Face video compression with generative networks

Yan Ye

Video Technology Lab, DAMO academy, Alibaba Group US

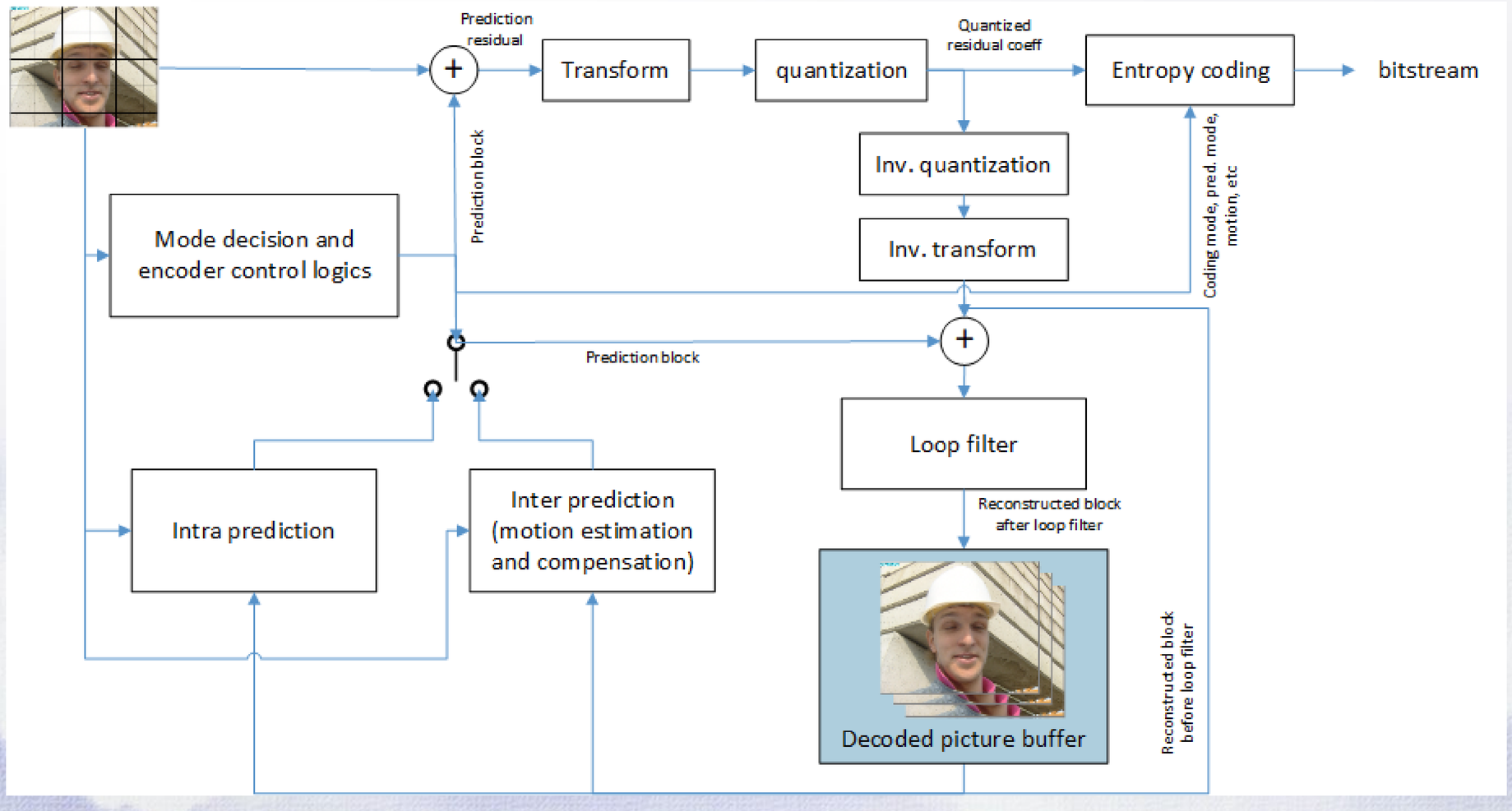
Outline

CONTENTS

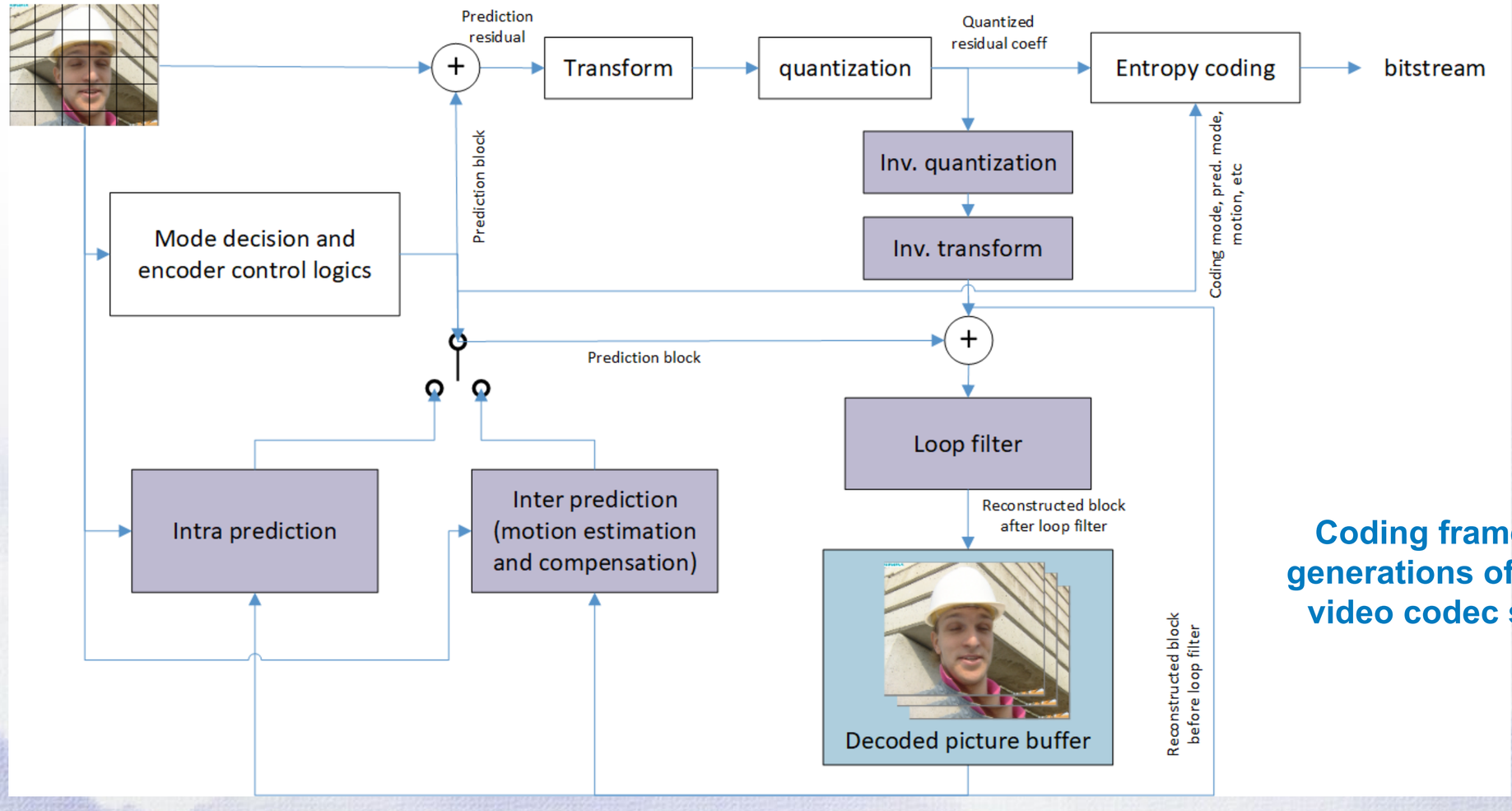
01

- Hybrid video coding framework
- Generations of video coding standards

Block-based hybrid video coding

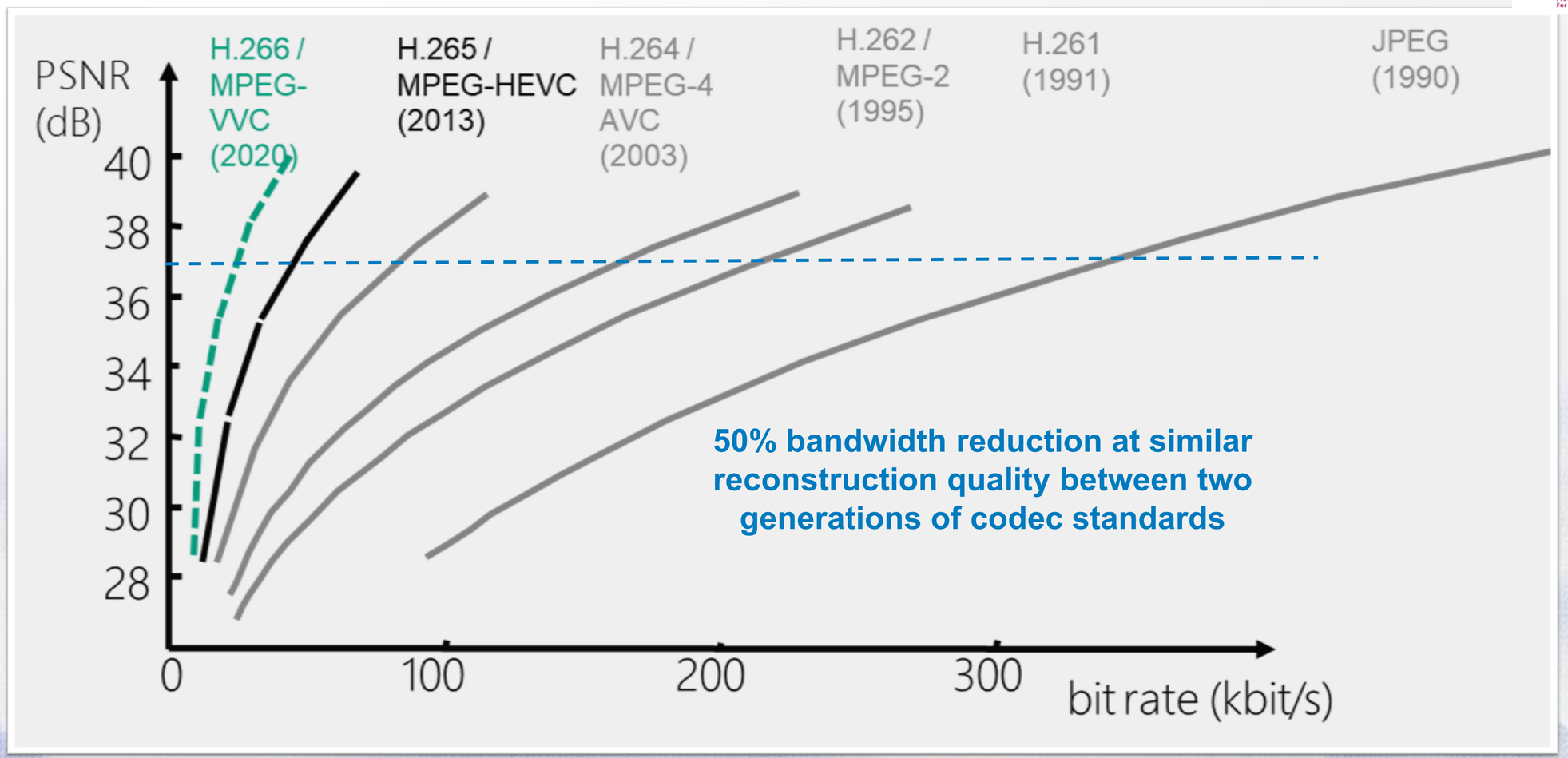


Block-based hybrid video coding



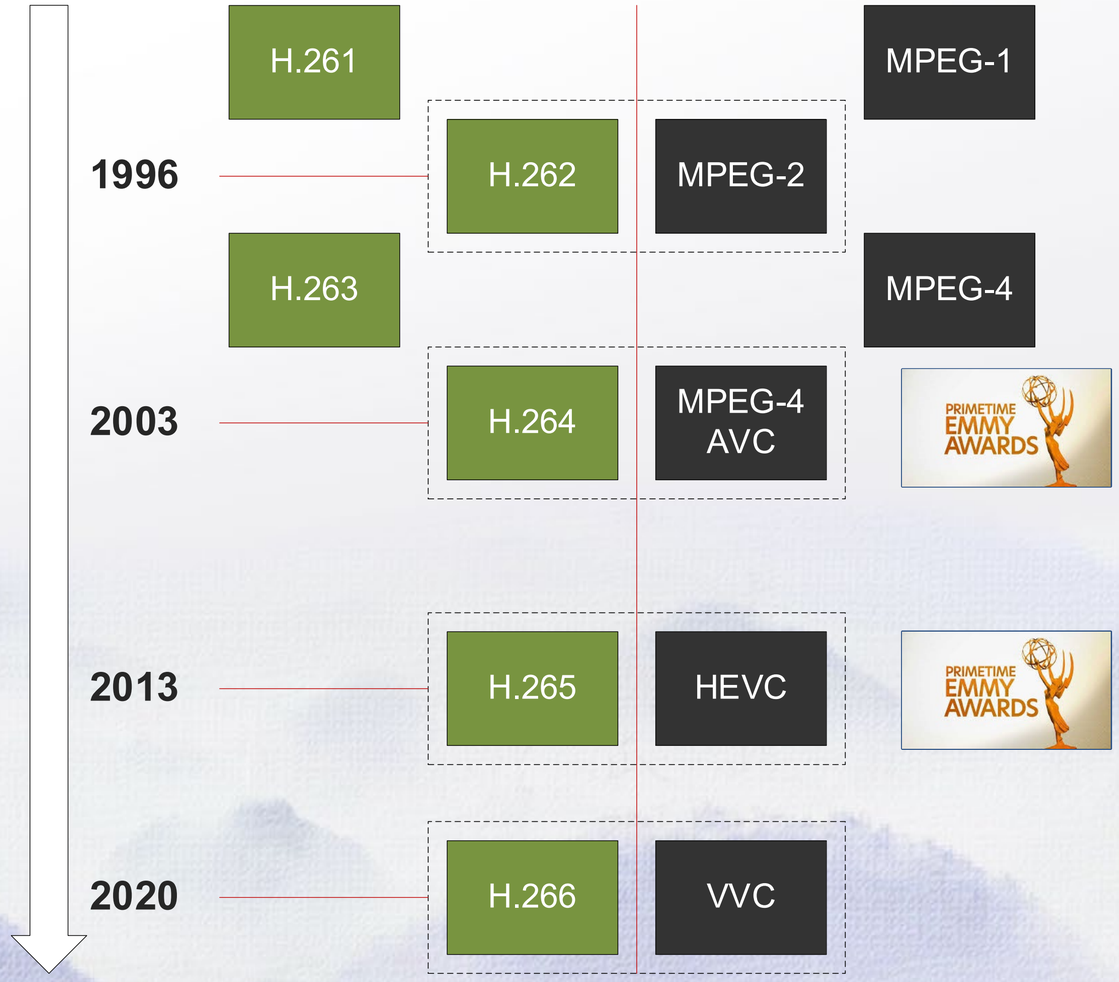
Coding framework for generations of image and video codec standards

Evolution of compression efficiency



Slide courtesy of B. Bross, “Versatile video coding (VVC) on the final stretch”, ITU Workshop on “The future of media,” Geneva, Switzerland, 8 October 2019

Deployment of video coding standards



ITU-T SG16 VCEG **ISO/IEC SC29 MPEG**

MPEG-2
- Digital TV broadcasting
- Cable TV
- DVD

H.264/AVC
- Digital TV, satellite TV
- Internet video, mobile video
- Video conferencing
- Blue-ray Disc

H.265/HEVC
- Streaming
- Mobile video
- HD and UHD, HDR

H.266/VVC
- AR/VR, 360-degree video
- UHD 4K & 8K

02

- Learning-based image coding
- Learning-based video coding

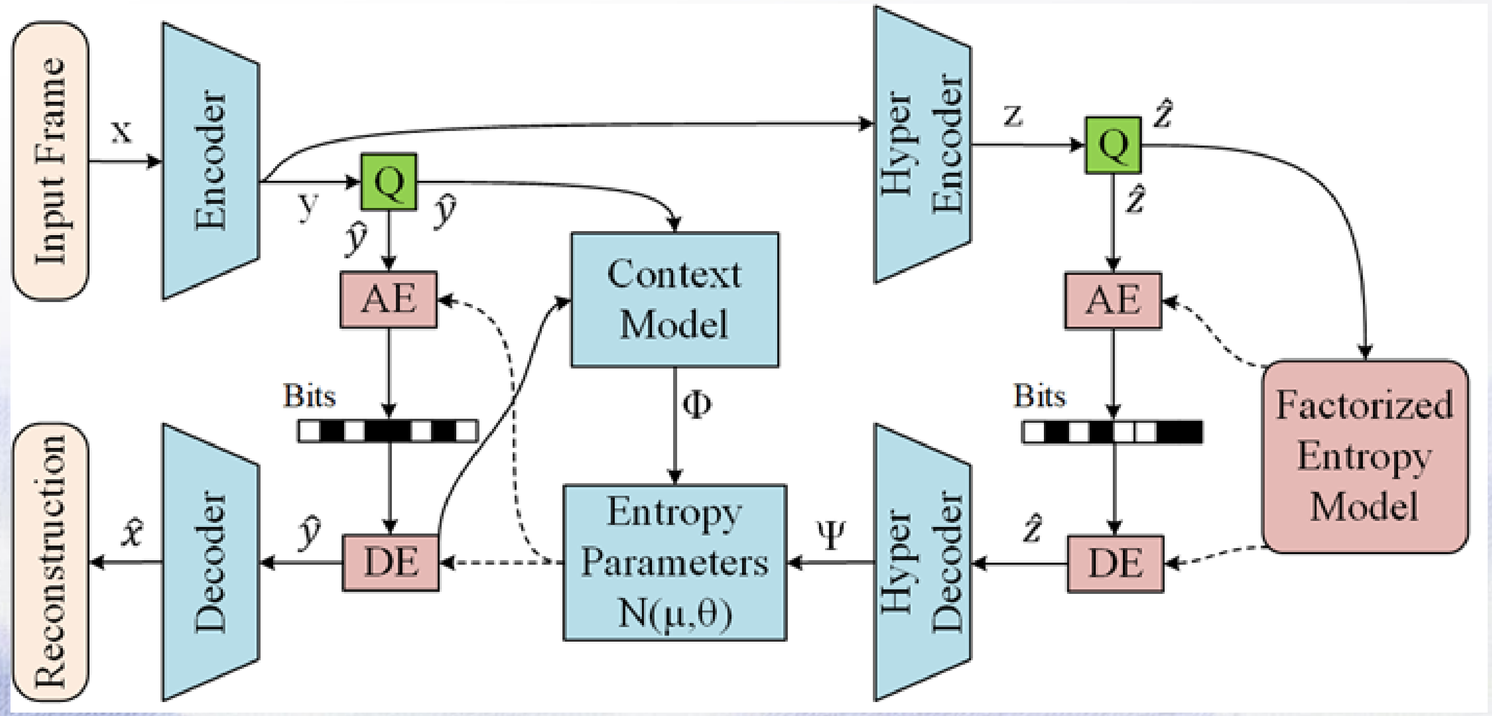
Competitive performance

Learning-based image and video coding

- Learning-based image and video coding has been a popular research topic in recent years
- Enhancing/replacing a coding tool within the hybrid framework
  - Intra coding, inter coding, loop filtering, entropy coding, etc.
