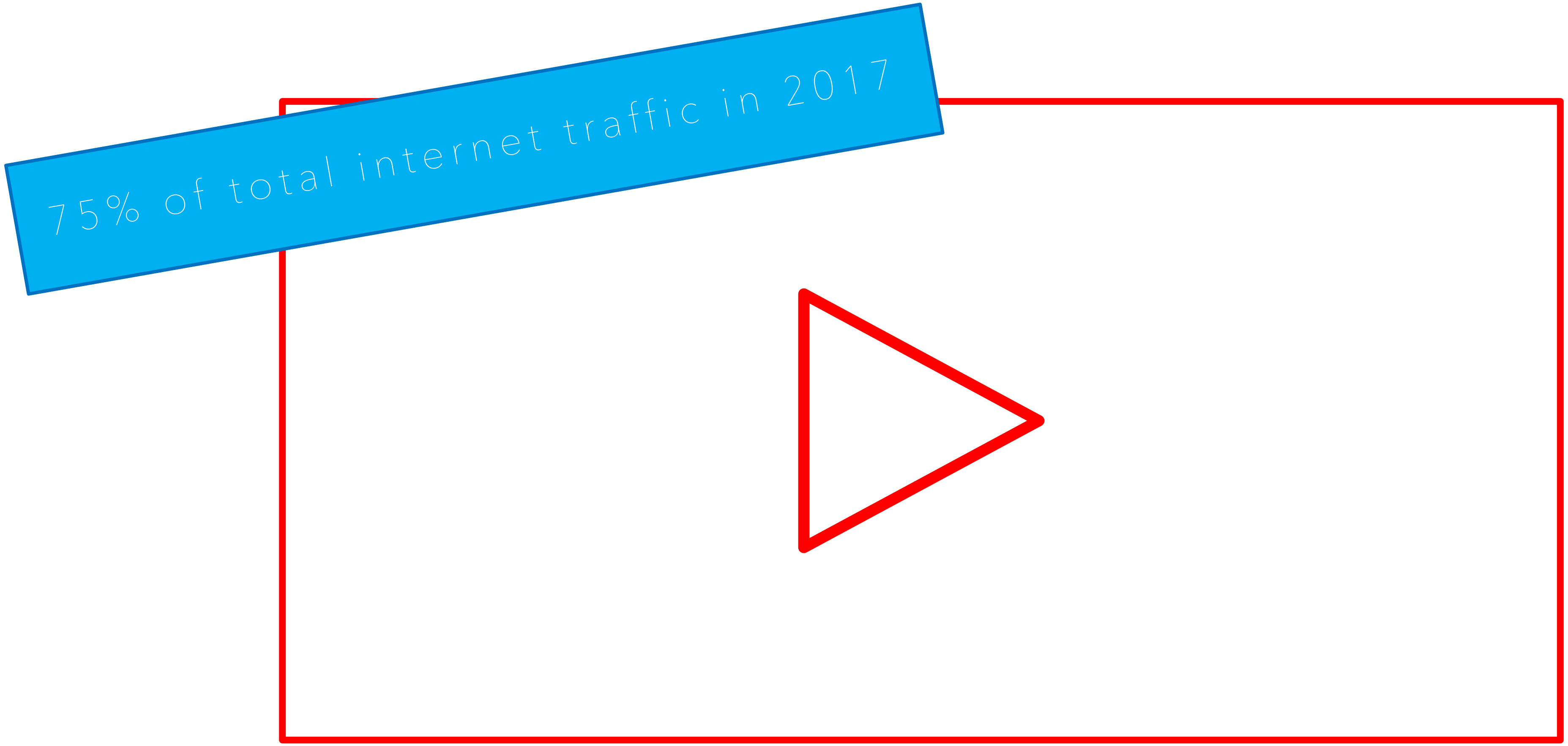
Neural Inter-Frame Compression for Video Coding

Presented by *Maxime Raafat*

Abdelaziz Djelouah, Joaquim Campos, Simone Schaub-Meyer, Christopher Schroers (Disney Research)

ICCV, October 27, 2019


75% of total internet traffic in 2017

Previous work

- Image compression : JPEG, JPEG2000, BPG, WebP
- Neural Image Compression : deep learning for image compression
- Video compression : H.264 (AVC), H.265 (HEVC)
- Neural Video Compression : deep learning for video compression

Pipeline

Interpolation based video compression technique, compatible with neural image compression methods, while expressing residual information in latent space.

