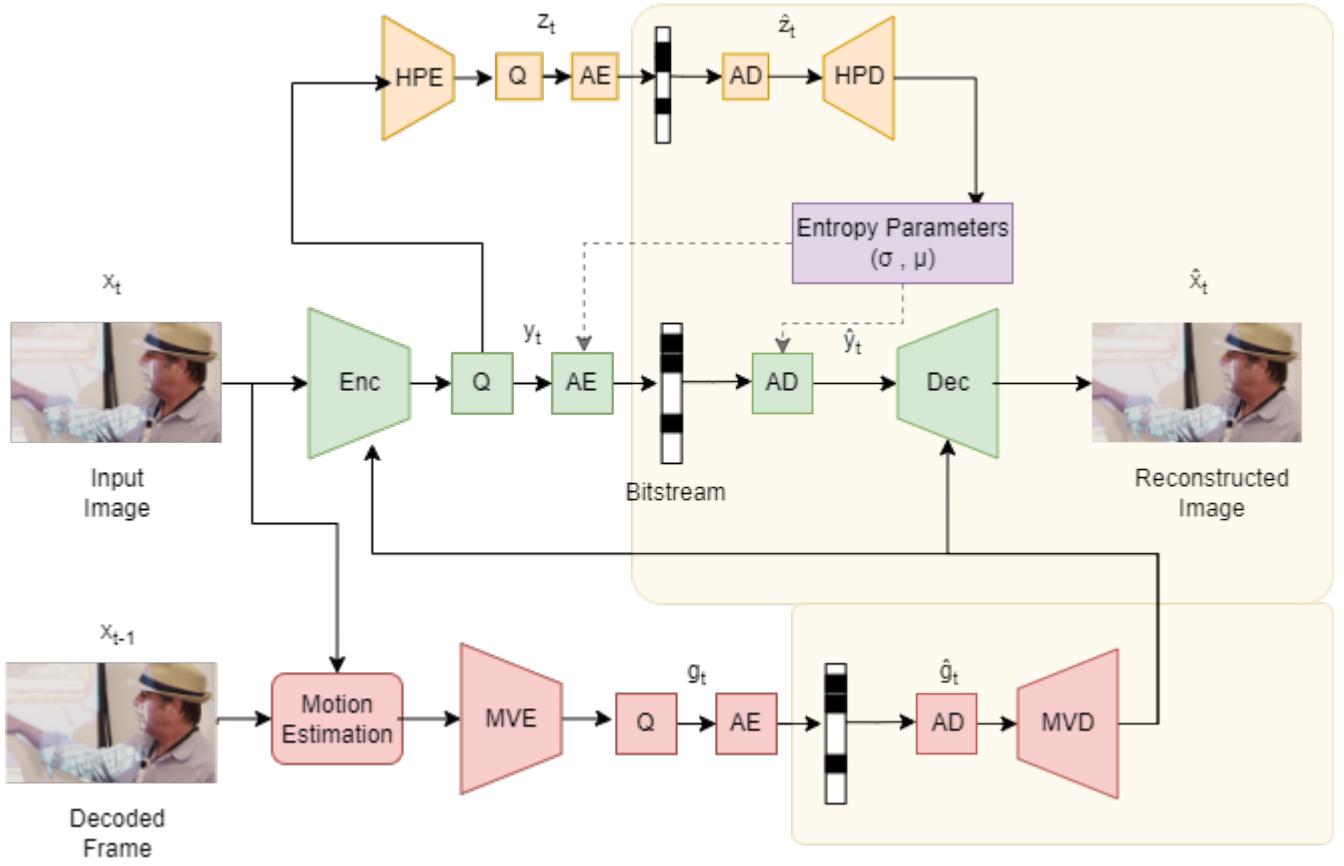
Masked Convolution Layer