- **End-to-end image and video compression**
  - Compression network is end-to-end trainable
  - Network types: CNN, RNN, transformer, generative
  - Distortions (in loss function): mean squared errors, structural similarity, perceptual metrics

End-to-end learning-based image coding

End-to-end trainable



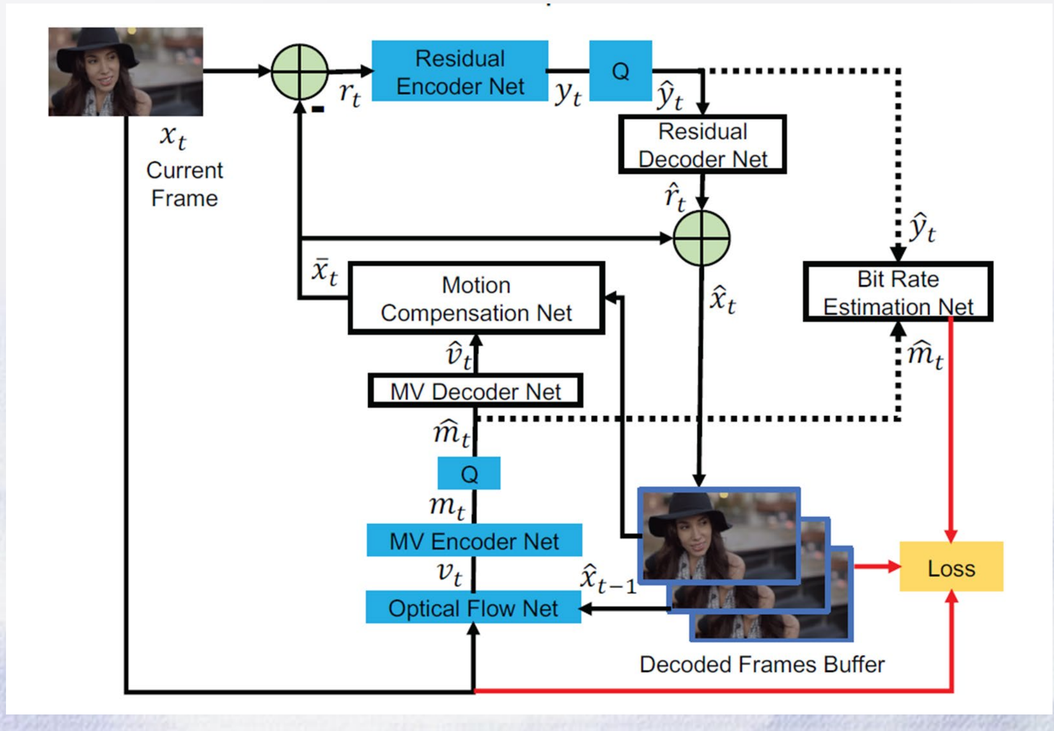
CNN network + entropy model

entropy model
- Factorized
- Hyperprior
- Autoregressive
- Coarse-to-fine
- …

Ballé, Johannes, et al. "End-to-end optimized image compression." in ICLR. 2017.
Ballé, Johannes, et al. "Variational image compression with a scale hyperprior." in ICLR. 2018.
Minnen, David, et al. "Joint autoregressive and hierarchical priors for learned image compression." in NeruIPS. 2018.
Mentzer, Fabian, et al. “Conditional Probability Models for Deep Image Compression”, in CVPR, 2018.

Deep Video Coding (DVC)

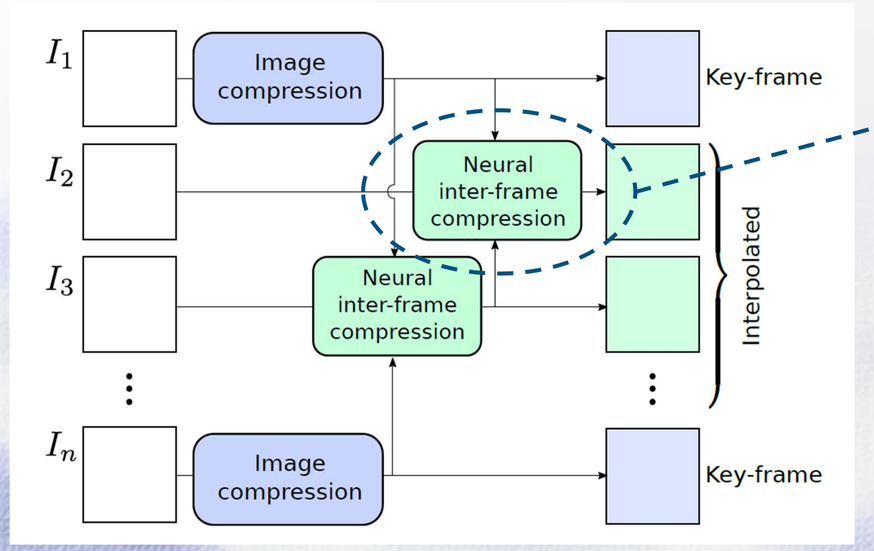
- Uses traditional block diagram but replaces main modules with deep networks
  - residual coding, motion network, advanced entropy model
- Entire system trained in an end-to-end manner
- First end-to-end video compression framework

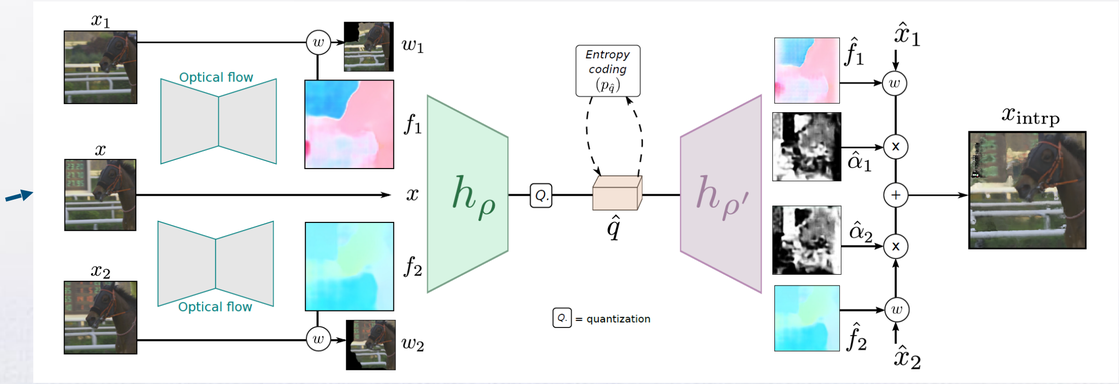


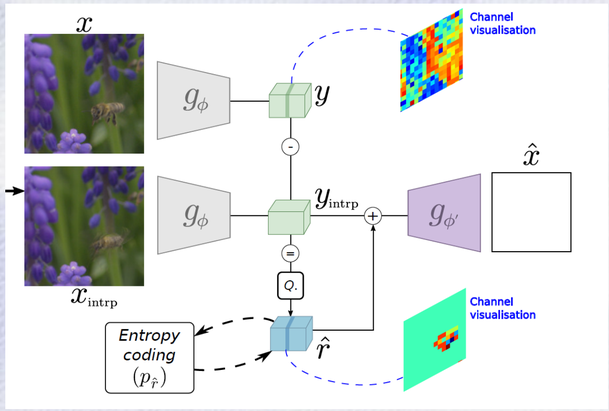
G. Lu, et. al., “DVC: an end-to-end deep video compression framework,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11006–11015, 2019.

