北京大學
PEKING UNIVERSITY

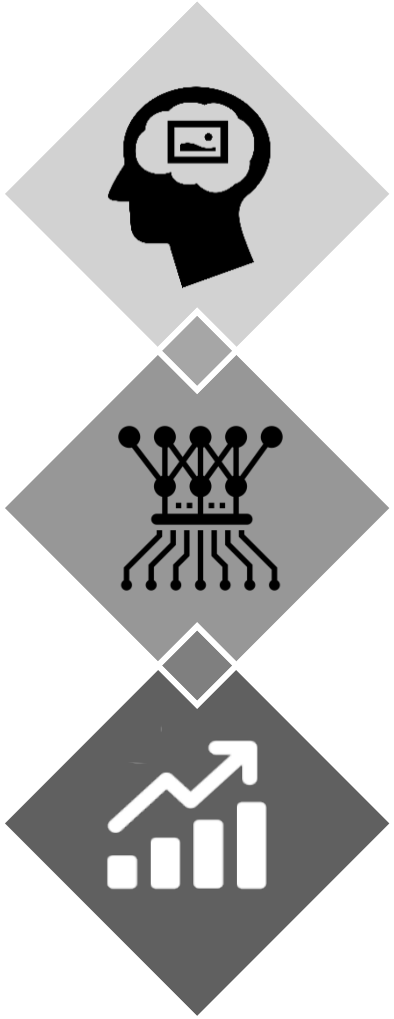
IEEE ICME 2020

Towards Coding for Human and Machine Vision: A Scalable Image Coding Approach

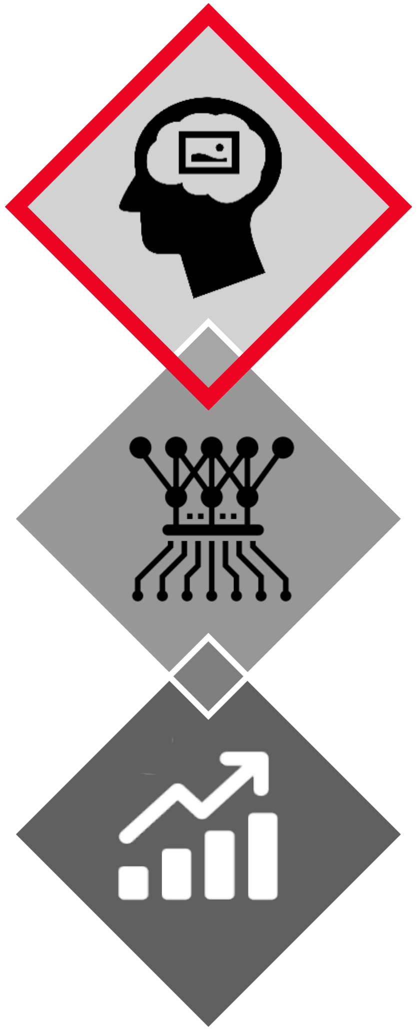
Yueyu Hu, Shuai Yang, Wenhan Yang, Ling-Yu Duan, Jiaying Liu