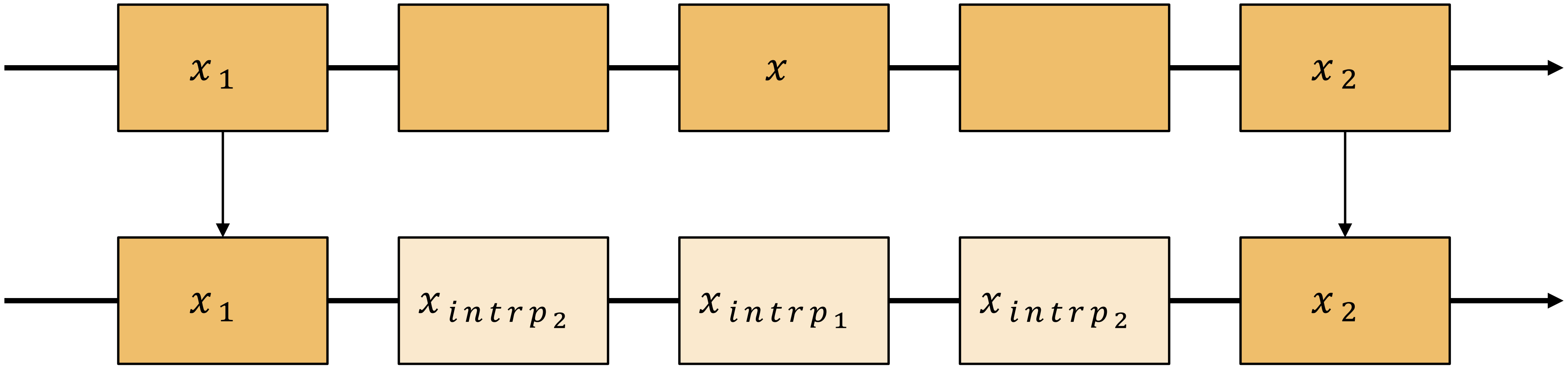
- Interpolation with compression constraints
- Experimental Results