Neural Inter Frame Compression

- Bi-directional “prediction” using frame interpolation
- Optical flow and interpolation coefficients decoded simultaneously
- Residual is coded in the latent space

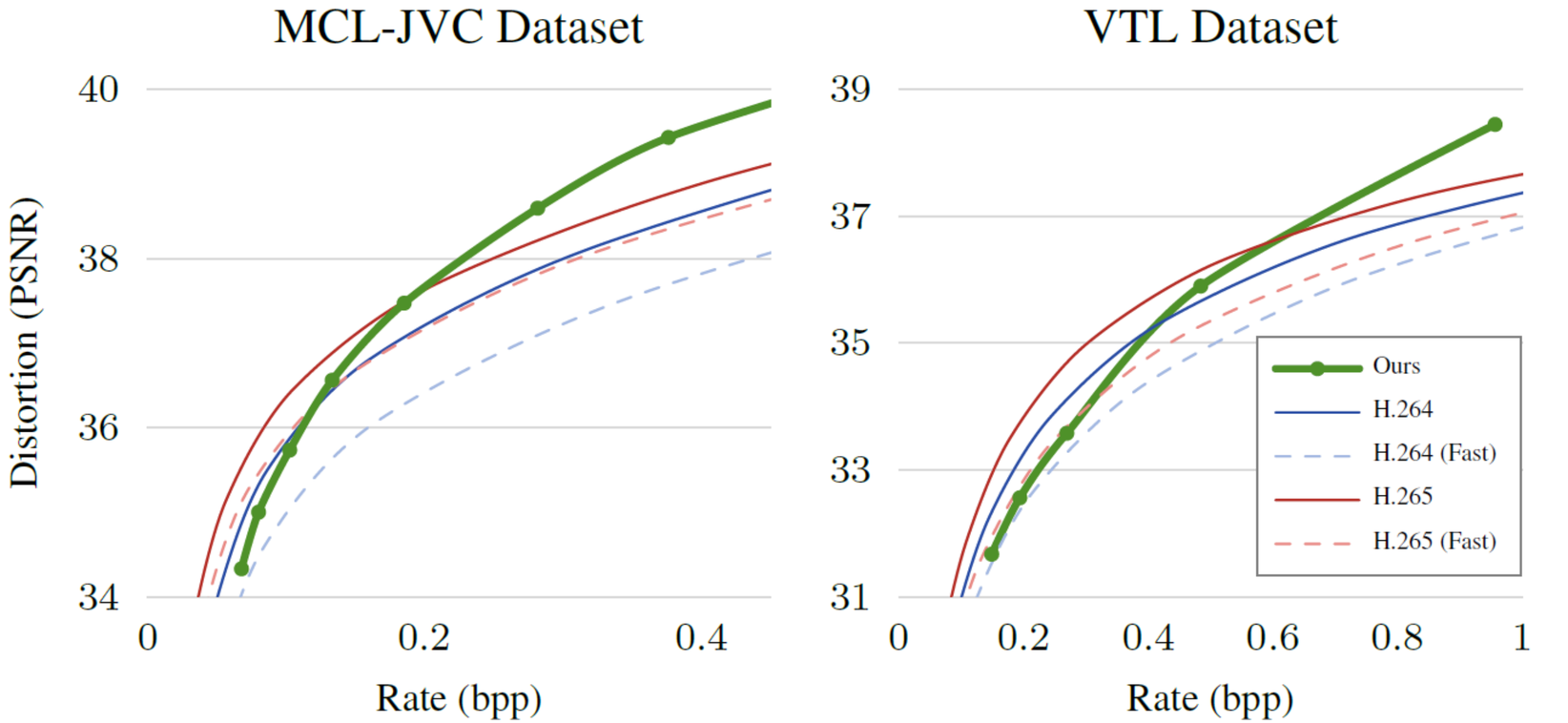






A. Djelouah, et al. "Neural Inter-Frame Compression for Video Coding." in ICCV. 2019.

Neural Inter Frame Compression: coding performance



Competitive performance compared with x265

A. Djelouah, et al. "Neural Inter-Frame Compression for Video Coding." in ICCV. 2019.

03

- Related work
- Our work
  - Compact temporal motion feature
  - 3D facial semantics
- Coding performance (vs. H.266/VVC)

Superior performance @ ultra low rate

Talking face video compression

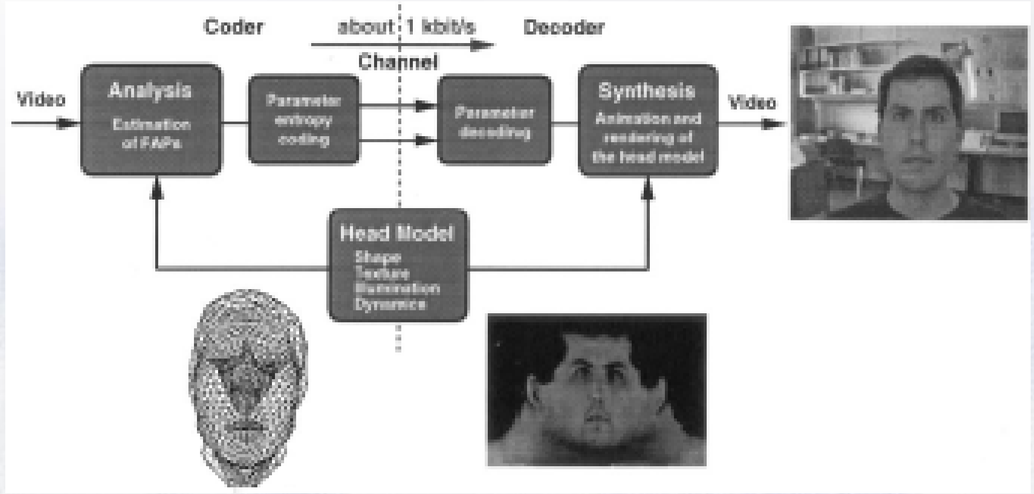
We focus on coding of human face video, where we find much inherent structure and prior knowledge, such as their shape, composition, and movement

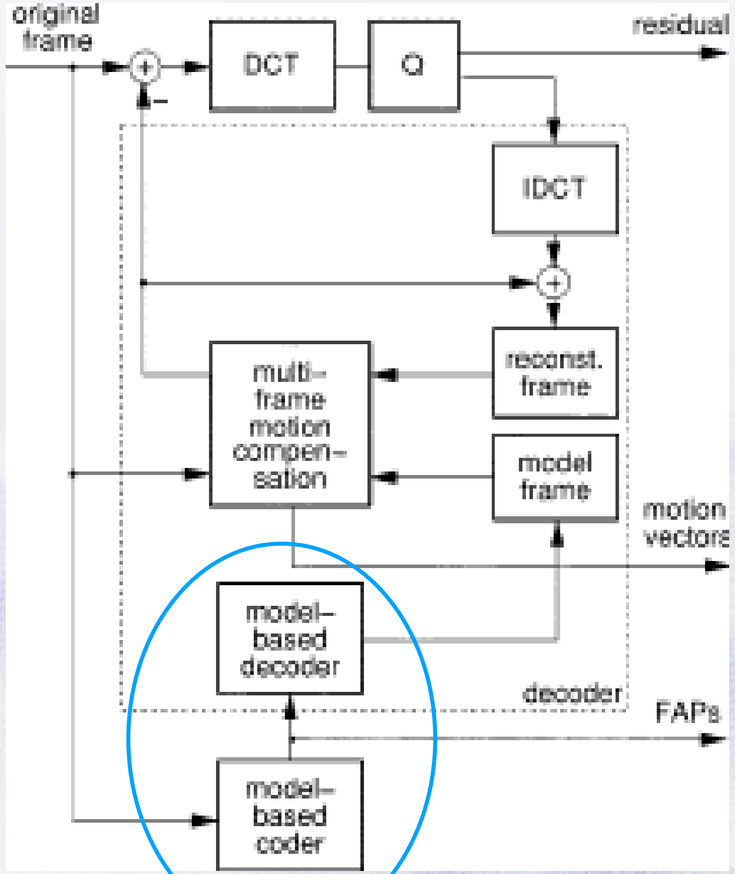
Related work

Model-based video compression

In the 1990's, model-based video compression was studied for video telephony

- Parameterized 3-D head model specifies shape and color of a person
- Facial animation parameters (FAP's) specifies motion and deformation in the temporal domain

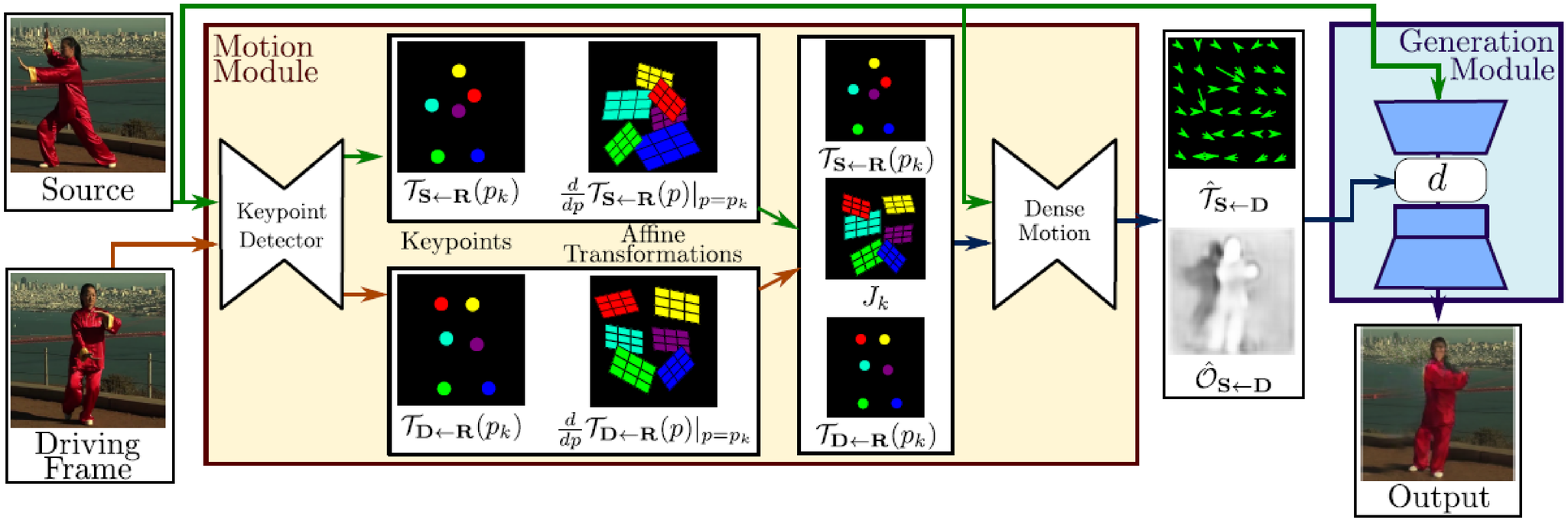




Today, with deep generative networks, reconstruction quality can be significantly improved

P. Eisert, T. Wiegand, and B. Girod, "Model-aided coding: a new approach to incorporate facial animation into motion-compensated video coding," IEEE Transactions on CSVT, vol. 10, no. 3, pp. 344–358, 2000