Peking University

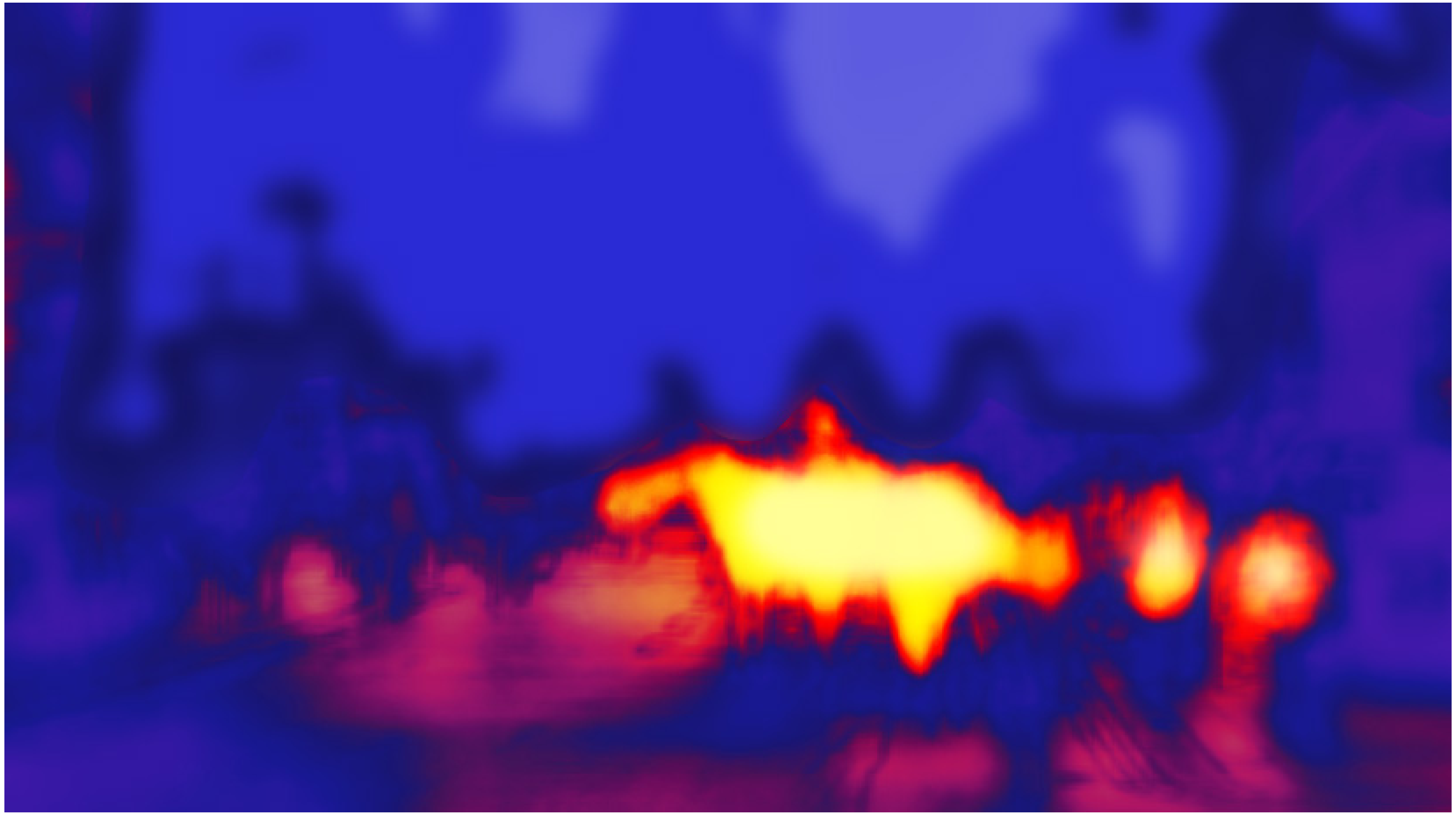
SCENE

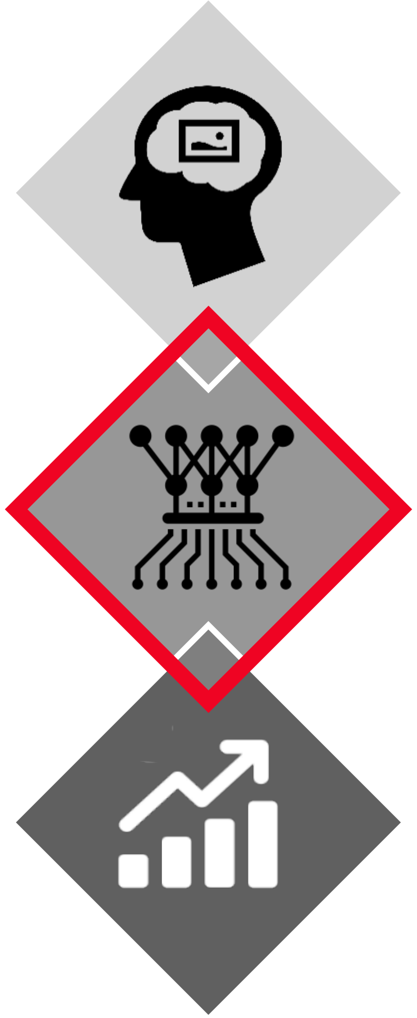


WHAT HUMANS SEE

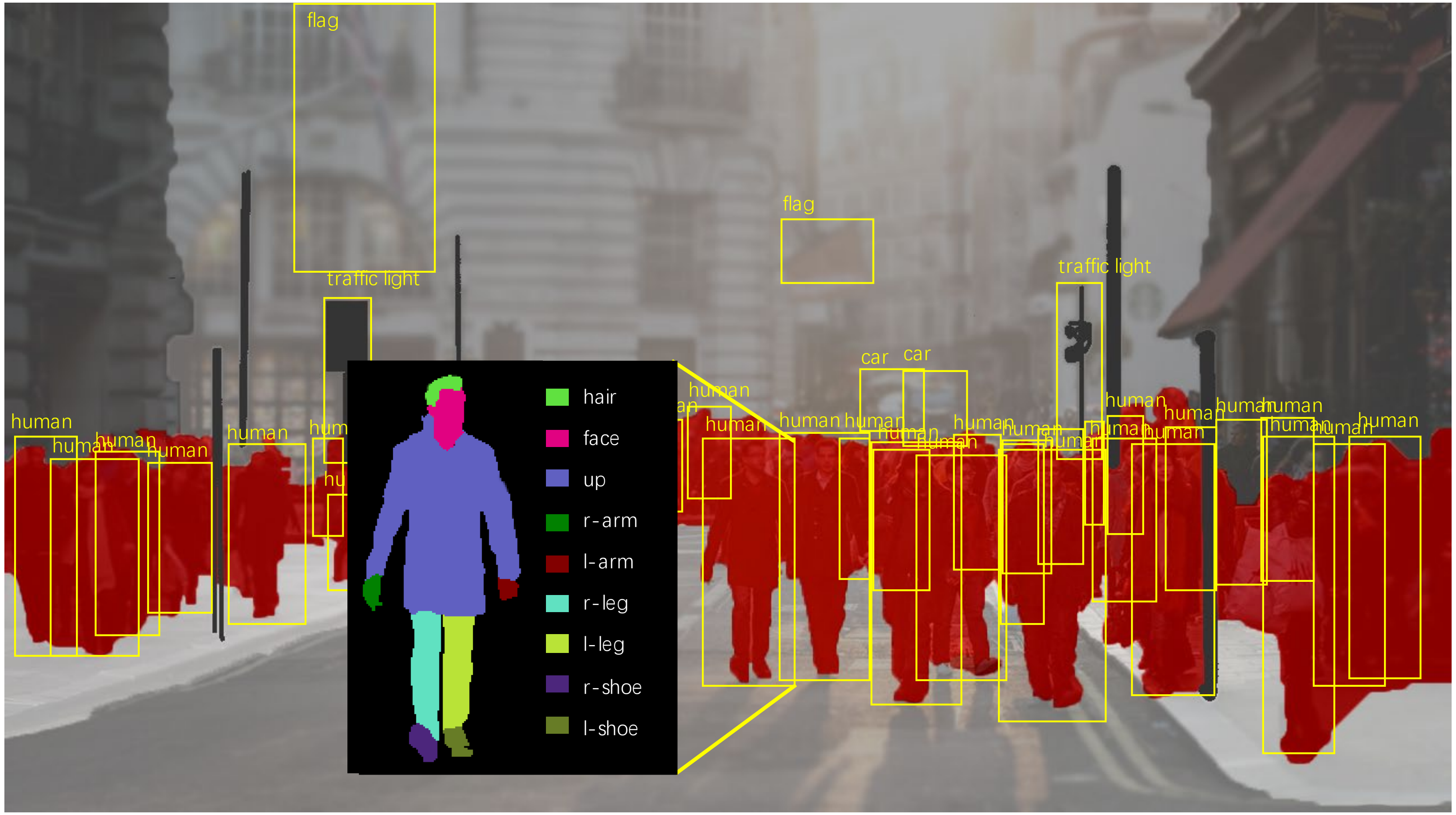


