Higher fidelity perceptual image and video compression with a latent conditioned residual denoising diffusion model

Jonas Brenig, Radu Timofte – Computer Vision Lab, University of Würzburg

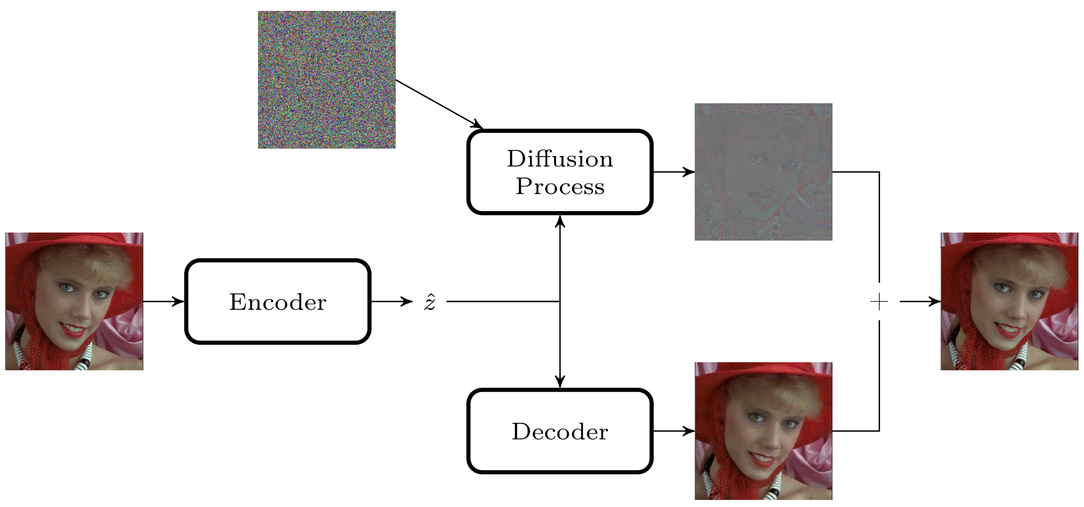
Motivation

- Learned image compression optimized for **perceptual quality** usually use GANs
- **Diffusion models** have shown impressive results in many generative tasks
- Existing diffusion-based compression method CDC [1] shows good perceptual quality, but at significantly lower PSNR

Contributions

- We propose a method **combining** the perceptual quality of diffusion with advantages of auto-encoder-based methods
- The approach can be **adapted** to various (image or video) learned compression methods

Method



- Train an **auto-encoder** to create a distortion optimized image reconstruction
- Use a **diffusion model** based on CDC [1] to improve the initial reconstruction, optimizing for perceptual quality
- The diffusion model is **conditioned on the encoder latent** and predicts the **residual** between the decoder output and ground-truth image

Training

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{bitrate} + (1 - \rho) \cdot \mathcal{L}_{dist} + \rho \cdot \mathcal{L}_{perc} + \mathcal{L}_{decoder}$$

- Encoder, decoder and diffusion-unet are **trained together**
- The decoder is explicitly optimized for MSE