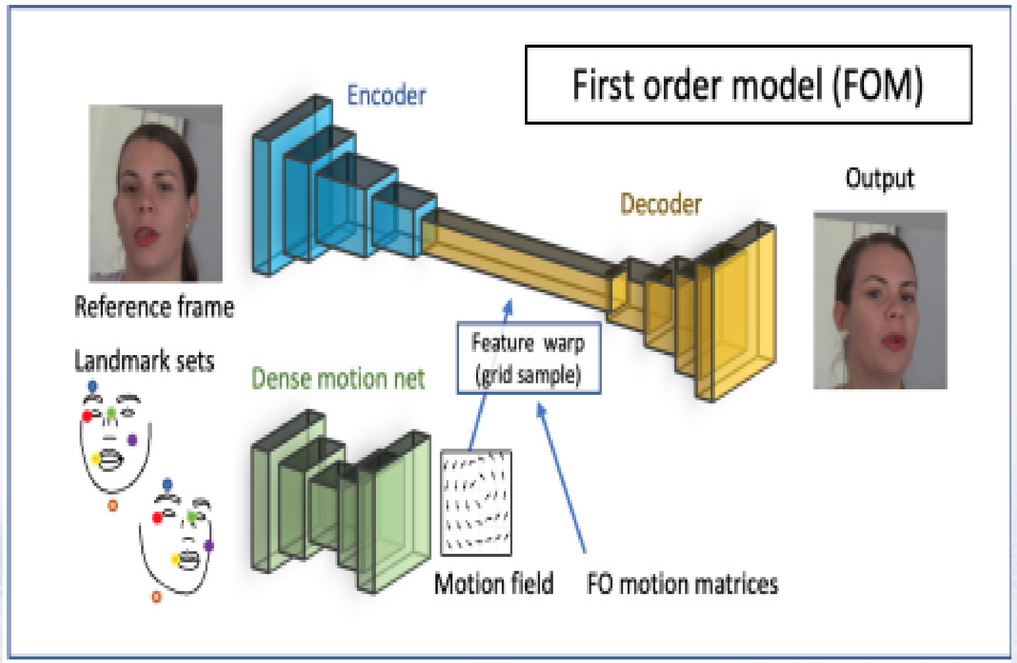
First order motion model (FOMM)

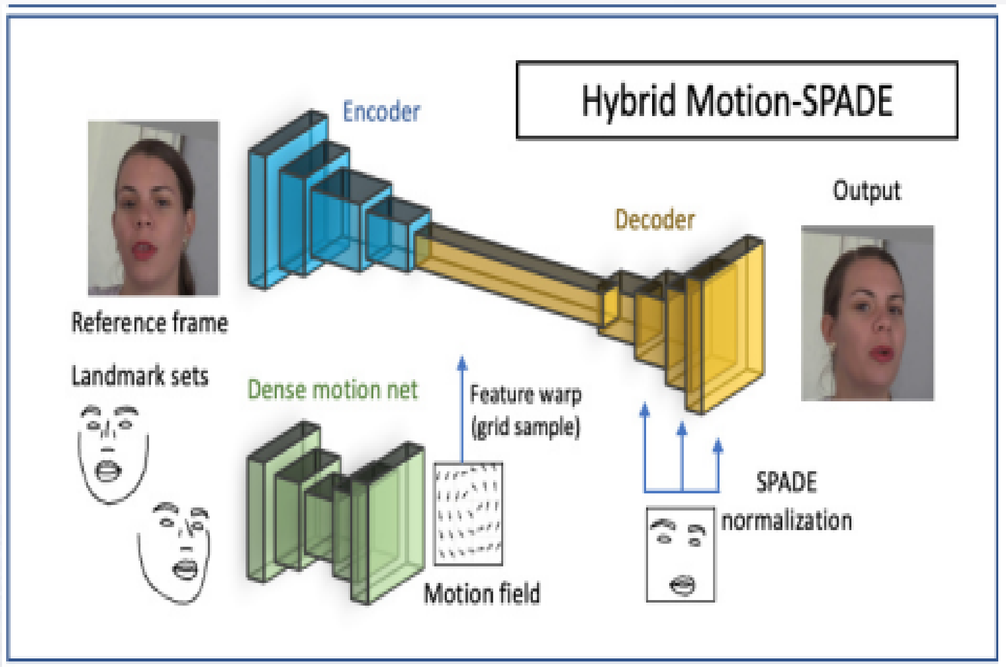


- Complex motions are represented using a set of keypoints & corresponding affine transformations
- Generator network combines the source image and the motion derived from the driving video
- Object in the source image is animated according to the motion of driving video

A. Siarohin, et. al., “First order motion model for image animation,” Advances in Neural Information Processing Systems, vol. 32, pp. 7137–7147, 2019.

Low bandwidth video-chat compression

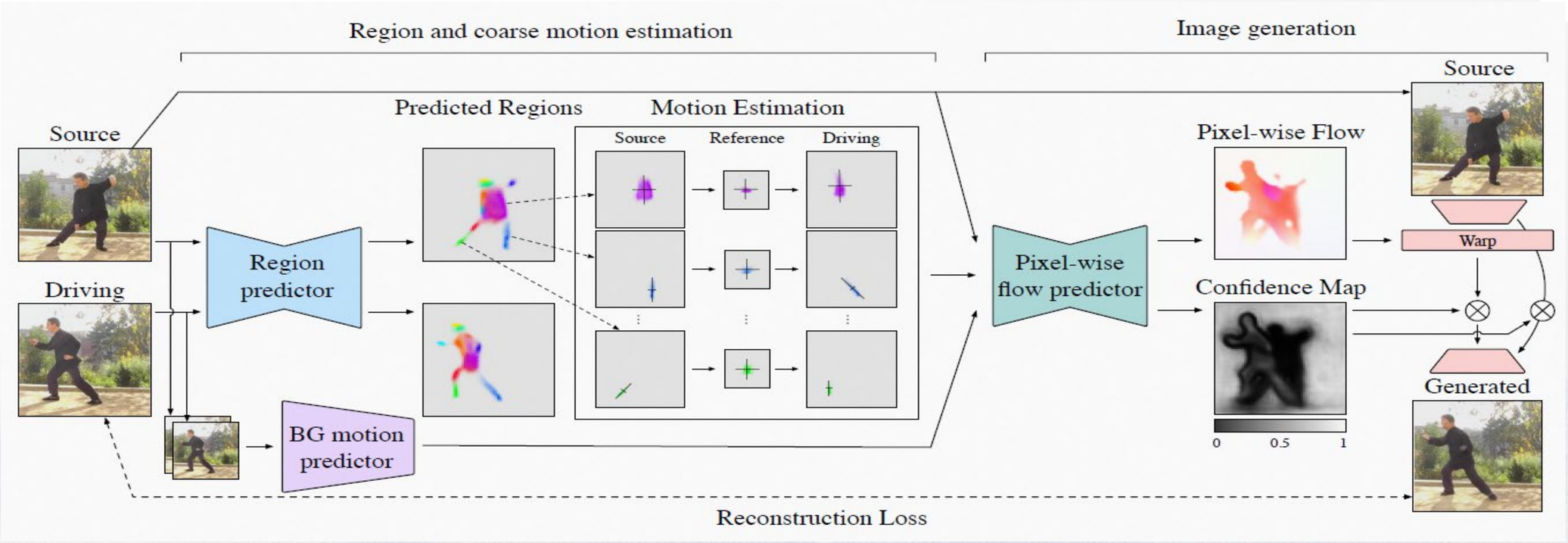




- Apply FOMM towards talking-head video compression
- Explore quality and bandwidth trade-offs for static landmarks (i.e., keypoints), dynamic landmarks or segmentation maps
- Runs real-time on mobile platform

M. Oquab, et. al., “Low bandwidth video-chat compression using deep generative models,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop, 2021.

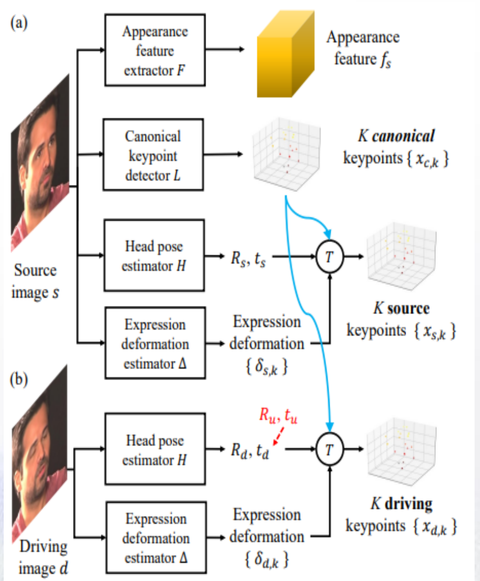
Motion Representations for Articulated Animation (MRAA)

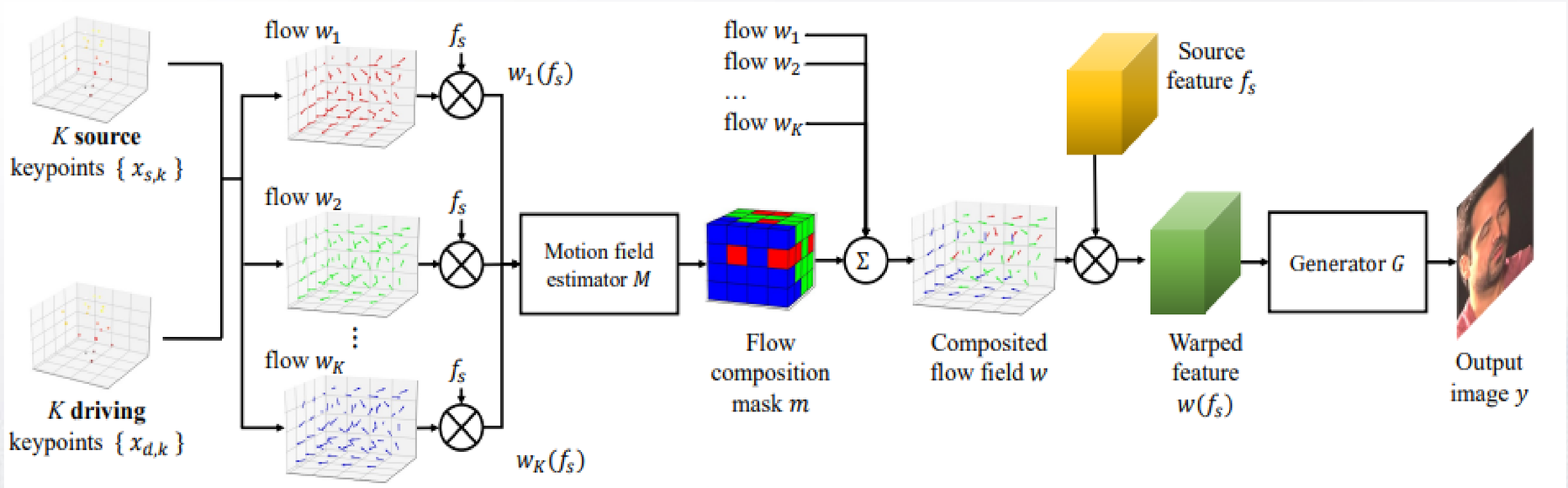


- Model consistent regions (e.g. objects) with first-order motion (i.e., locations, shape, and pose)
- Model non-object related global motion with an additional affine transformation to decouple foreground from background

Siarohin A, et al., “Motion representations for articulated animation,” IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13653-13662.

Free-view neural talking-head synthesis





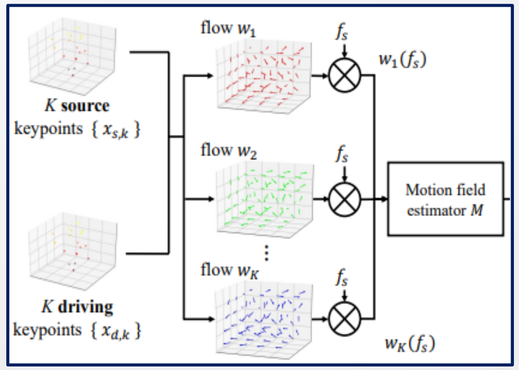
- Motion information represented using compact 3D keypoints
- Source image containing the target person's appearance and driving video dictates the motion in the output
- 3D keypoints allows to rotate the head during synthesis

T.C. Wang, et. al., "One-shot free-view neural talking-head synthesis for video conferencing," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10039–10049.