MACHINE FEATURES





MACHINE ANALYTICS



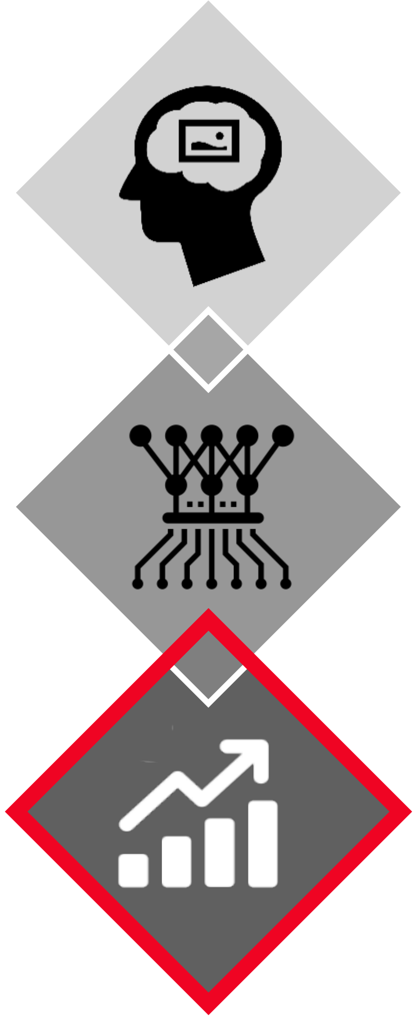


IMAGE CODING FOR WHOM?

HIGH-LEVEL FEATURES FOR MACHINE VISION

?