Interpolation with compression constraints

Interpolation model

Reference frames (keyframes) $\mathcal{K}_x = \{x_1, x_2, \dots, x_k\}$

Predict intermediate frame $\boldsymbol{x}$



Interpolation model

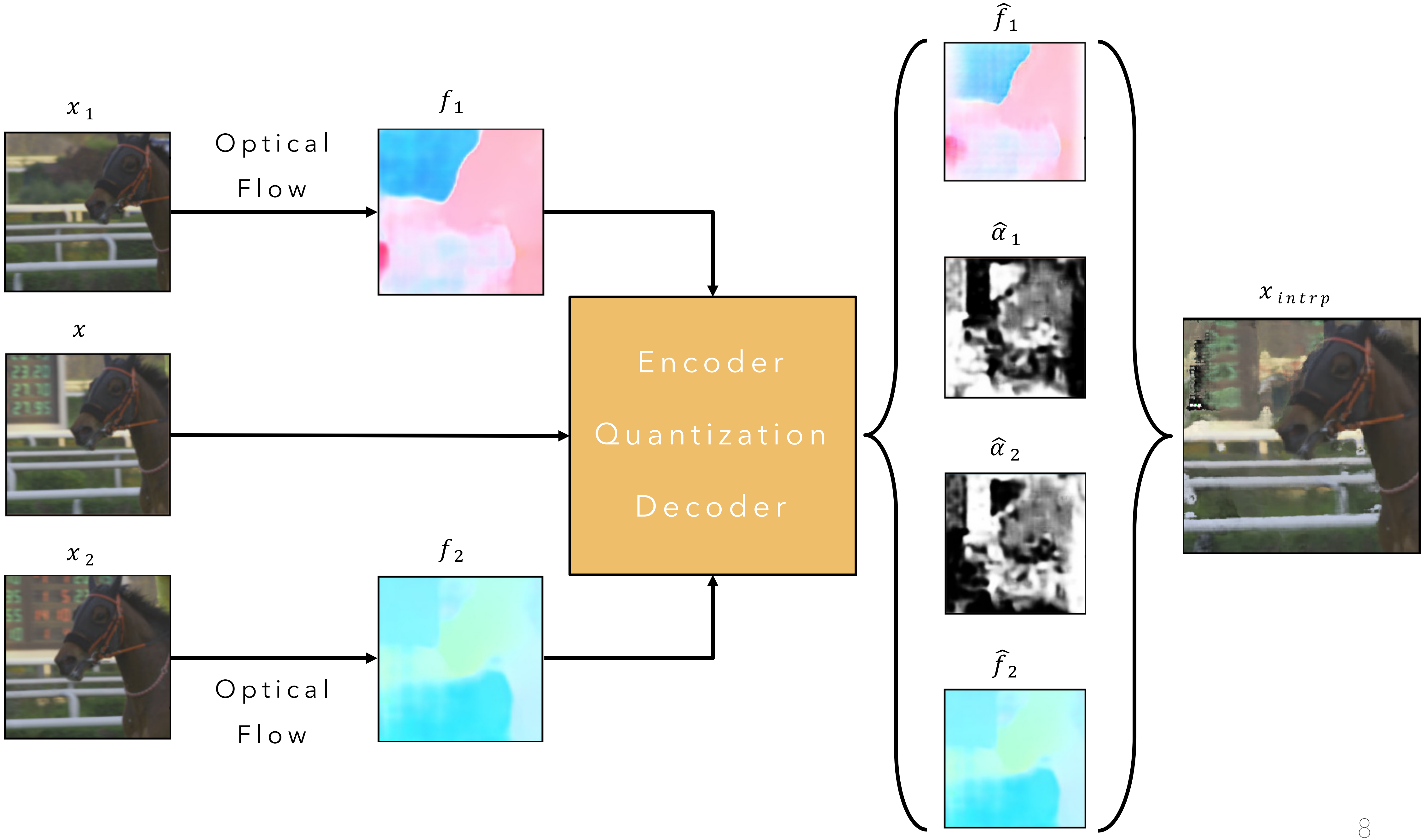
Reference frames (keyframes) $\mathcal{K}_x = \{x_1, x_2, \dots, x_k\}$

Predict intermediate frame $\boldsymbol{x}$

$$\boxed{x_{intrp} = \sum_{i=1}^{k} \hat{\alpha}_i \, \omega(x_i, \hat{f}_i), \text{ with } \sum_{i=1}^{k} \hat{\alpha}_i = 1}$$

$\hat{f}_i$ = quantized displacement map of $\boldsymbol{x}_i$ w.r.t. to $\boldsymbol{x}$

$\hat{\alpha}_i$ = quantized blending coefficient of $\boldsymbol{x}_i$


x_1
Optical
Flow
f_1
x
Encoder
Quantization
Decoder
x_2
Optical
Flow
f_2
$\hat{f}_1$
$\hat{\alpha}_1$
$\hat{\alpha}_2$
$\hat{f}_2$
x_{intrp}

Compression constraints

Quantized latent representation $\hat{q}$ should occupy as little storage as possible, while minimizing distortion on the interpolation result.

$$\boxed{L(\phi, \phi', p_{\hat{q}}) = \mathbb{E}_{x \sim p_x}[-log_2 p_{\hat{q}}(\hat{q}) + \lambda * d(x, x_{intrp})]}$$