Part 1:

Coding compact temporal motion features

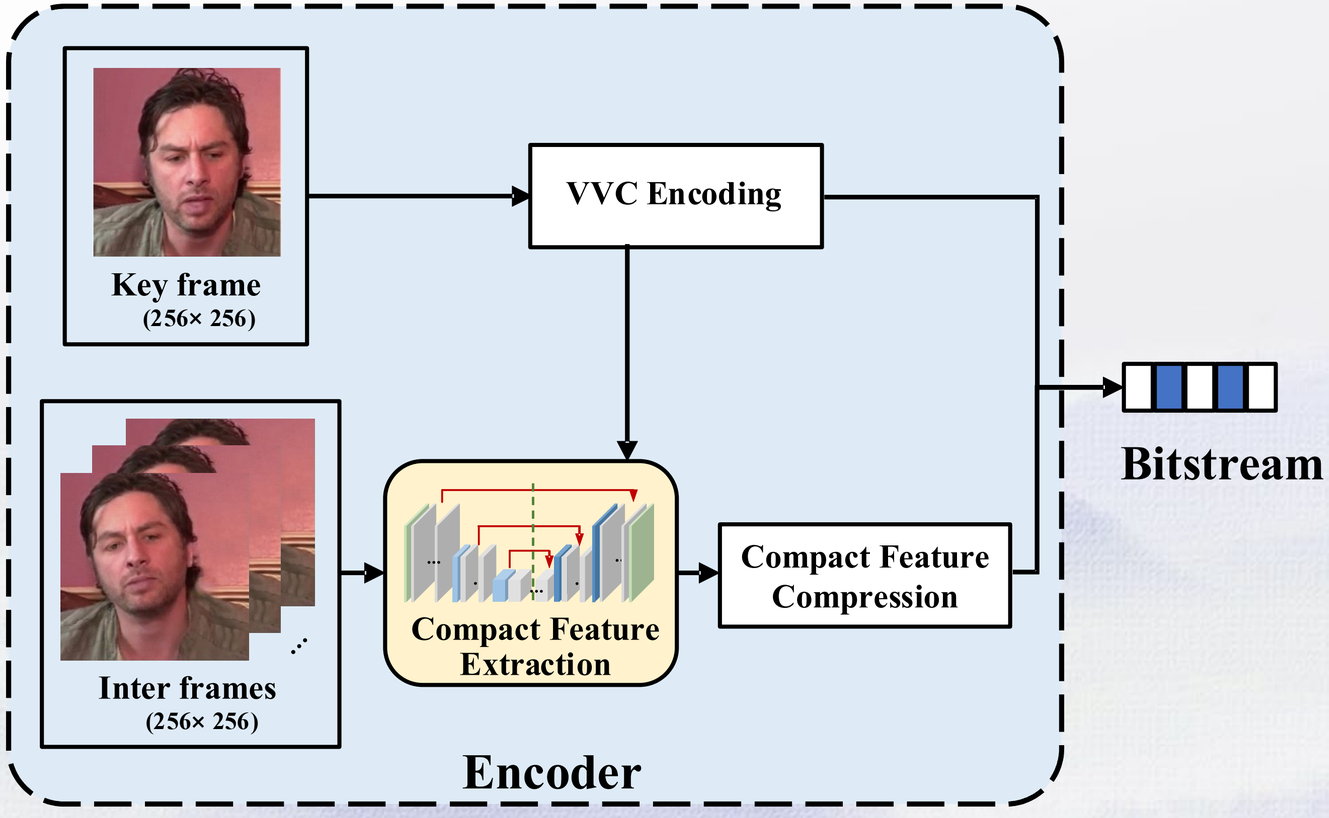
Going beyond keypoints



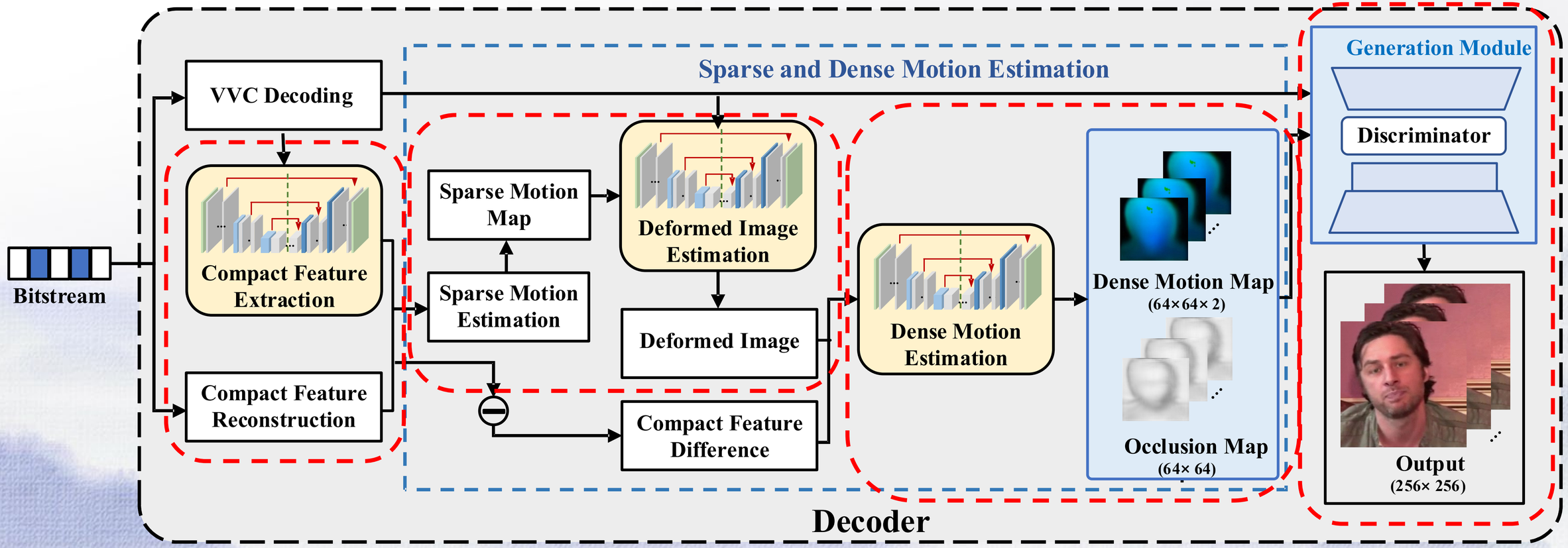
Loosely correlated with facial features

Separately drives motion flow

CTMF encoder



CTMF decoder



CTMF Training

Perceptual loss

$$L_{per-initial} = \sum_{n=1}^{i} \frac{1}{C_i \times H_i \times W_i} \left\| VGG_i(F_{cdf}) - VGG_i(\phi(I)) \right\|$$

$$L_{per-final} = \sum_{n=1}^{i} \frac{1}{C_i \times H_i \times W_i} \left\| VGG_i(\hat{I}) - VGG_i(I) \right\|$$

Adversarial loss

$$L_G(\hat{I}) = -\sum_{i=1}^{k} E_{\hat{I} \sim P_g}(D_i(\hat{I}))$$

$$L_D(\hat{I}, I) = \sum_{i=1}^{k} E_{\hat{I} \sim P_g}\left(D_i(\hat{I})\right) - \sum_{i=1}^{k} E_{\hat{I} \sim P_r}\left(D_i(I)\right)$$

Total loss

$$L_{total} = \lambda_{initial} \cdot L_{per-initial} + \lambda_{final} \cdot L_{per-final} + \lambda_{adv} \cdot (L_G + L_D)$$

CTMF decoding flow visualization

Key frame | Current frame | Feature map | Coarse deformed frame | Dense motion map | Occlusion map | Final output

Part 2:

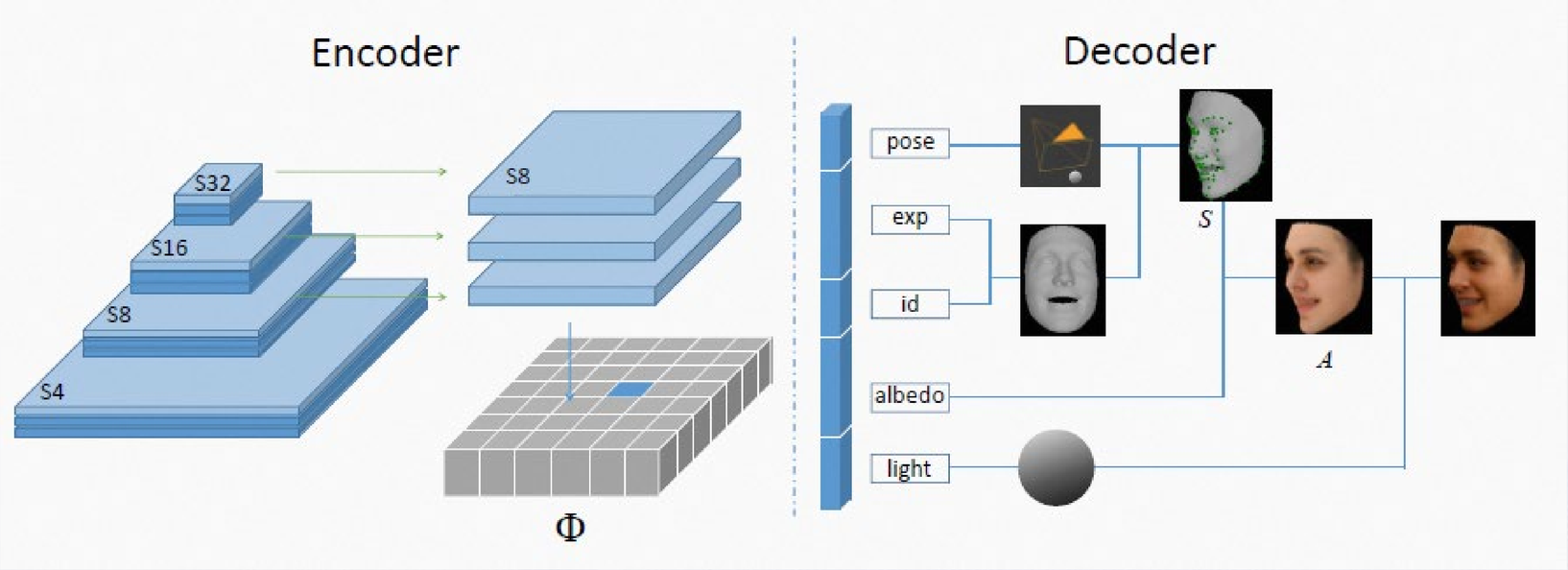
Coding 3D facial semantics

Why 3D facial semantics

- **Compact temporal motion features:**
  - Reduces parameter signaling overhead compared to keypoint-based methods, but
  - Mainly represents global motion, and can't be used to control head posture
- **Neural talking head: 3D key points can be used to rotate the head, but not to alter expression**

T.C. Wang, et. al., "One-shot free-view neural talking-head synthesis for video conferencing," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10039–10049.
Video from https://nvlabs.github.io/face-vid2vid/

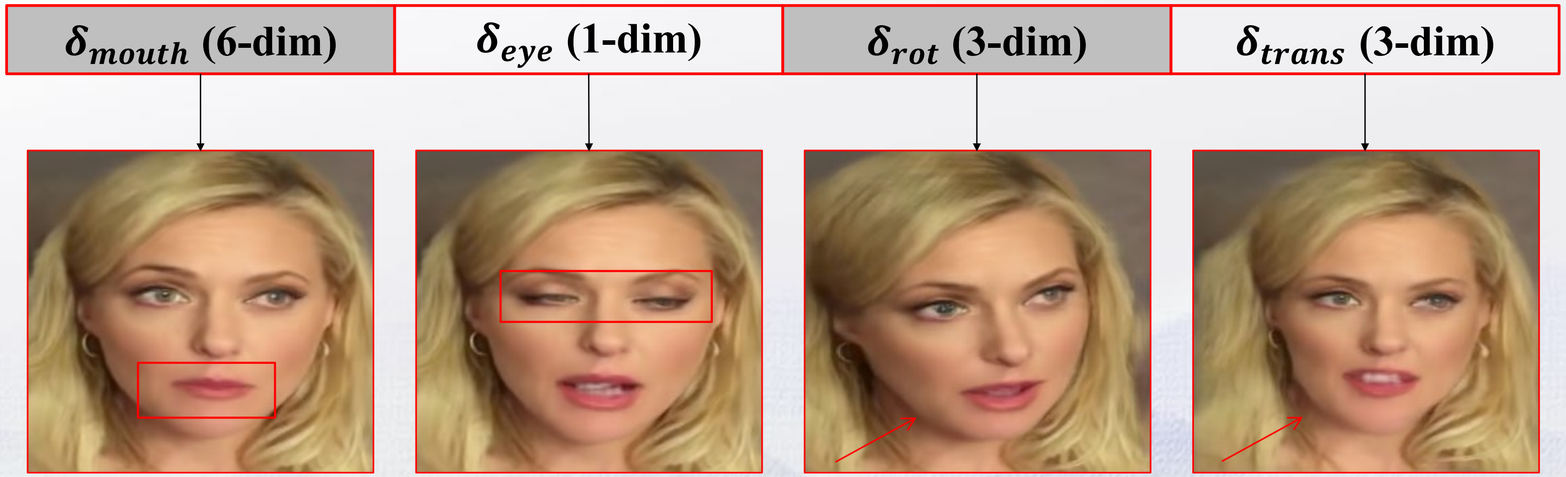
Expressing 3D face



- Utilize a parametric face model to separate facial parameters
  - 3D head position, head rotation, face expression, eye gaze, and eye blinking
- Can be transferred from a source actor to a target actor for controllable face reconstruction