RECONSTRUCTED IMAGES FOR HUMAN VISION

AMOUNT OF INFORMATION

MACHINE VISION

1) Feature Extraction
- Redundant → Compact
- **Compressed** Information

2) **Regression**
- Feature → Label
- Further Compressed

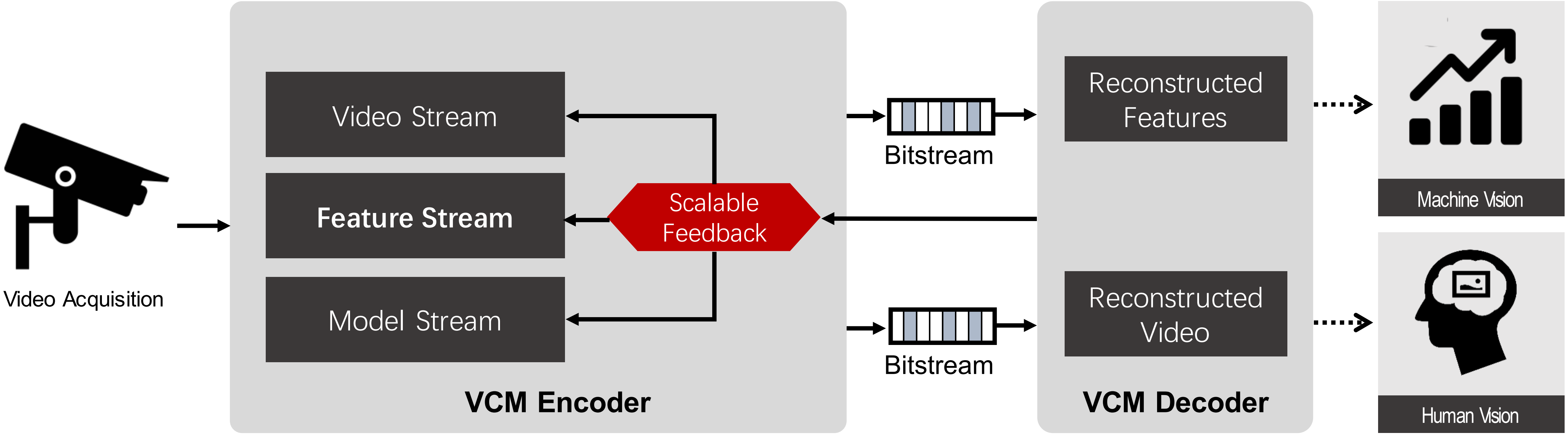
HUMAN VISION

1) Feature Extraction
- Degraded→ High-Quality
- **Enhanced** Information

2) Guided **Reconstruction**
- Feature → Image
- Information Generation

IMAGE CODING NEXTGEN

- Scalable (according to utilizations)
- Efficient compression for joint human and machine vision



Ling-Yu Duan, Jiaying Liu, Wenhan Yang, Tiejun Huang, Wen Gao. Video Coding for Machines: A Paradigm of Collaborative Compression and Intelligent Analytics. arXiv:2001.03569, 2020

INFORMATION DENSITY SPECTRUM

- Descriptor coding for efficient machine vision analytics (low bit-rate)
- Sophisticated video codecs for improved human vision (high bit-rate)