ϕ and ϕ' = encoder-decoder network parameters

$p_{\hat{q}}$ = entropy model ; λ = compression-rate vs distortion regularizer

Latent space residuals

Minimize the transmitted residual information between x_{intrp} and x.

$$r = y - y_{intrp} = g_\phi(x) - g_\phi(x_{intrp})$$

Quantize $r \rightarrow \hat{r}$

$$\hat{x} = g_{\phi'}\left(y_{intrp} + \hat{r}\right)$$

g_ϕ and $g_{\phi'}$ = encoder and decoder

Latent space residuals

Minimize the transmitted residual information between $\boldsymbol{x_{intrp}}$ and $\boldsymbol{x}$.

$$\boxed{L\left(\phi, \phi', p_{\hat{q}}\right) \leftarrow L\left(\phi, \phi', p_{\hat{q}}\right) + \mathbb{E}_{x \sim p_x}[-log_2 p_{\hat{r}}(\hat{r}) + \lambda * d(x, \hat{x})]}$$

$\boldsymbol{\phi}$ and $\boldsymbol{\phi'}$ = encoder-decoder network parameters

$\boldsymbol{p_{\hat{r}}}$ = entropy model for residual values

$\boldsymbol{\lambda}$ = regularizer

Network architectures

- Encoder g_{ϕ} : 5 blocks (each one convolutional* and one Generalized Normalization Transformation layer).

- Decoder $g_{\phi'}$: 3 RGB output channels for the image $\hat{x}$ and 5 output channels for x_{intrp} ($\hat{f}_1$ and $\hat{f}_2$, as well as $\hat{\alpha}_1$ and $\hat{\alpha}_2 = 1 - \hat{\alpha}_1$).

- Training performed with the mean squared error (MSE) for the distortion function d.

* kernel size $k = 5$, stride size $s = 2$

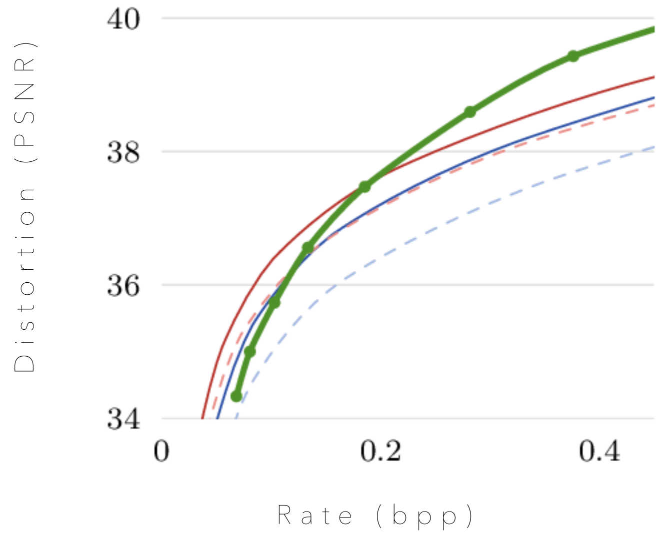
Experimental Results

Experimental Setting

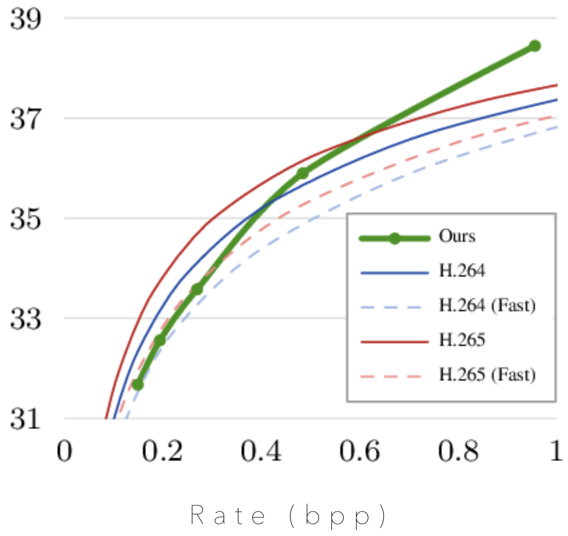
- Keyframes positioned every 12 frames
- Peak Signal to Noise Ratio (for distortion measures)
- Training on 3 datasets (max length = 300 frames)
  - MCL-JVC (resolution : 1920 x 1080)
  - VTL : Video Trace Library (resolution : 352 × 288)
  - UVG : Ultra Video Group (resolution: 1920 x 1080)