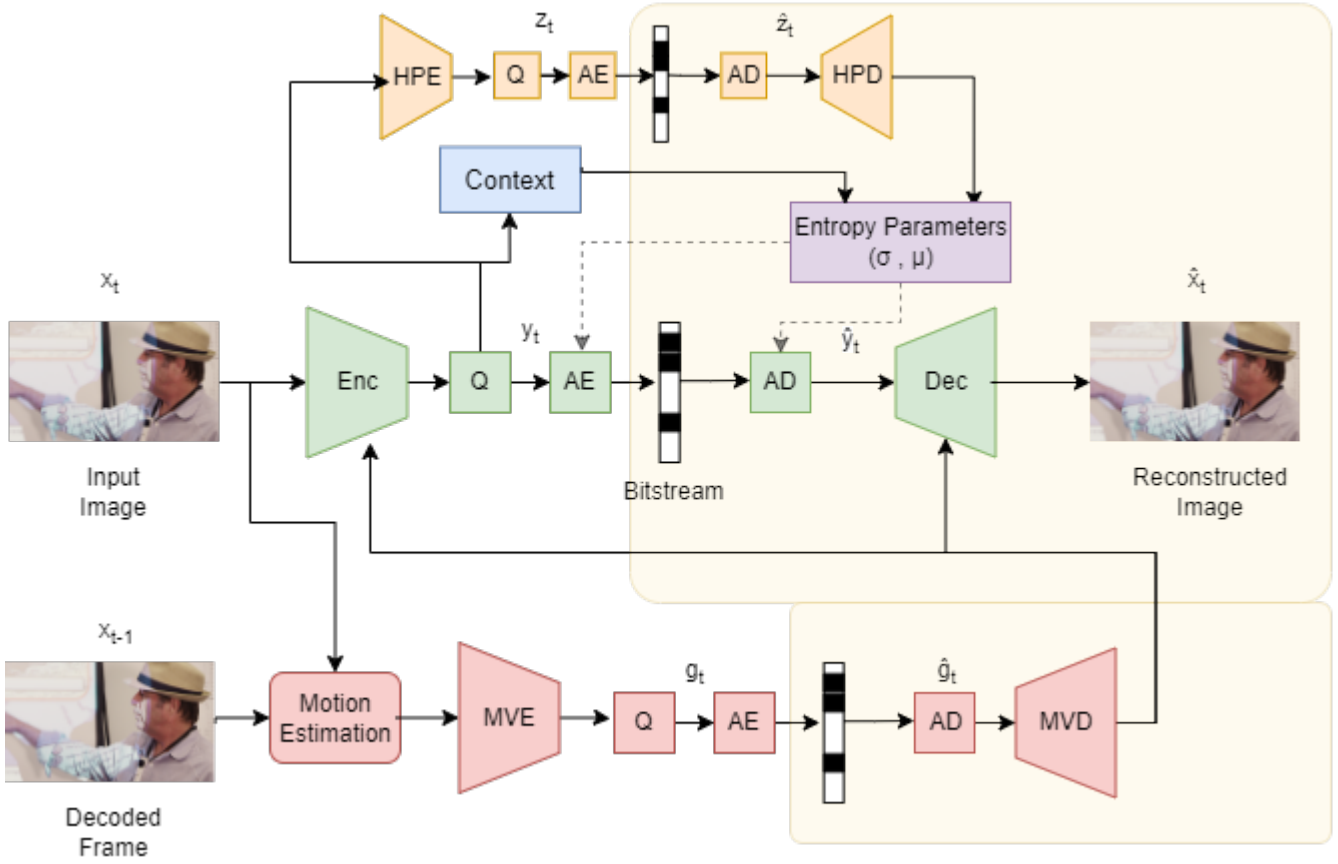
PWC Network using PWC Net to predict Motion Vectors