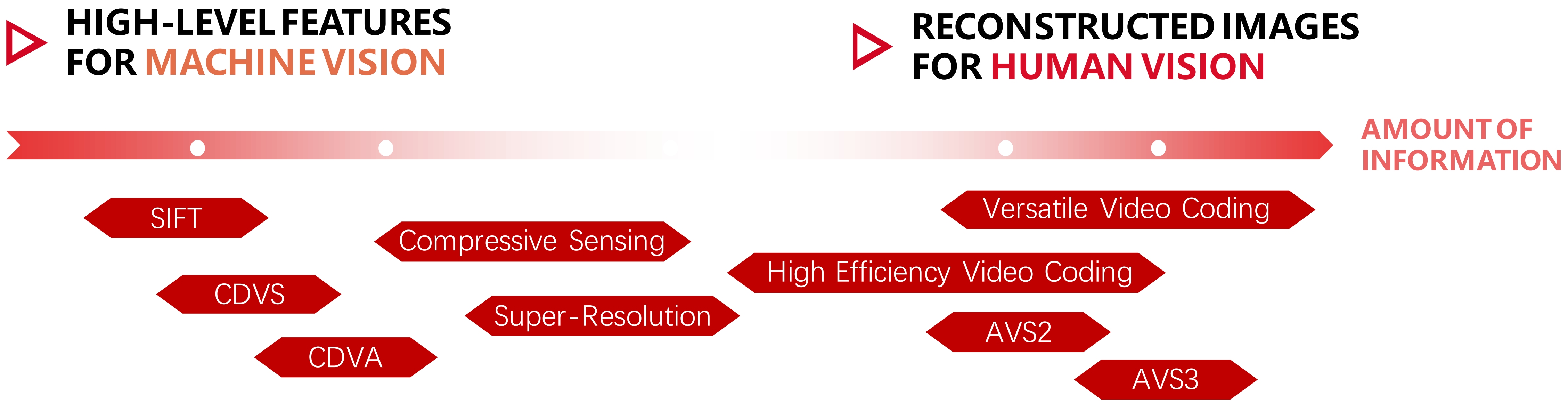
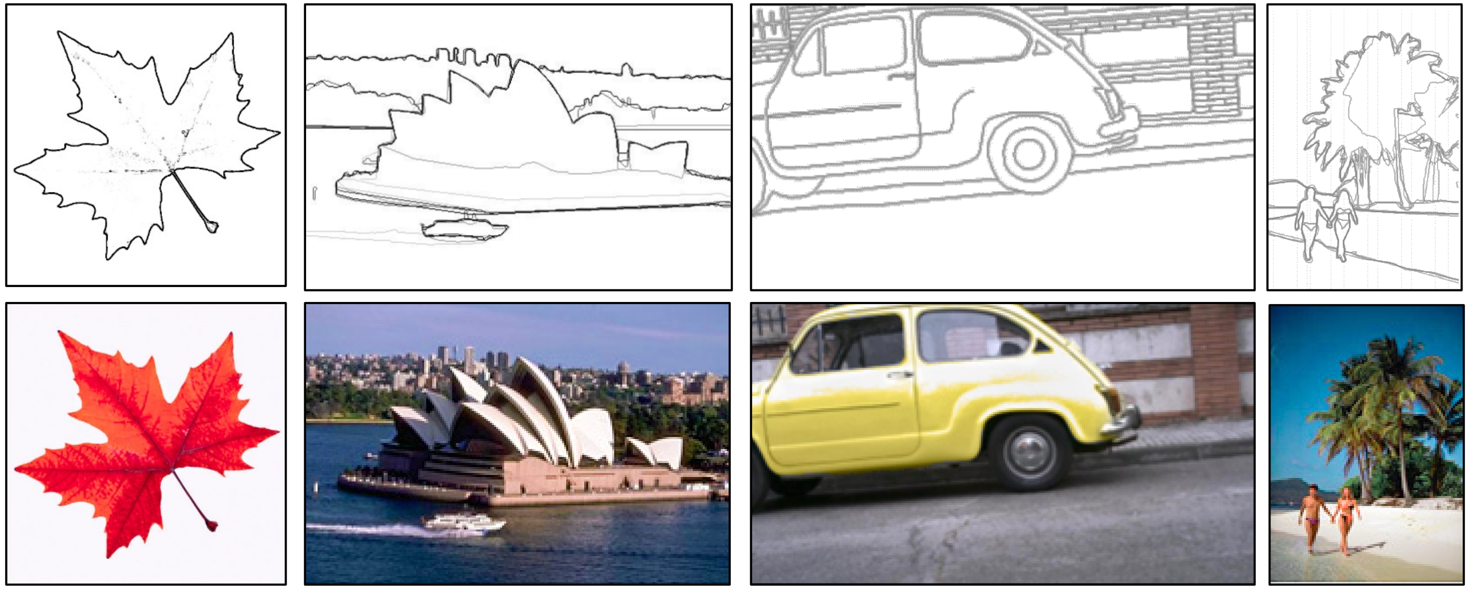


IMAGE REPRESENTATIONS

EDGES



PROS

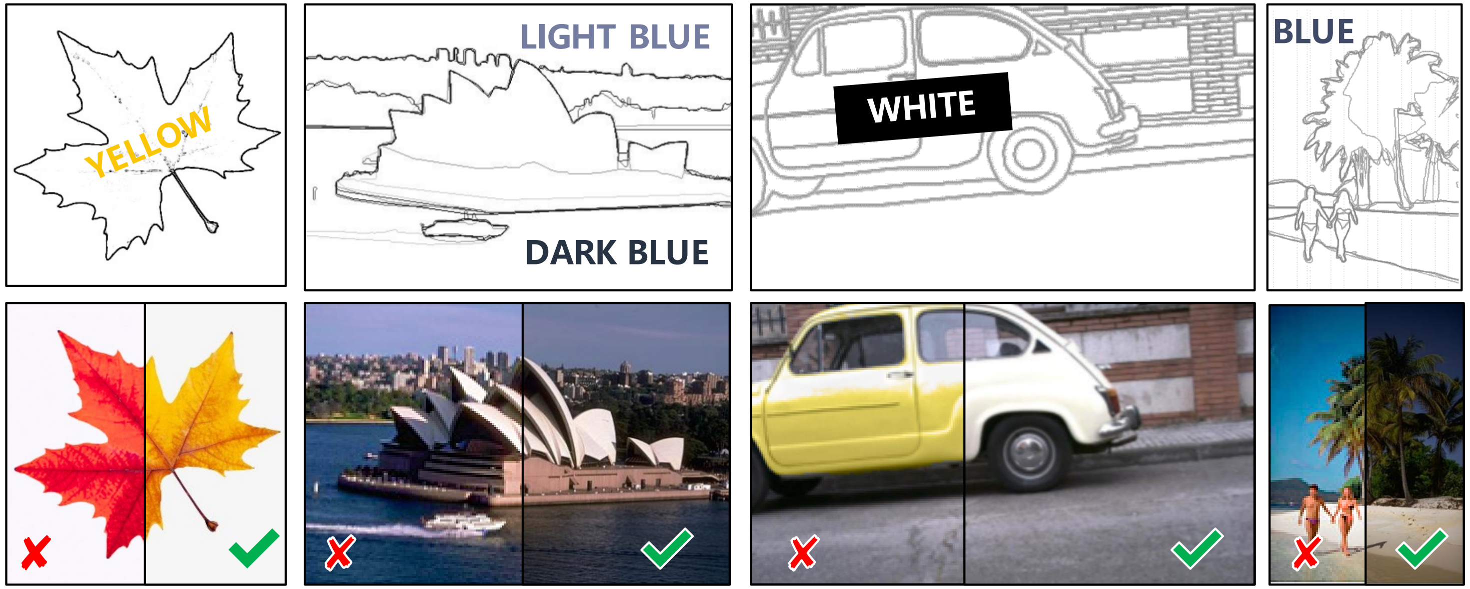
- Efficient for structural information
- Maintain scalability
- Sparse and light-weight
- Supports smooth scaling