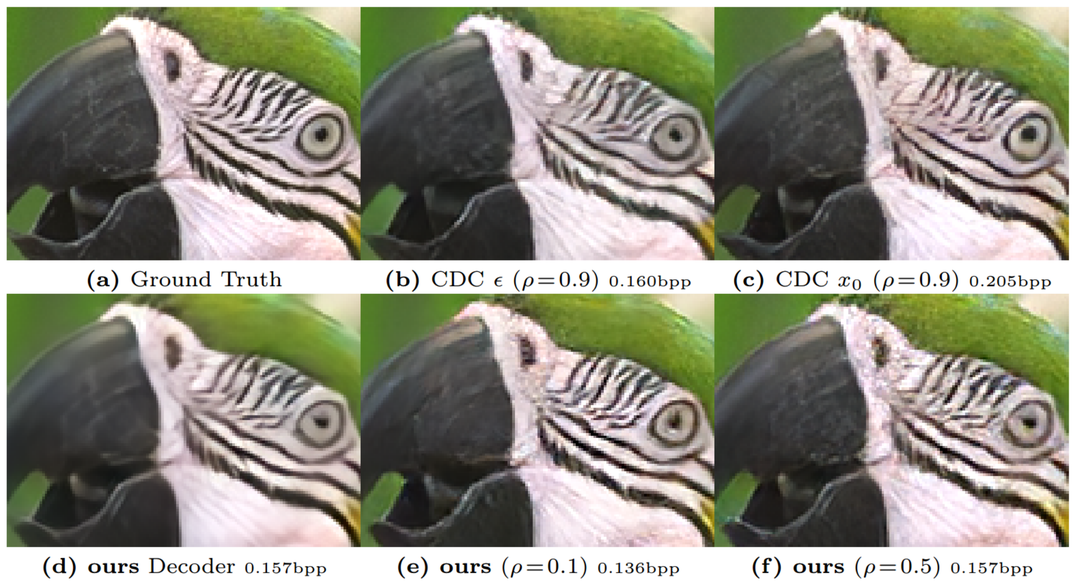
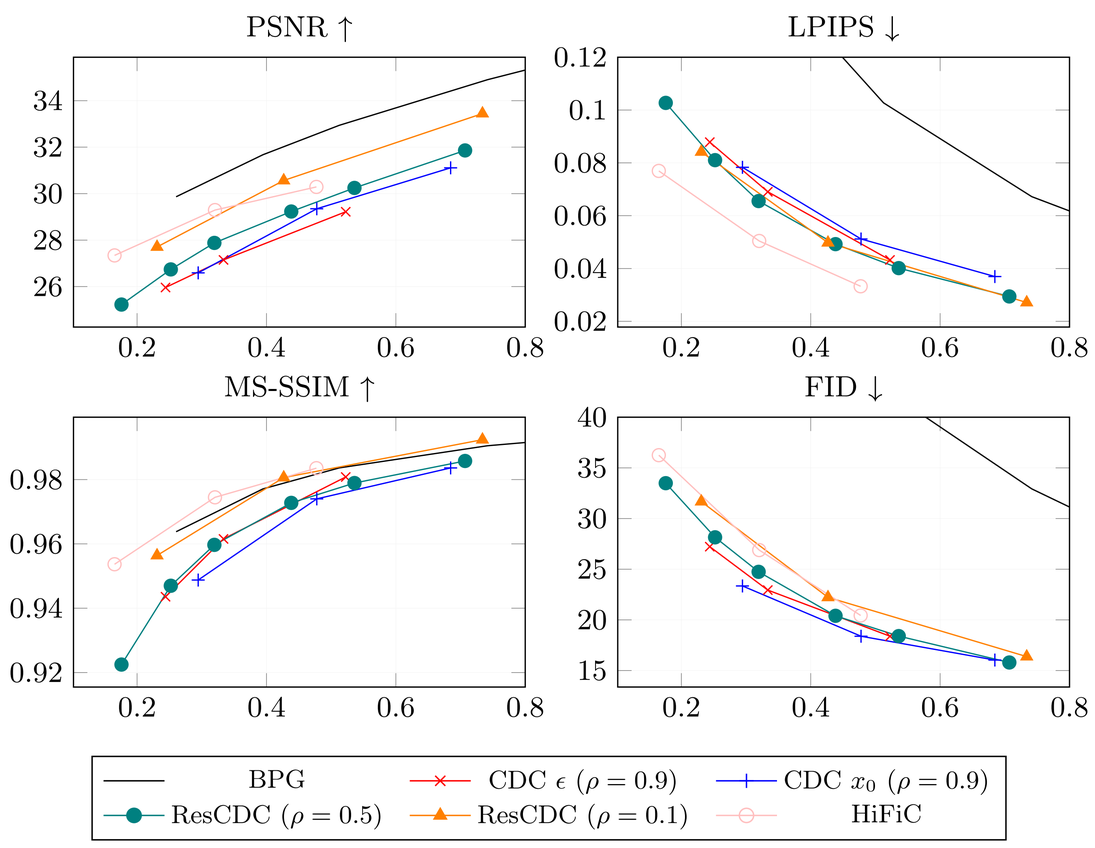


Image Compression

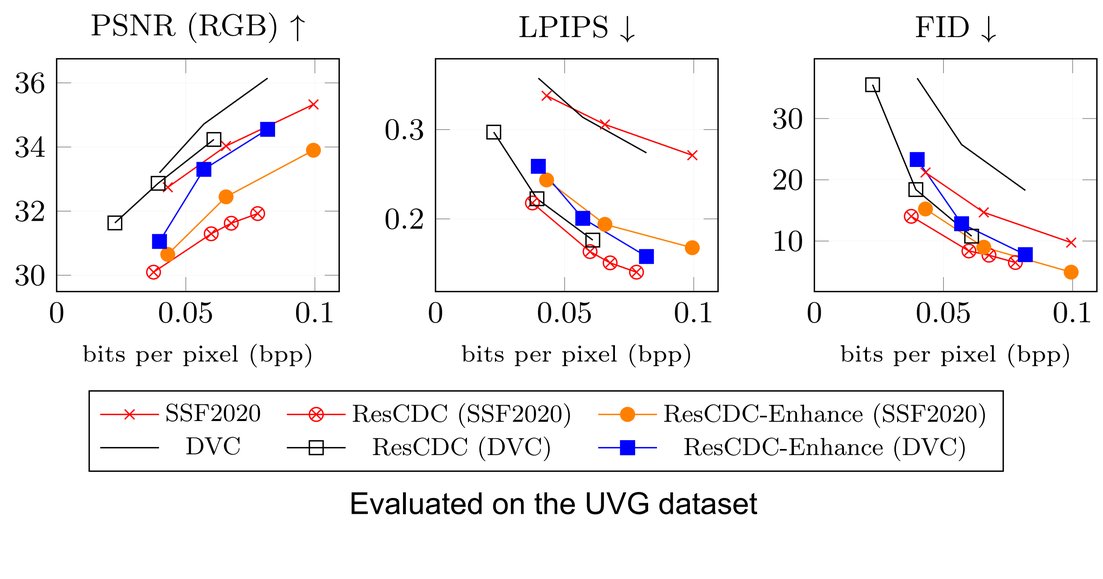
- The proposed model significantly **improves PSNR** over CDC
- With comparable perceptual quality (depending on the metric)
- Results achieved with 100 decoding steps
- (For the CDC baseline 500 and 17 decoding steps)



Evaluated on the Div2k dataset

Video Compression / Enhancement

- The model can easily be **extended to video** compression
- The encoder-decoder backbone is replaced by a **video compression method**
- Includes **previous frame** in diffusion unet input for improved temporal consistency
- **Finetuned** on pretrained checkpoints of existing video compression methods
- Improving **perceptual quality** at the cost of PSNR
- Results achieved with 10 decoding steps
- Can be used as an enhancement method by freezing the video-codec during training



Evaluated on the UVG dataset

[1] Yang, R., Mandt, S.: Lossy image compression with conditional diffusion models. Advances in Neural Information Processing Systems (2024)