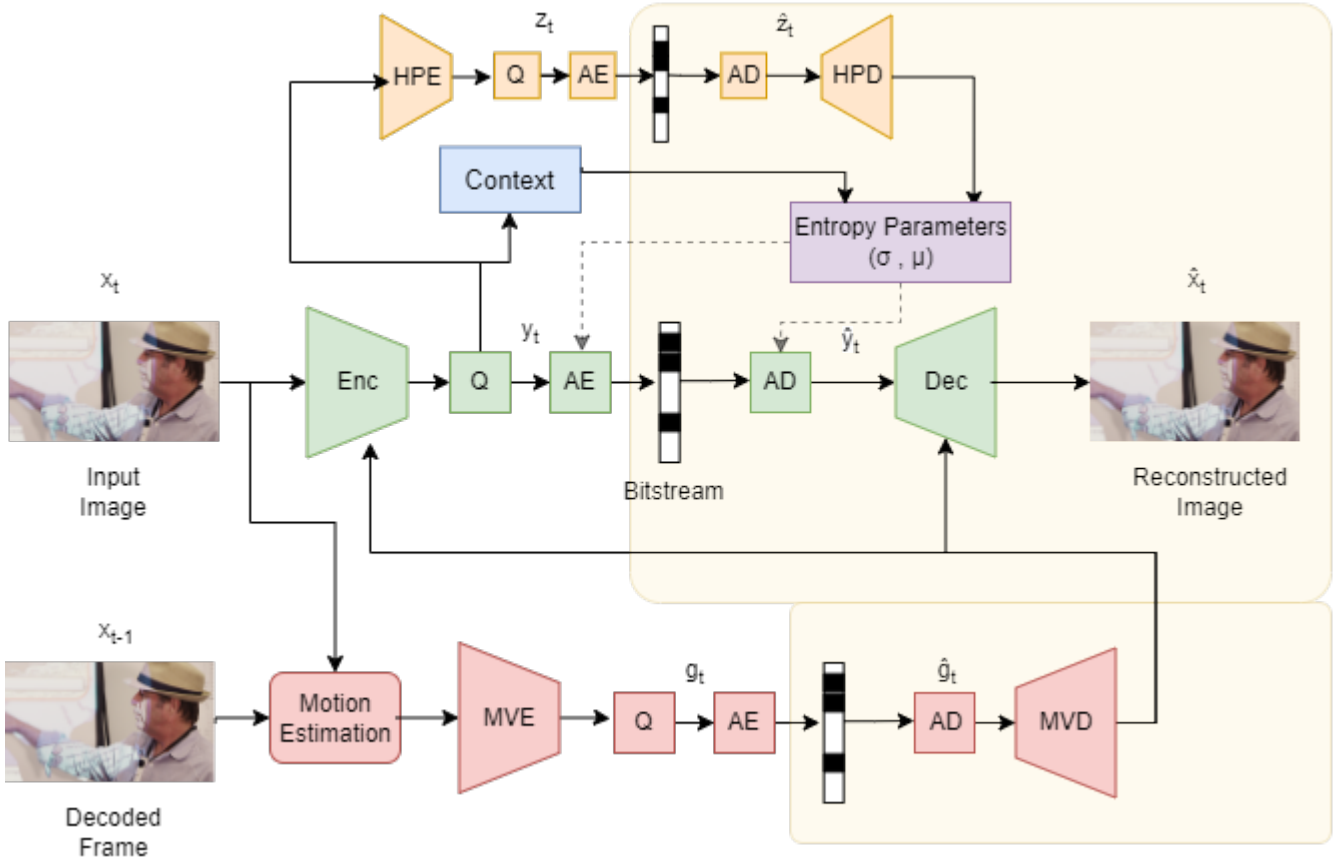
- Moving to video compression models from the image compression
- The motion vectors are predicted using the previous decoded frame and the input frame using PWC network



HyCoVC Network using Context Model on the PWC Network

- Produces better results compared to the PWC network due to the addition of context model





Final Architecture

Training : Loss function, Dataset & Parameters

Training of Hyperprior Contextual Video Compression (HyCoVC) Network

Loss Function

Rate

Distortion

$$\mathcal{L}_{EGP} = \mathbb{E}_{x\sim \mathrm{p}_x}\left[\lambda\, r(y) + d(x, \quad \hat{x}) - \beta \log\left(D(\hat{x}, y)\right)\right]$$

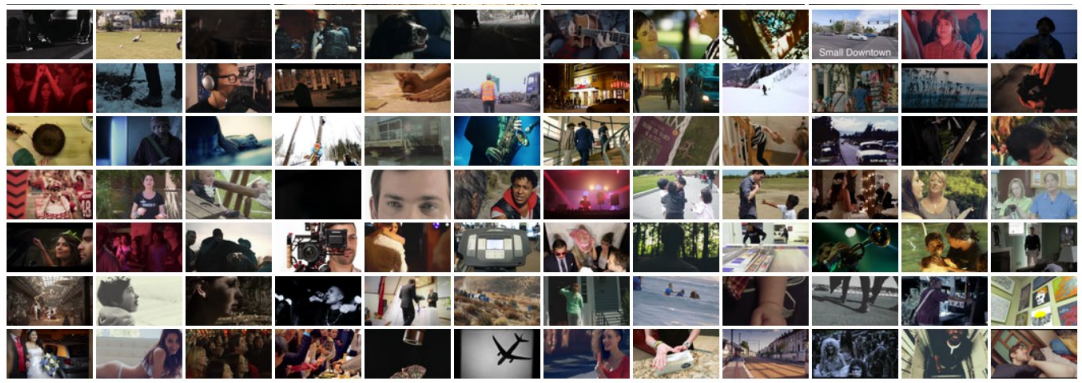
$$\mathcal{L}_D = \mathbb{E}_{x\sim \mathrm{p}_x}\left[-\log\left(1 - D(\hat{x}, y)\right)\right] + \mathbb{E}_{x\sim \mathrm{p}_x}\left[-\log\left(D(x, y)\right)\right]$$