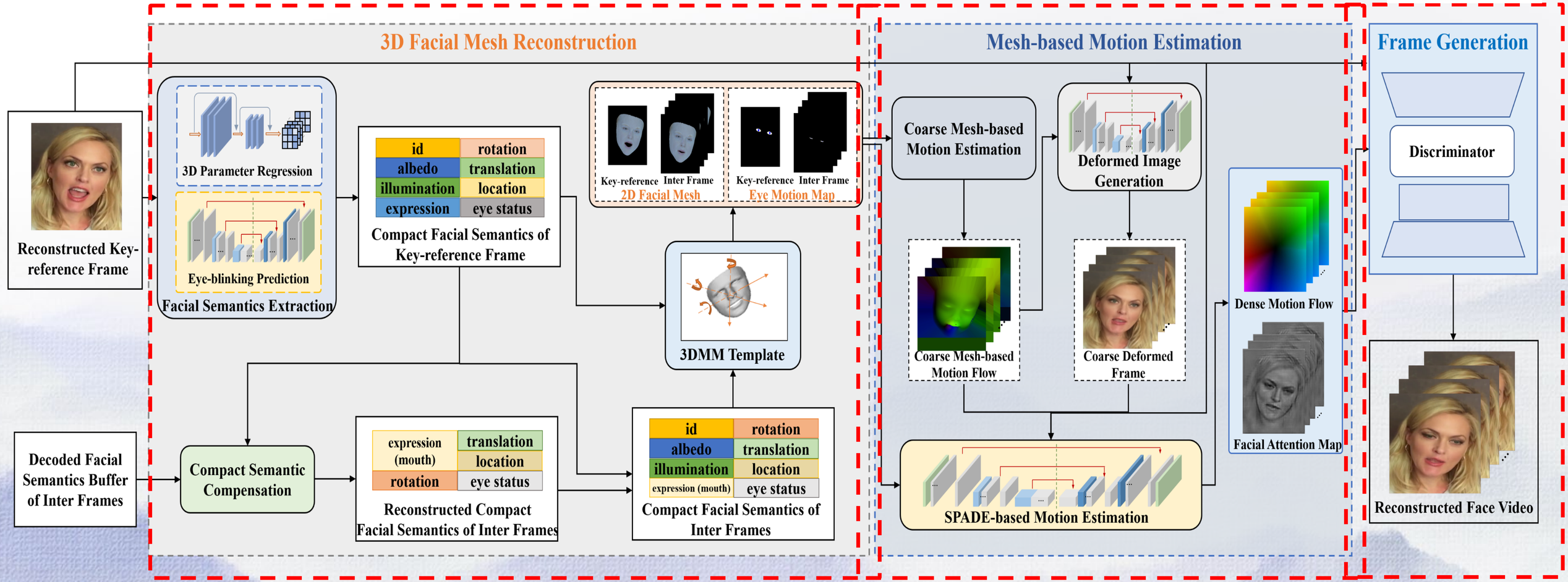
Zhang J, Lin L, Zhu J, et al. Weakly-supervised multi-face 3d reconstruction[J]. arXiv preprint arXiv:2101.02000, 2021.

3D face semantics (3DFS)



13-dimensional semantics to represent face

3DFS decoder



3DFS Training

Perceptual loss

$$\mathcal{L}_{per1} = \sum_{i=1}^{5} \frac{\left\| VGG_i\left(\mathcal{F}_{cdf}^{I_l}\right) - VGG_i(I_l) \right\|}{C_i \times H_i \times W_i}$$

$$\mathcal{L}_{per2} = \sum_{i=1}^{5} \frac{\left\| VGG_i\left(\hat{I}_l\right) - VGG_i(I_l) \right\|}{C_i \times H_i \times W_i}$$

Adversarial loss

$$\mathcal{L}_{adv1} = \mathcal{L}_{G1}\left(\mathcal{F}_{cdf}^{I_l}\right) + \mathcal{L}_{D1}\left(\mathcal{F}_{cdf}^{I_l}, I_l\right)$$

$$\mathcal{L}_{adv2} = \mathcal{L}_{G2}\left(\hat{I}_l\right) + \mathcal{L}_{D2}\left(\hat{I}_l, I_l\right)$$

Texture loss

$$\mathcal{L}_{tex} = \sum_{i=1}^{5} \frac{\left\| \mathbf{Gram}\left(VGG_i\left(\hat{I}_l\right)\right) - \mathbf{Gram}\left(VGG_i(I_l)\right) \right\|}{C_i \times H_i \times W_i}$$

Pixel loss

$$\mathcal{L}_{pixel} = \left\| I_l - \hat{I}_l \right\|$$

Flow loss

$$\mathcal{L}_{flow} = \left\| \Gamma_{original}^{I_l} - \Gamma_{fine}^{I_l} \right\|$$

Identity loss

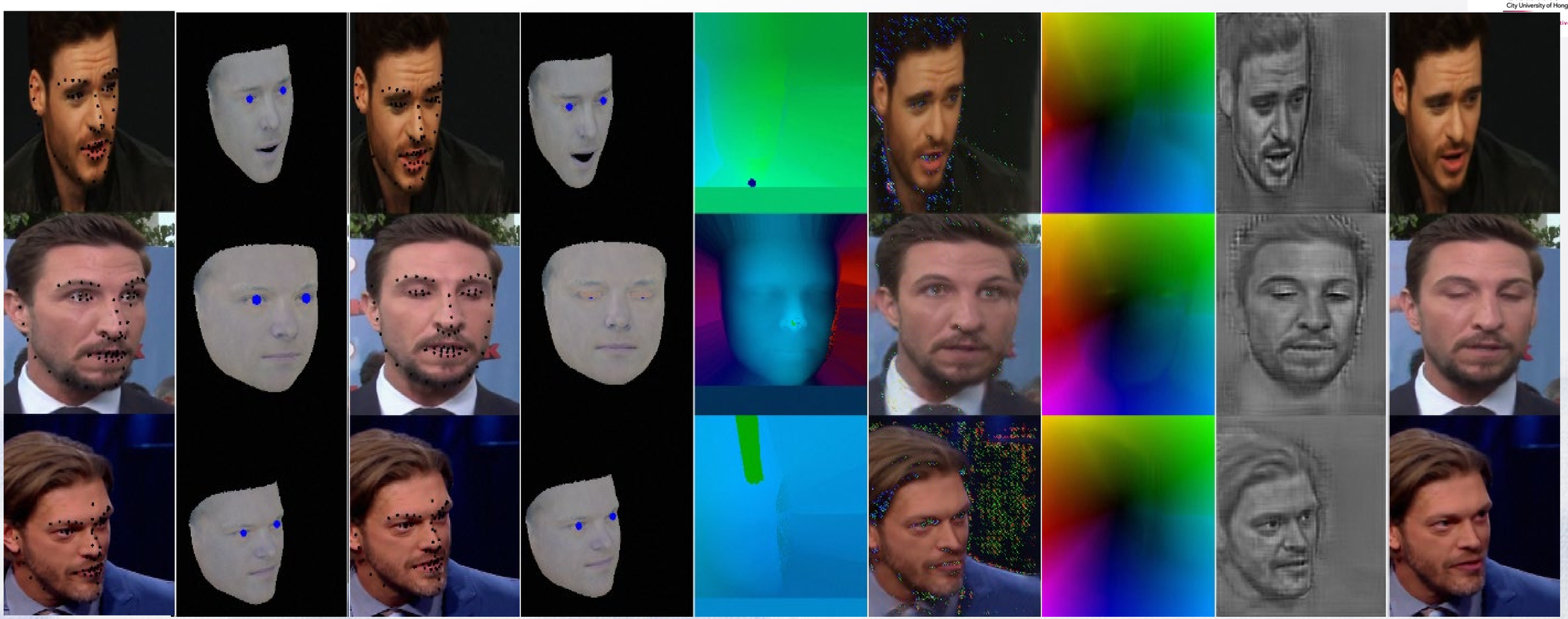
$$\mathcal{L}_{id} = \sum_{i=1}^{4} \frac{\left\| Arc_i\left(\hat{I}_l\right) - Arc_i(I_l) \right\|}{C_i \times H_i \times W_i}$$

Total loss

$$\mathcal{L}_{total} = \lambda_{per1}\mathcal{L}_{per1} + \lambda_{per2}\mathcal{L}_{per2} + \lambda_{adv1}\mathcal{L}_{adv1} + \lambda_{adv2}\mathcal{L}_{adv2}$$

$$+\lambda_{flow}\mathcal{L}_{flow} + \lambda_{pixel}\mathcal{L}_{pixel} + \lambda_{tex}\mathcal{L}_{tex} + \lambda_{id}\mathcal{L}_{id}$$

3DFS decoding flow visualization



Key frame | Key mesh | Current frame | Current mesh | Mesh flow | Coarse deformed frame | Dense motion map | Attention map | Final output

Experimental results

Test sequences

Resolution: 256x256
Frame rate: 25 fps
Duration: 10 sec
Format: RGB444

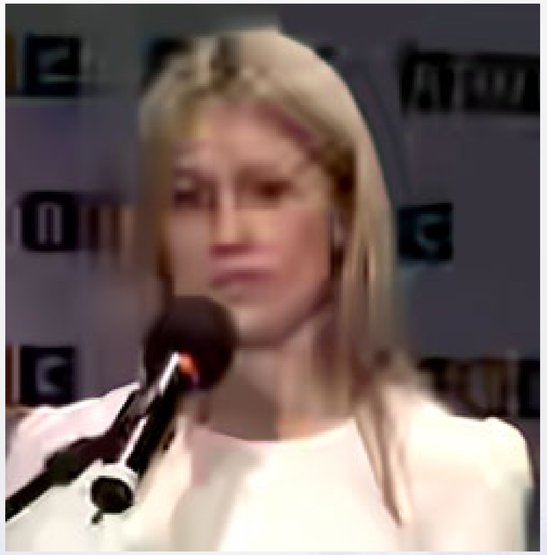
Cropped from open source database:
https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html

Distortion metrics

- Conventional metrics:
  - Widely used in video coding research
  - PSNR: Peak Signal to Noise Ratio
  - SSIM: Structural SIMilarity index
- Learning-based perceptual metrics:
  - More suitable for generative methods that do not optimize for pixel-level fidelity
  - LPIPS: *Learned Perceptual Image Patch Similarity*
  - DISTS: *Deep Image Structure and Texture Similarity*
- All metrics calculated with the open-source implementation from https://github.com/dingkeyan93/IQA-optimization

R. Zhang, et al. "The unreasonable effectiveness of deep features as a perceptual metric." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.
K. Ding, et. al., “Image quality assessment: Unifying structure and texture similarity," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020
K. Ding, et al. "Comparison of full-reference image quality models for optimization of image processing systems." International Journal of Computer Vision 129.4 (2021): 1258-1281

