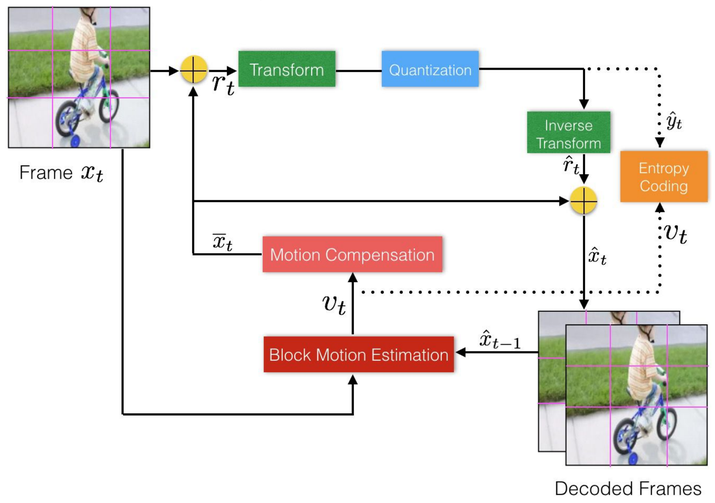
Deep Generative Video Compression

Jun Han*[1] , Salvator Lombardo*[1], Christopher Schroers[1] and Stephan Mandt[2]
1: Disney Research 2: UC Irvine
*: shared first authorship

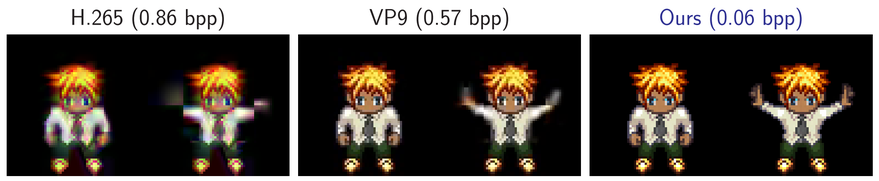
Video Codecs & Motivation

- **Traditional Codecs:** H.264/H265; VP9.



- **Pipelines**
  1. Motion v_t from x_t and $\hat{x}_{t-1}$
  2. Predict $\bar{x}_t$ and residual $r_t = x_t - \bar{x}_t$
  3. Transform and quantize r_t to $\hat{y}_t$
  4. Inverse $\hat{y}_t$ to residual $\hat{r}_t$
  5. Entropy coding v_t and $\hat{y}_t$
  6. Reconstruct $\hat{x}_t = \bar{x}_t + \hat{r}_t$
- **Limitations:**
  1. Simple motion estimation; simple transform
  2. Blocky artifacts in low-bit rate regime

Codecs Comparison



bpp: bit per pixel; original: $8 \times 3 = 24$ bits per pixel.

- **Ours**: much lower bit rate; no blocky artifact.

Deep Image Compression (Balle, ICLR 2017)

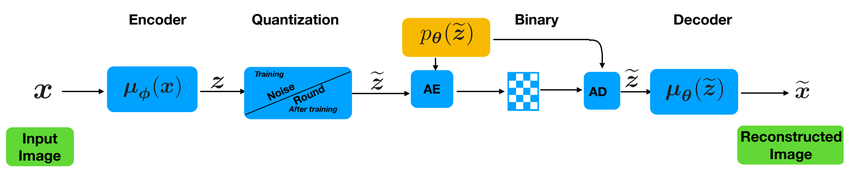


Figure: $\tilde{z}$: adding noise (training) or rounding (after training). Prior $p_\theta(\tilde{z})$: parametric form to fit data distribution for entropy coding.

- Inference model: $\tilde{z} \sim q_\phi(z \mid x) = \mathcal{U}\left(\hat{z} - \frac{1}{2}, \hat{z} + \frac{1}{2}\right)$, encoder $\hat{z} = \mu_\phi(x)$.
- Loss: $\mathcal{L}(\phi; \theta) = \underbrace{\mathbb{E}_{\tilde{z}\sim q}[\log p_\theta(x|\tilde{z})]}_{\text{distortion}} - \beta \underbrace{\left(0 - \mathbb{E}_{\tilde{z}\sim q}[\log p_\theta(\tilde{z})]\right)}_{\text{cross entropy: bit rate}}$,
  where β adjusts rate-distortion ratio.
- Outperforms best traditional image codec.