Where r is the Shannon entropy function to calculate bitrate,
$d = k_M MSE + k_p d_p$ where k_m and k_p are weights and d_p is perceptual loss
And $\lambda\ and\ \beta$ are hyperparameters to control rate and distortion trade off

- There two lambdas which are selected based on whether the rate is above or below the required bitrate.

Dataset Used and Parameters

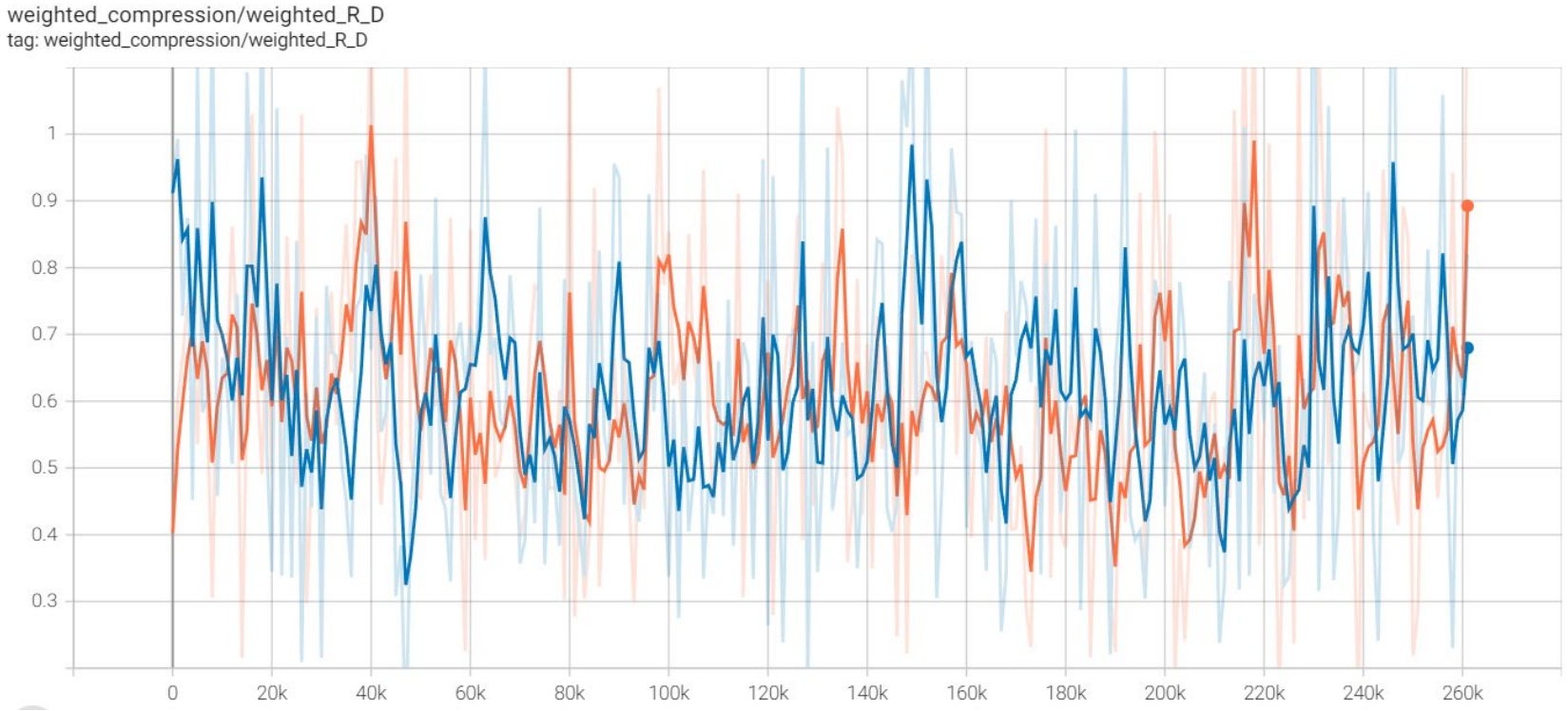
- Vimeo 90k- dataset was used for the project.
- It consists of 6 frames from 90k videos each of resolution 477x256 split into train and test set.