CONS

- Inefficient for details in images
- Ambiguous in color

IMAGE REPRESENTATIONS

COLOR



PROS

- Avoid color ambiguity
- Sparse and compact
- Related to visual fidelity

CONS

- Usually randomly distributed
- Inefficient for further compress

HUMAN FACES

Analytics of Faces

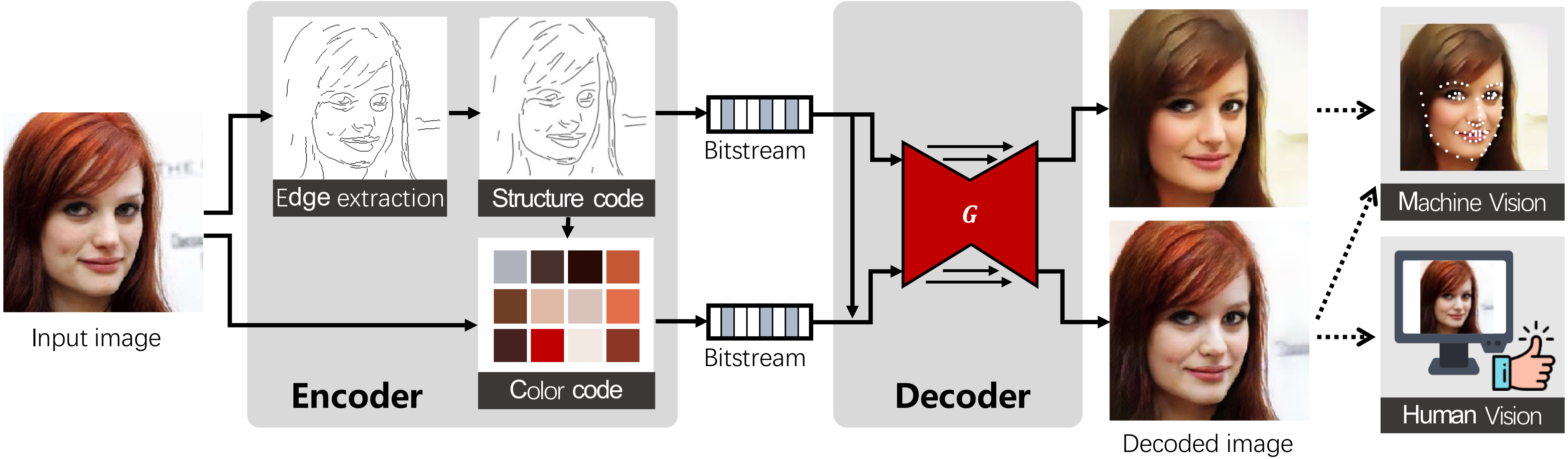
Faces are naturally salient area in images we are looking at.

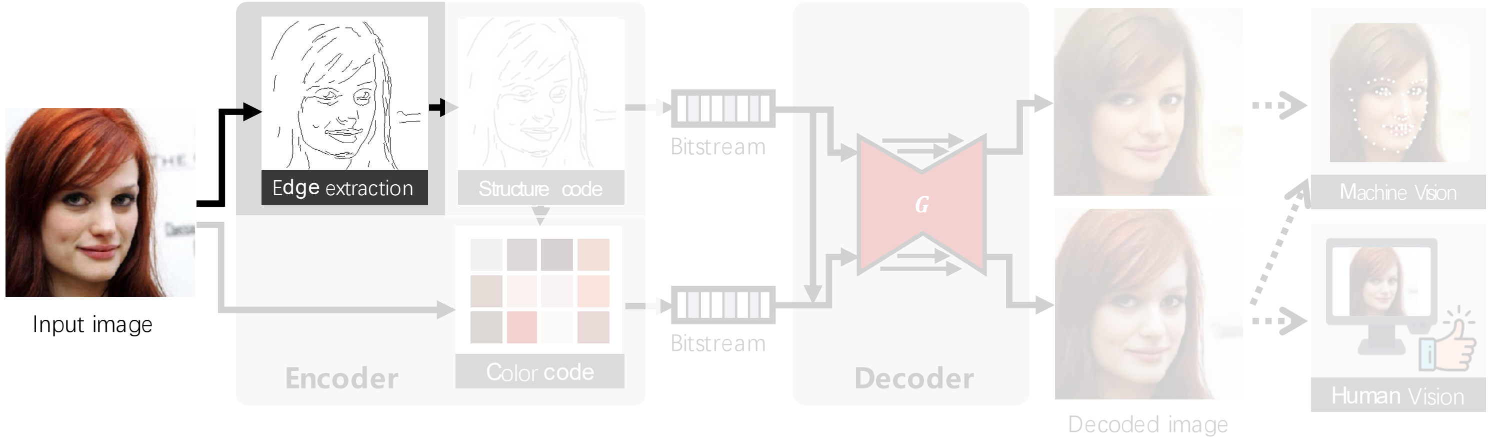
Machine vision systems to analysis faces have been widely developed.