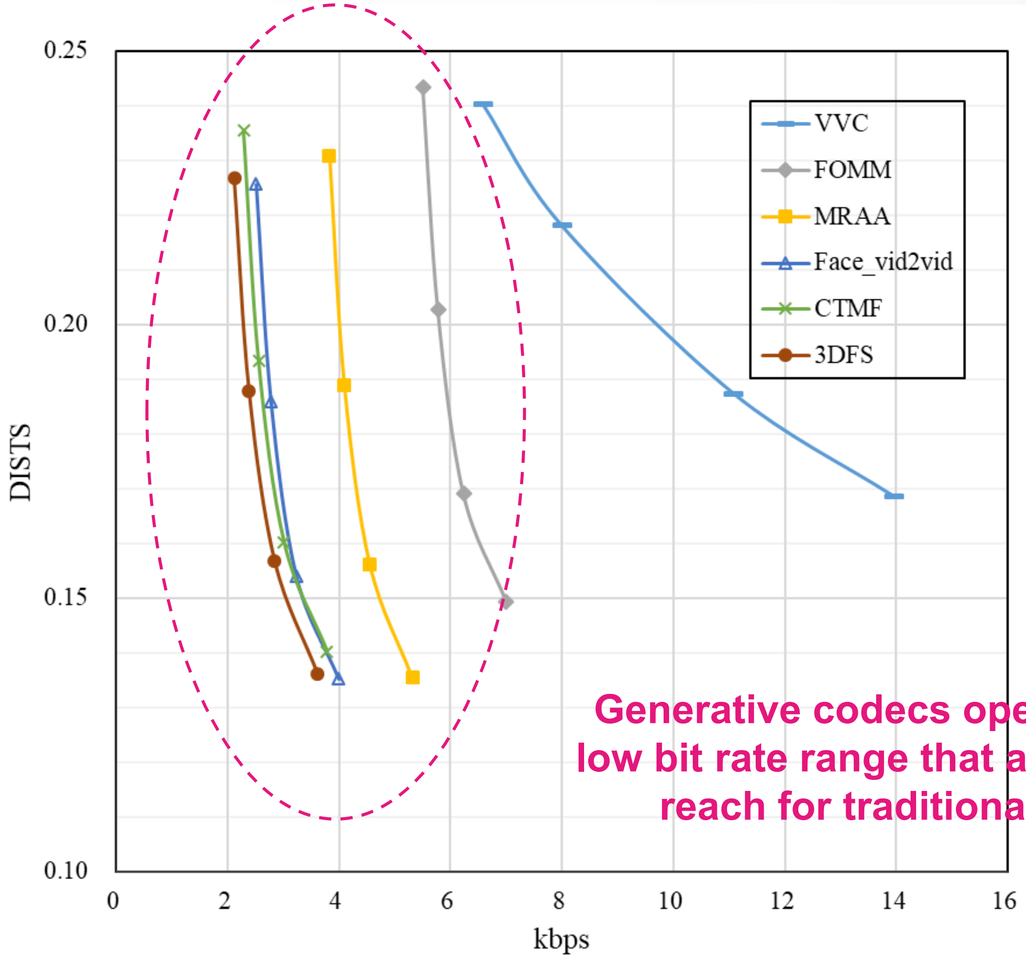
Distortion metrics: a visualization

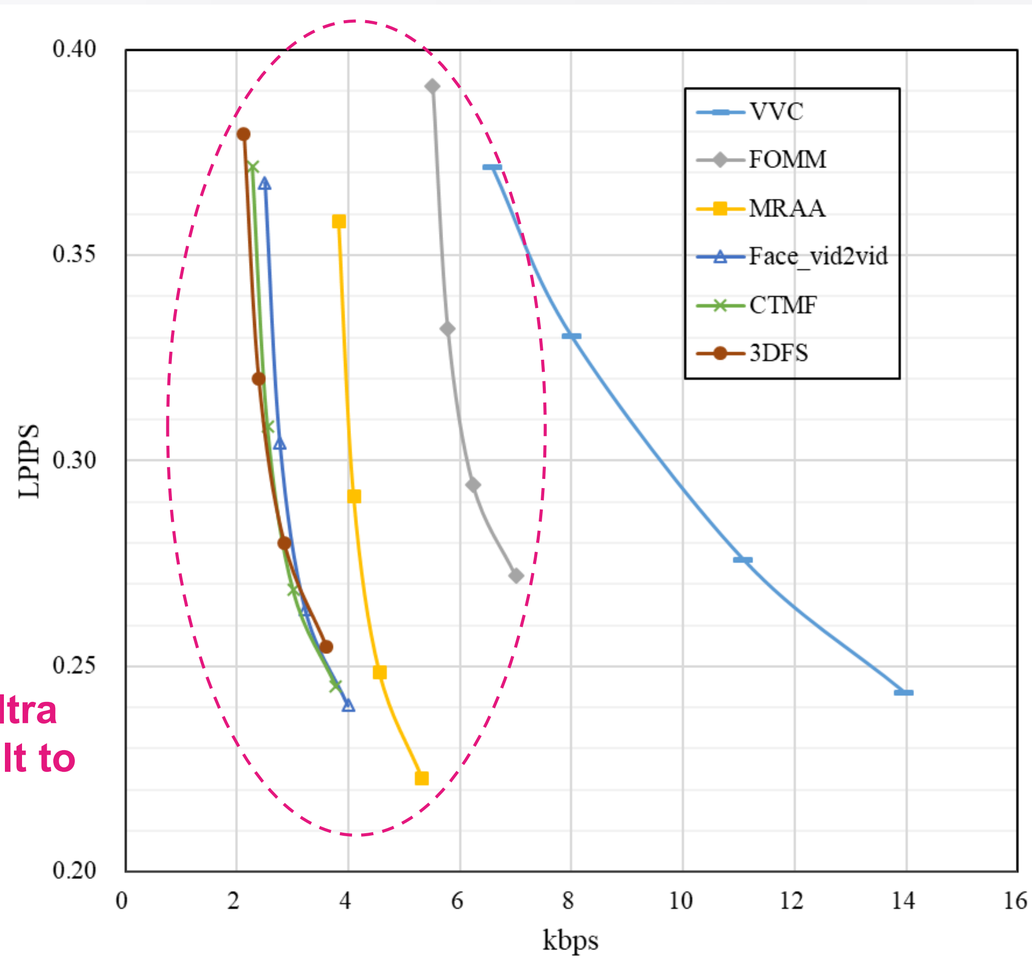


Original, frame #2	VVC, QP 47	VVC, QP 52	FOMM
PSNR (↑)	30.18	27.31	24.36
SSIM (↑)	0.8816	0.8107	0.8139
LPIPS (↓)	0.2275	0.3399	0.1637
DISTS (↓)	0.1458	0.2007	0.1092

Focus on perceptual metrics, as generative methods do ***not*** optimize for pixel-level fidelity

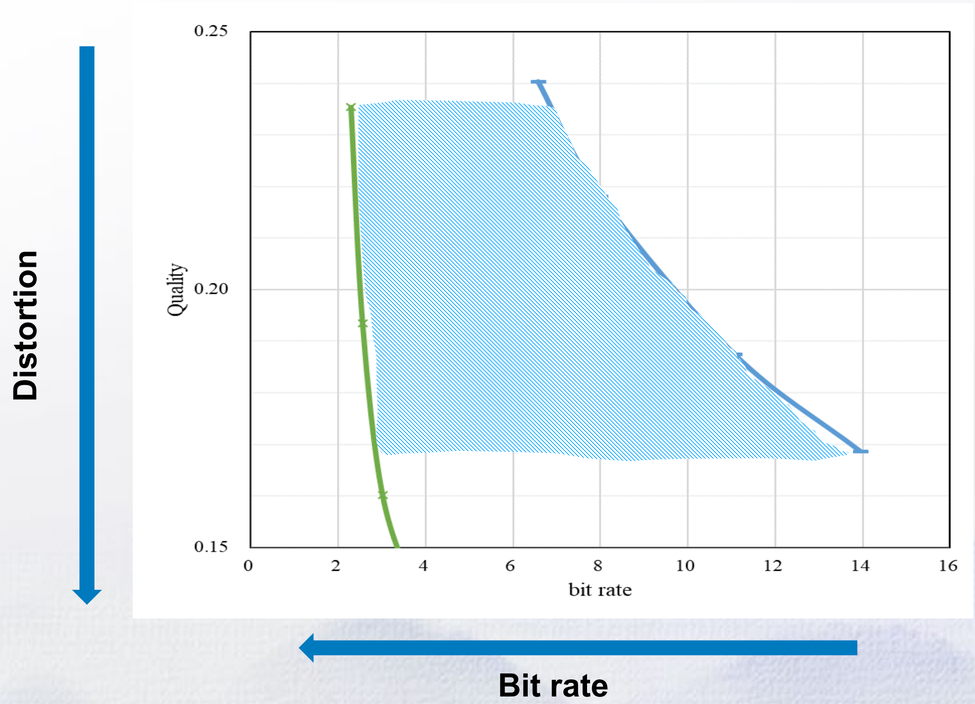
Rate-distortion performance: DISTS and LPIPS





Generative codecs operate at ultra low bit rate range that are difficult to reach for traditional codecs

Quantifying the efficiency of different coding schemes



Bjøntegaard delta rate (BD rate) calculates the average rate savings for the shaded area between two rate-distortion curves (in log domain)

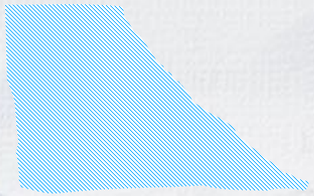
It is widely used to quantitatively measure the average rate savings between two coding schemes

BD-rate savings of generative methods

Anchor: VTM-10.0 (VVC reference encoder), low-delay B config

	FOMM	MRAA	Face_vid2vid	CTMF (ours)	3DFS (ours)
DISTS	-35.3%	-57.3%	-71.8%	-72.6%	-75.4%
LPIPS	-26.6%	-55.5%	-68.1%	-69.7%	-70.3%

Significant rate savings compared to VVC in terms of perceptual metrics



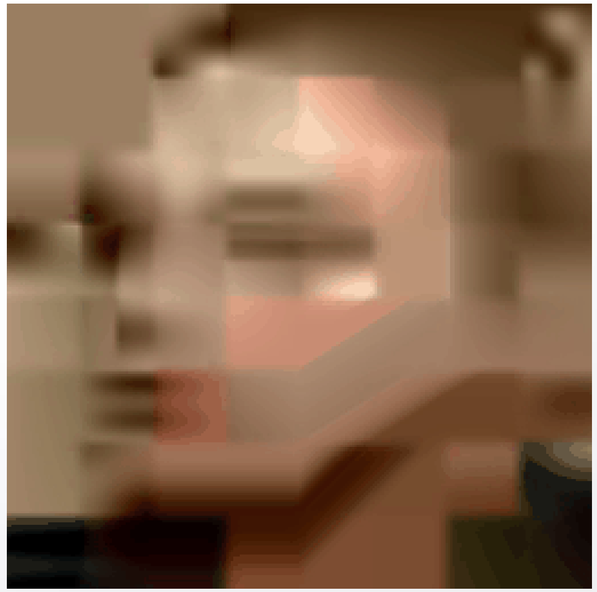
Quality comparison at similar bit rates

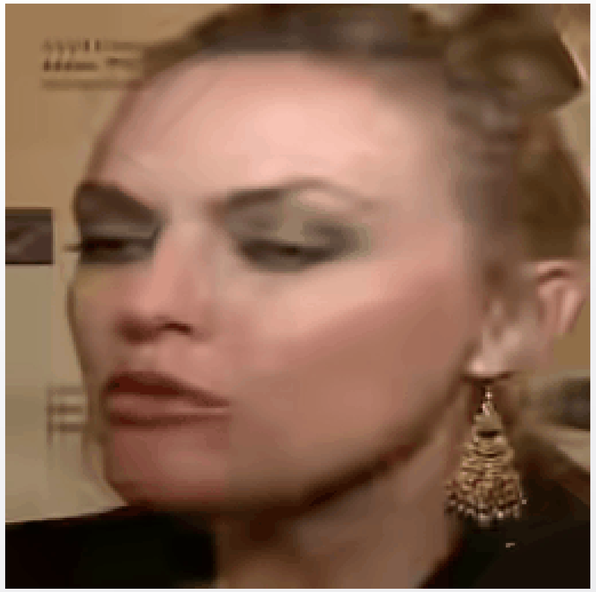
達摩院
ALIBABA DAMO ACADEMY

CityU 香港城市大學 City University of Hong Kong
專業 創新 胸懷全球 Professional·Creative For The World