- The input images were cropped to **256x256**
- Each model was run for **4 epochs**
- The results were evaluated on **PSNR, MS-SSIM and LPIPS metrics**

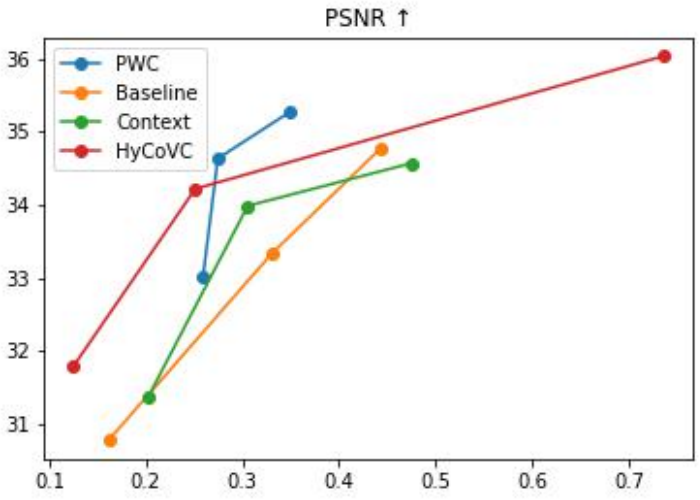
Results

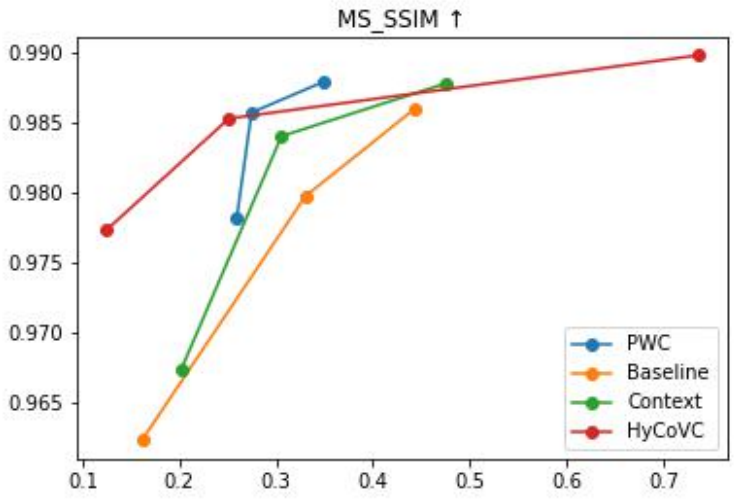
Performance of Hyperprior Contextual Video Compression (HyCoVC) Network