It is the reflection of humanity in technology.

SCALABLE FRAMEWORK

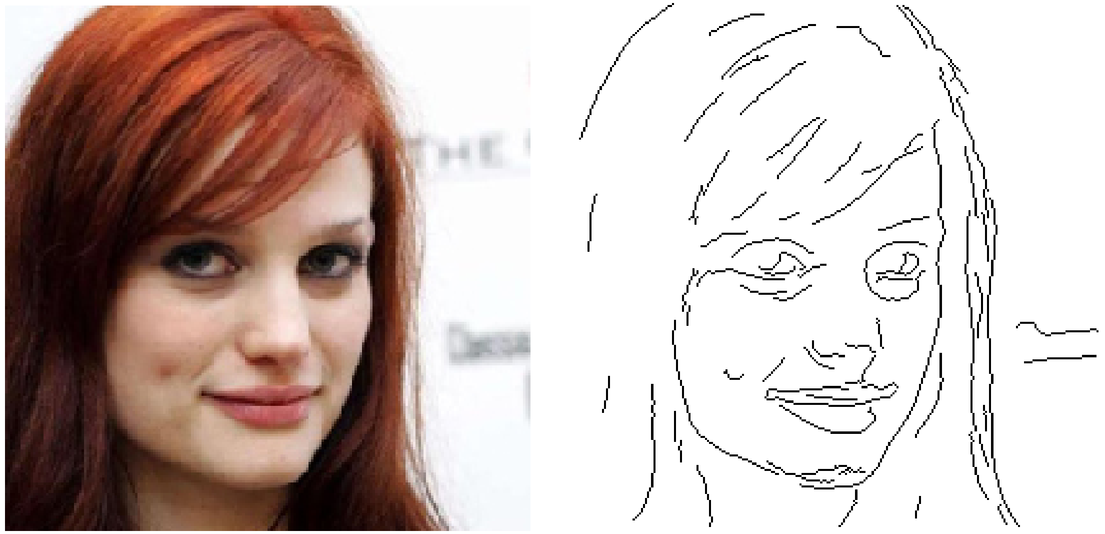
- Conceptual compression to achieve high quality with compact features
- Scalable bit-stream for different tasks
- Vectorized Edges + Sparse Pixels