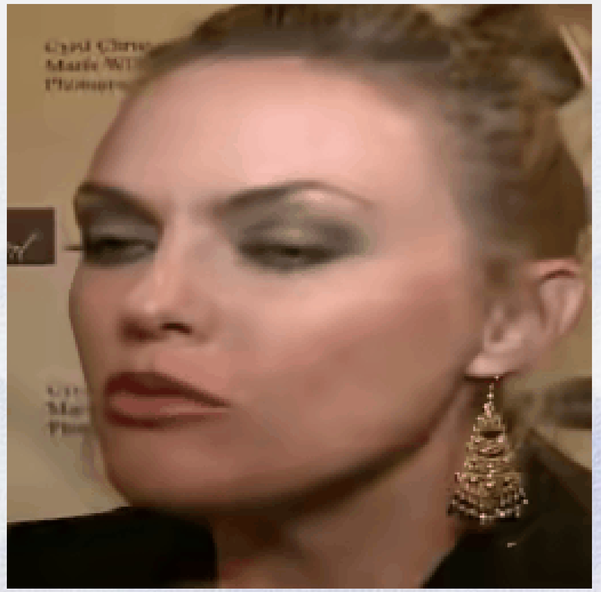
VVC

Bit rate	3.328k
DISTS	0.372
LPIPS	0.578
PSNR	25.414
SSIM	0.744





Original



CTMF

Bit rate	2.942k
DISTS	0.155
LPIPS	0.284
PSNR	22.874
SSIM	0.722

Face_Vid2Vid

Bit rate	2.814k
DISTS	0.190
LPIPS	0.321
PSNR	22.876
SSIM	0.712

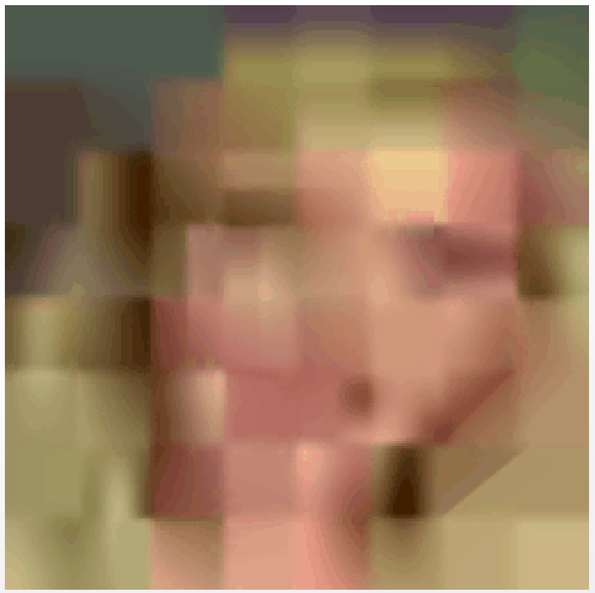
3DFS

Bit rate	2.751k
DISTS	0.152
LPIPS	0.297
PSNR	21.895
SSIM	0.703

Quality comparison at similar bit rates

VVC

Bit rate	3.715k
DISTS	0.381
LPIPS	0.597
PSNR	25.193
SSIM	0.727



Face_Vid2Vid

Bit rate	3.098k
DISTS	0.180
LPIPS	0.316
PSNR	21.461
SSIM	0.672

Original

CTMF

Bit rate	2.729k
DISTS	0.183
LPIPS	0.320
PSNR	21.349
SSIM	0.676

3DFS

Bit rate	2.954k
DISTS	0.153
LPIPS	0.296
PSNR	20.997
SSIM	0.668

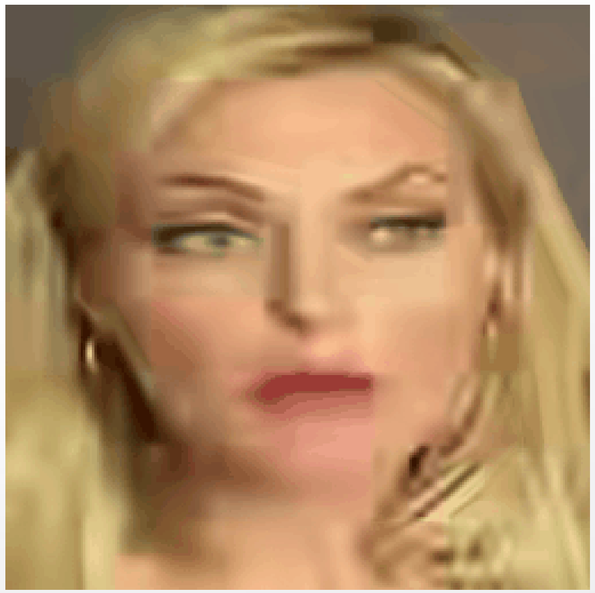
Bit rate comparison at similar quality

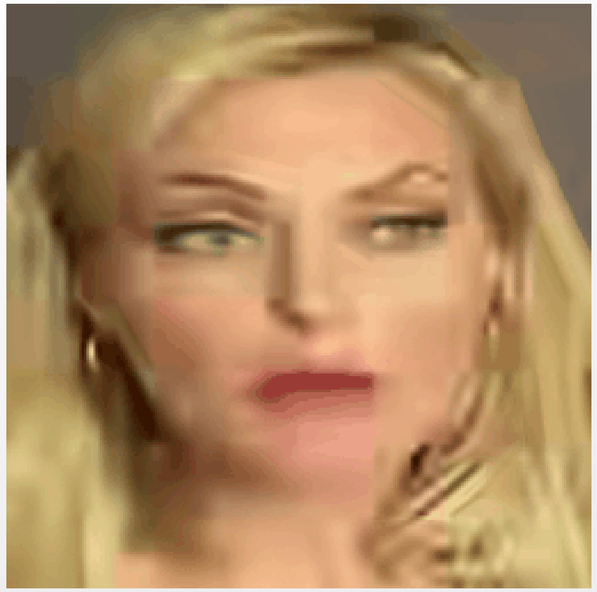
達摩院
ALIBABA DAMO ACADEMY

CityU 香港城市大學 City University of Hong Kong
專業 創新 胸懷全球 Professional·Creative For The World

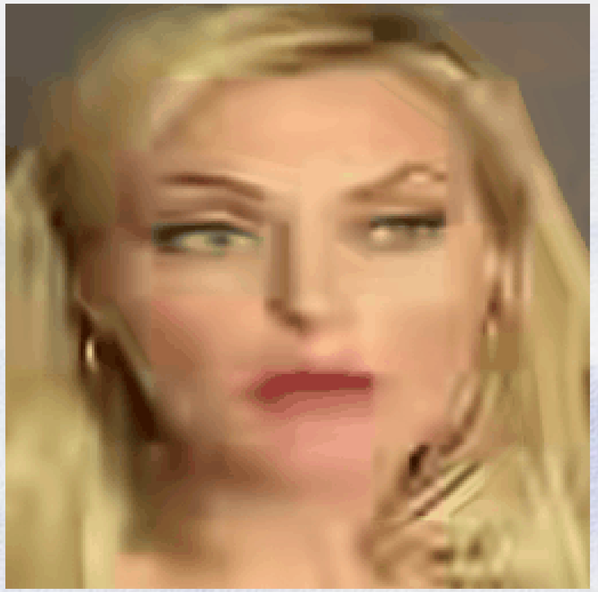
VVC

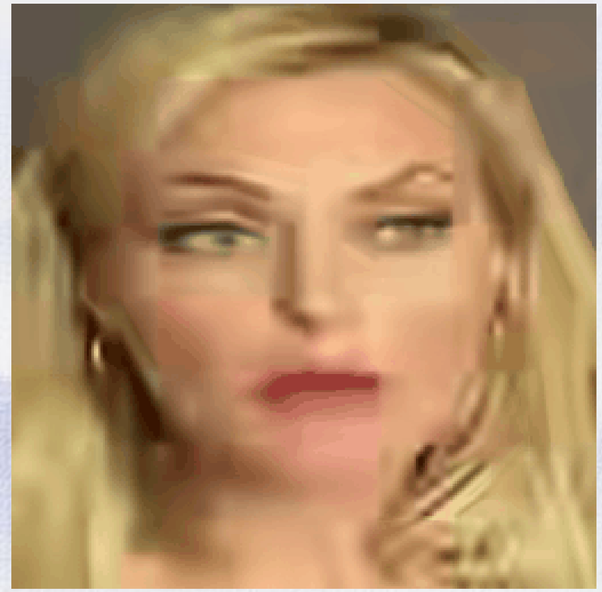
Bit rate	6.528k
DISTS	0.219
LPIPS	0.346
PSNR	32.389
SSIM	0.902





original





Face_Vid2Vid

Bit rate	2.460k
DISTS	0.204
LPIPS	0.356
PSNR	22.964
SSIM	0.707

~1/3 of VVC's bandwidth

CTMF

Bit rate	2.256k
DISTS	0.225
LPIPS	0.365
PSNR	23.091
SSIM	0.715

3DFS

Bit rate	2.105k
DISTS	0.213
LPIPS	0.367
PSNR	22.408
SSIM	0.696

Bit rate comparison at similar quality

達摩院
ALIBABA DAMO ACADEMY

CityU
香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional·Creative
For The World

original

VVC

Bit rate	5.909k
DISTS	0.227
LPIPS	0.344
PSNR	32.365
SSIM	0.909

Face_Vid2Vid

Bit rate	2.308k
DISTS	0.209
LPIPS	0.338
PSNR	22.584
SSIM	0.747

~1/3 of VVC's bandwidth

CFTE

Bit rate	2.170k
DISTS	0.212
LPIPS	0.330
PSNR	23.117
SSIM	0.773

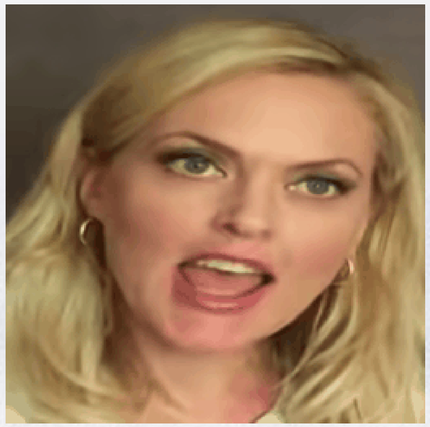
IFVC

Bit rate	2.059k
DISTS	0.205
LPIPS	0.358
PSNR	22.229
SSIM	0.742

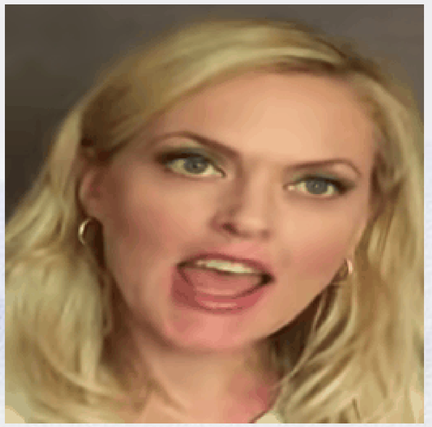
Interacting with facial expression: eyes

Original

$\widehat{\delta}_{eye} = 0.5$



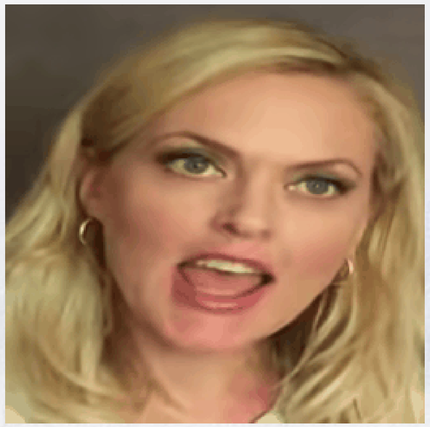
$\widehat{\delta}_{eye} = 3.5$



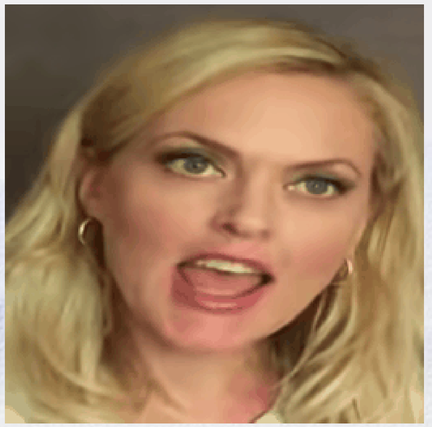
Interacting with facial expression: mouth

Original

$\widehat{\delta}_{mouth} \mathrel{-}= 0.75$



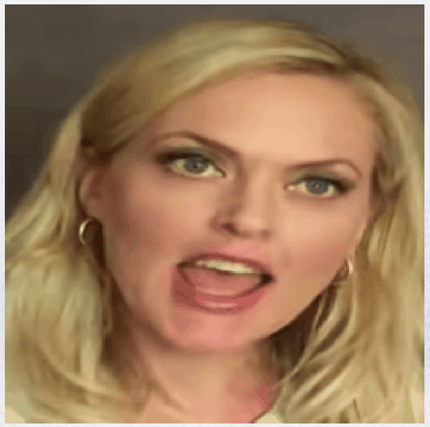
$\widehat{\delta}_{mouth} \mathrel{+}= 0.25$



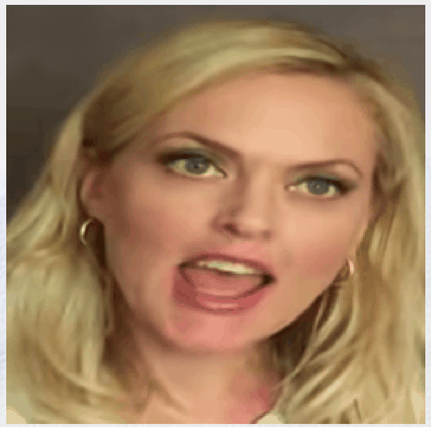
Interacting with head position: rotation

Original

$\widehat{\delta}_{rot} \mathrel{-}= 0.2$



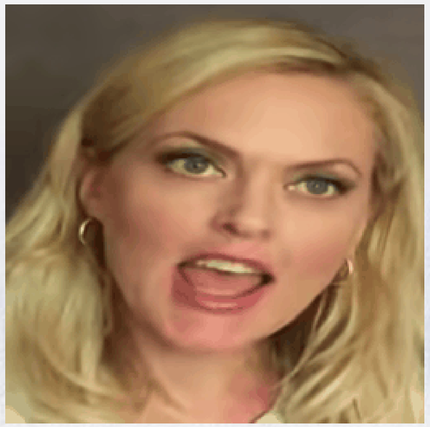
$\widehat{\delta}_{rot} \mathrel{+}= 0.125$



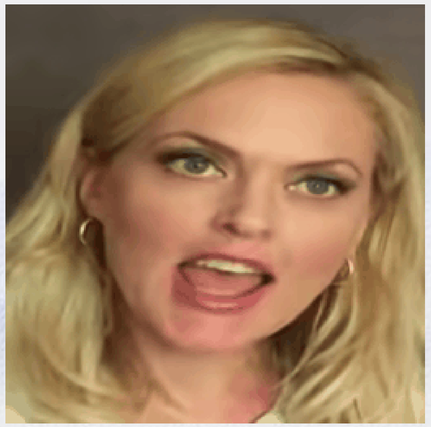
Interacting with head position: zoom

Original

$\widehat{\delta}_{trans} -= 1.0$



$\widehat{\delta}_{trans} += 1.0$



Animating virtual faces

Concluding remarks

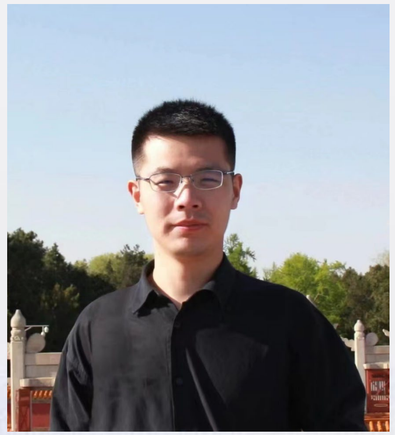
Traditional video coding	• Foundation for generations of video codec standards • Widely deployed in many commercial products and services
Learning-based video coding	• End-to-end trainable, performance competitive to H.265/HEVC • On-going active research will bring further improvements
Generative networks	• Superior performance for talking-face video • Powerful generative capability with very compact representation • Extensibility to generic video content to be explored

Acknowledgment

Dr. Shiqi Wang
Assistant Prof.
City University of Hong Kong

Bolin Chen
Ph.D. student
City University of Hong Kong

Binzhe Li
Ph.D. student
City University of Hong Kong



Dr. Zhao Wang
Assoc. researcher
Peking University

達摩院
ALIBABA DAMO ACADEMY

CityU
香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional·Creative
For The World

Questions?