Video codec comparisons

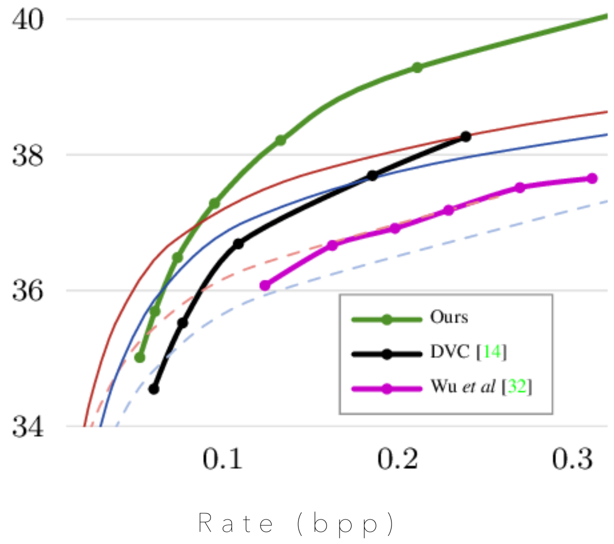
MCL-JVC Dataset



VTL Dataset



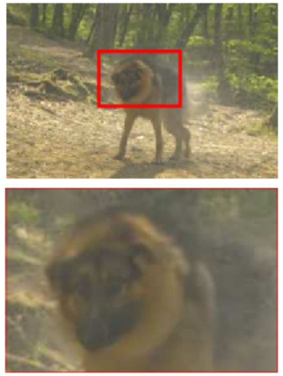
UVG Dataset



Advantages of the proposed interpolation

Flow + Interpolation

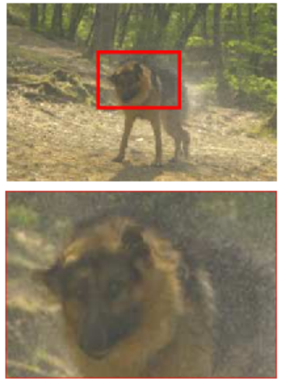
0.028bpp



0.024bpp

Ours (lower bit-rate)

0.027bpp

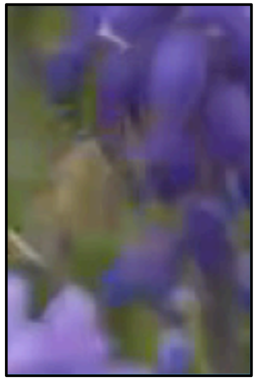


0.021bpp

Ours (higher bit-rate)

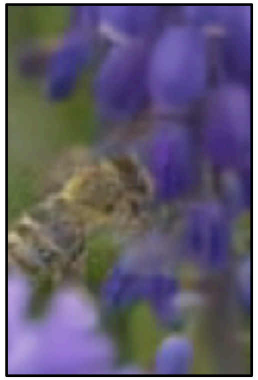
0.24bpp

0.27bpp

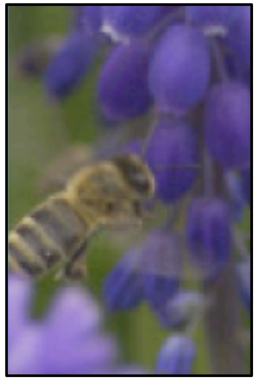


H.264
0.02bpp

H.265
0.02bpp



Ours
0.02bpp



Ground
Truth

Questions?