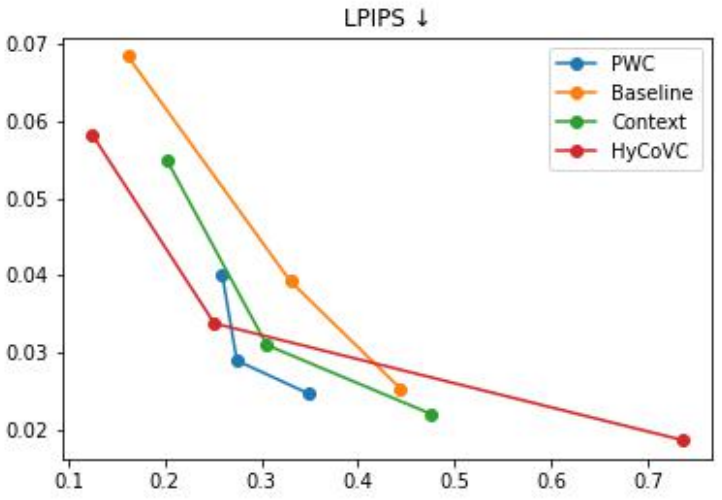


Rate-Distortion loss over 4 epochs

Blue is Train dataset and Orange is Test dataset

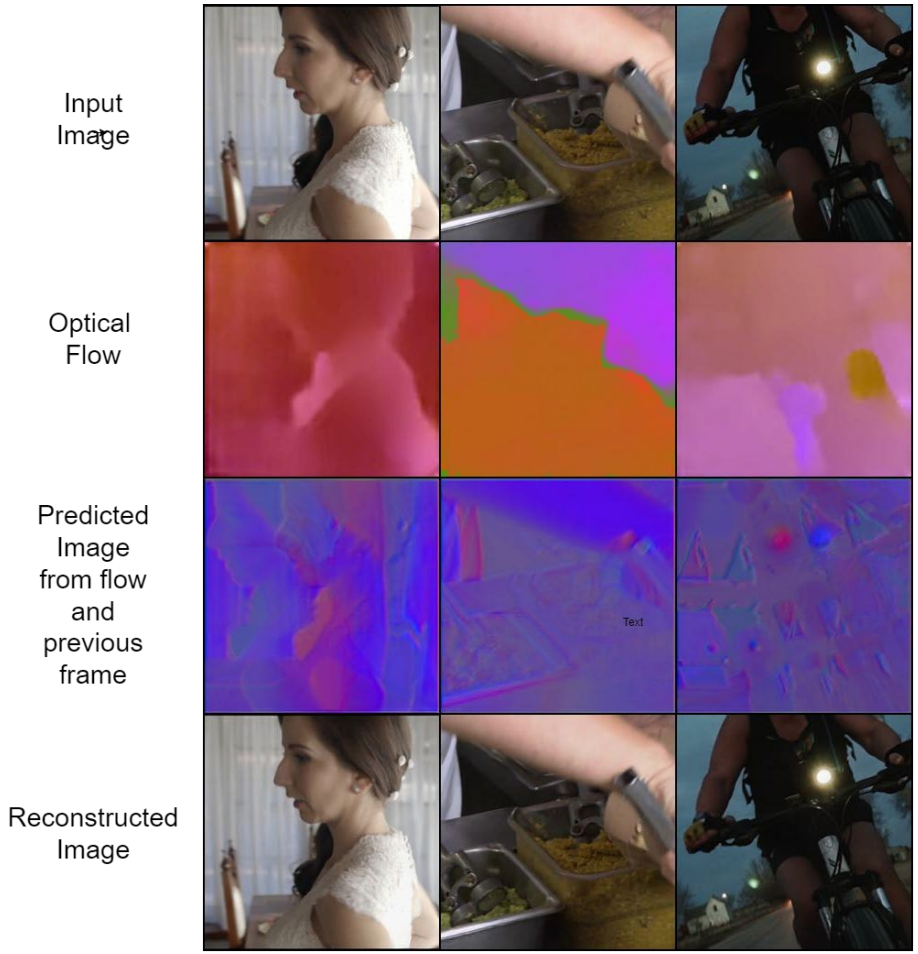






Results of each network architecture

Baseline is the hyperprior network alone, Context is Baseline with contextual layer, PWC is the hyperprior network with PWC prediction network and HyCoVC is PWC network with contextual layer



Visual Samples after each step

The first row shows the input image, the second image shows the optical flow predicted by PWC net, the third row shows the predicted image from flow and final row is resulting generated image

Further Improvements

Areas not yet explored

Improvements

- The information of the predicted frame can be added to the entropy parameter network.
- Taking for decoded frames into consideration can reduce the bitrate even more.
- Training the models for more time with produce better and more consistent results.

References

References

- Minnen, David, Johannes Ballé, and George D. Toderici. "Joint autoregressive and hierarchical priors for learned image compression." *Advances in neural information processing systems* 31 (2018).
- Mentzer, Fabian, et al. "High-fidelity generative image compression." *Advances in Neural Information Processing Systems* 33 (2020): 11913-11924.
- Lu, Guo, et al. "Dvc: An end-to-end deep video compression framework." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
- Sun, Deqing, et al. "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018

THANK YOU