Proposed Baseline Model

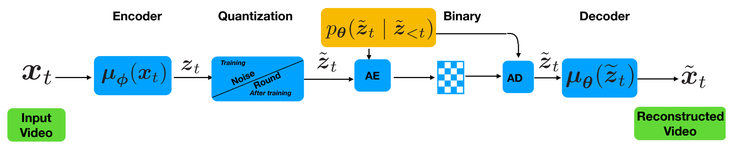


Figure: Two choices of **predictive models** for entropy coding: $p_\theta(\tilde{z}_t \mid \tilde{z}_{t-1})$ Kalman Filter; $p_\theta(\tilde{z}_t \mid \tilde{z}_{<t})$ LSTM.

- Generative model: $p(x_{1:T}, z_{1:T}) = \prod_{t=1}^{T} p_\theta(z_t|z_{<t}) p_\theta(x_t|z_t)$.
- Encode/decode frame x_t using image encoder/decoder
- Leverage predictive models to capture temporal redundancy

Improved Video Compression Model

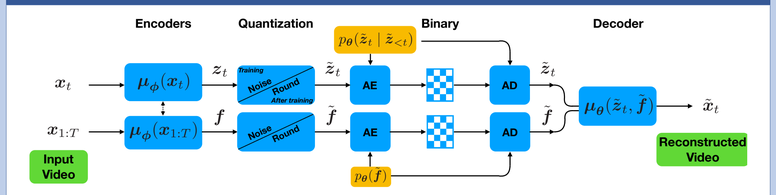


Figure: Global state f: per-segment, inferred from a segment T of video by LSTM after encoder μ_ϕ; local state z_t: per-frame, inferred from x_t after encoder μ_ϕ.

- Split latent states of a video segment into global state (coding once) and local states (predictive model for coding)
- **Loss:** distortion + cross-entropy

$$\mathbb{E}_{\tilde{f},\tilde{z}_{1:T}\sim q}[\log p_\theta(x_{1:T}|\tilde{f},\tilde{z}_{1:T})] + H[q_\phi(\tilde{f}|x_{1:T}), p_\theta(\tilde{f})] + H[q_\phi(\tilde{z}_{1:T}|x_{1:T}), p_\theta(\tilde{z}_{1:T})].$$

- **Entropy Coding:** arithmetic coding by learned prior models

Choice of Probability Models

- **Generative Model:**

$$p_\theta(x_{1:T}, z_{1:T}, f) = p_\theta(f) p_\theta(z_{1:T}) \prod_{t=1}^{T} p_\theta(x_t|f, z_{1:T})$$

- **Encoder Model:**

$$q_\phi(z_{1:T}, f \mid x_{1:T}) = q_\phi(f \mid x_{1:T}) \prod_{t=1}^{T} q_\phi(z_t \mid x_t).$$

$$\tilde{f} \sim q_\phi(f \mid x_{1:T}) = \mathcal{U}\left(\hat{f} - \frac{1}{2}, \hat{f} + \frac{1}{2}\right)$$

$$\tilde{z}_t \sim q_\phi(z_t \mid x_t) = \mathcal{U}\left(\hat{z}_t - \frac{1}{2}, \hat{z}_t + \frac{1}{2}\right).$$

where encoders $\hat{f} = \mu_\phi(x_{1:T})$ and $\hat{z}_t = \mu_\phi(x_t)$.

- **Prior Model:**

$$p_\theta(f) = \prod_{i}^{\dim(f)} p_\theta(f^i) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2});$$

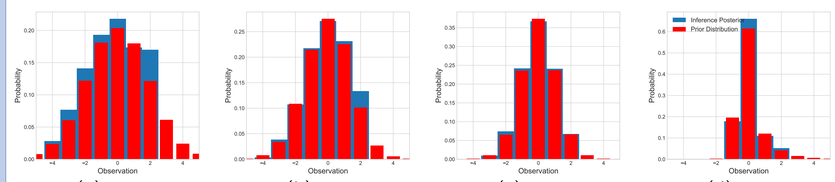
$$p_\theta(z_{1:T}) = \prod_{t}^{T} \prod_{i}^{\dim(z)} p_\theta(z_t^i \mid z_{<t}^i) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2}),$$

Gaussian convoluted with uniform; closed-form evaluation.

Verification Experiments

Datasets: Sprites; BAIR; Kinetics.

Latent Variable Distribution Visualization



(a) $p_{\theta(f^i)}$ (b) $p_{\theta(f^j)}$ (c) $p_{\theta(z_1^i)}$ (d) $p_{\theta(z_1^j)}$

Figure: Empirical distributions of **inference posteriors** and learned **prior model**.

Latent Variable Entropy Visualization

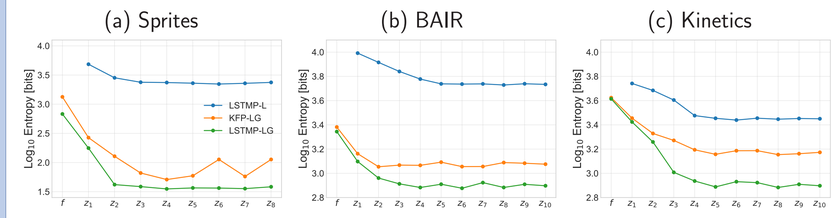
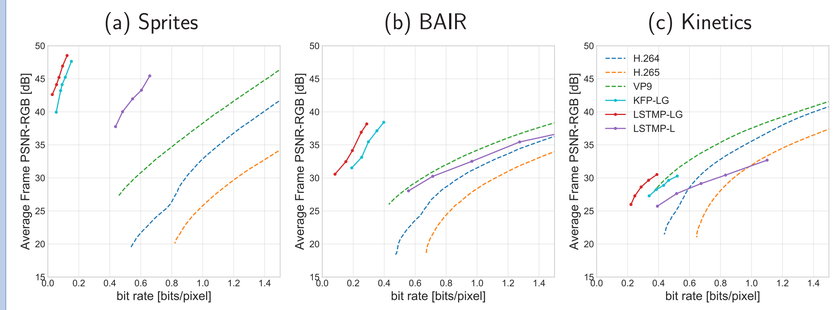


Figure: Average bits of information stored in f and $z_{1:T}$

Summary: global state stores much more information than local state; predictive model save bits; LSTM works better than Kalman filter.

Comparison Experiments

Quantitative Comparison (Rate-distortion curves)



Qualitative Comparison

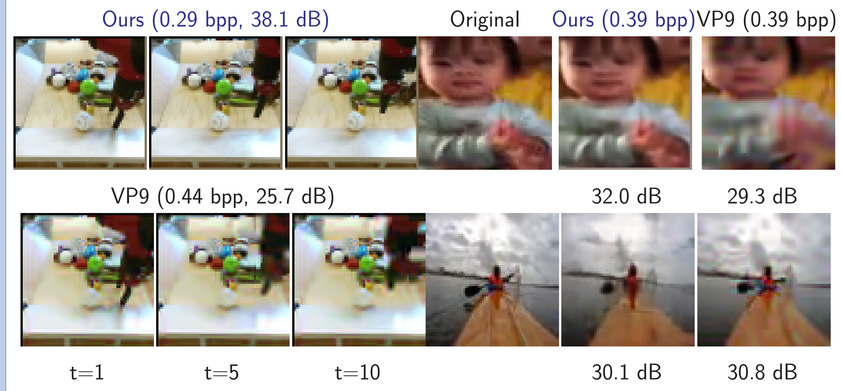


Figure: Compressed videos by our LSTMP-LG and VP9 in low-bit rate regime.