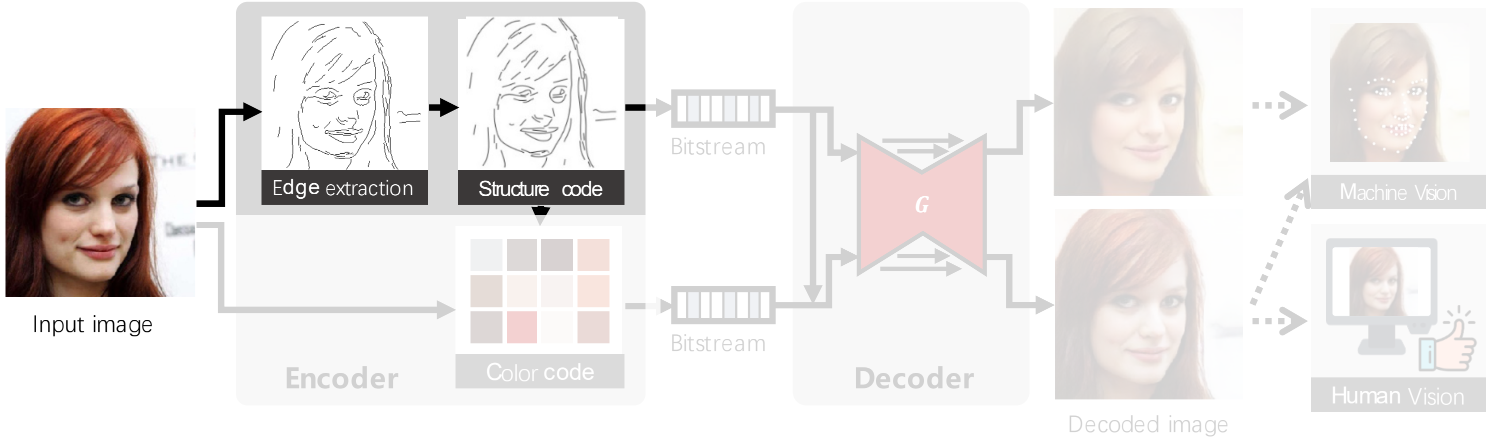


ENCODER • EDGE

- Edge detection via structured forests

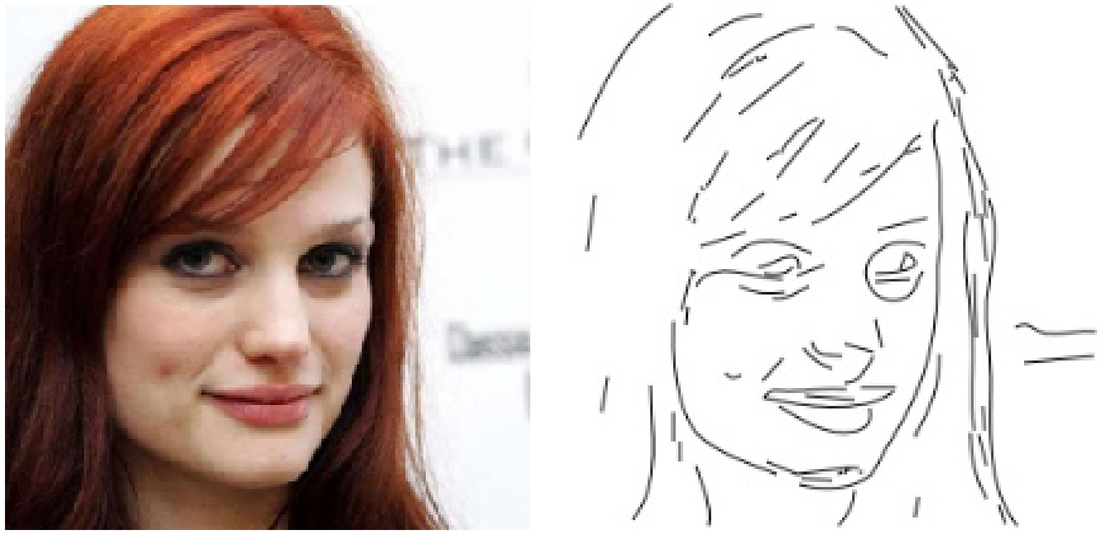


P. Dollar and C. L. Zitnick. Structured forests for fast edge detection. ICCV, 2013.

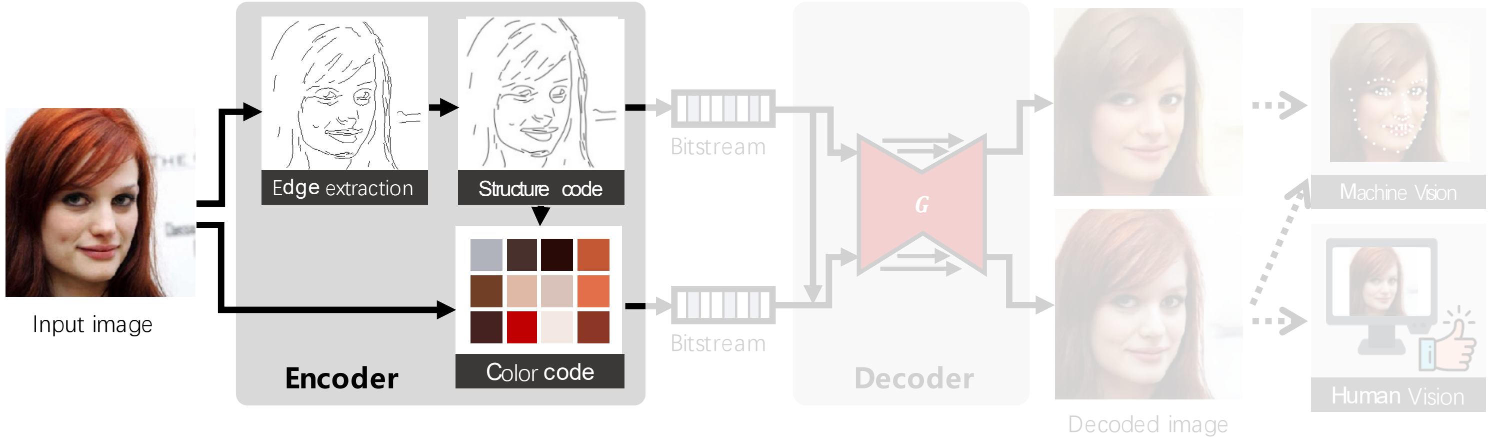


ENCODER • EDGE

- Edge detection via structured forests
- *AutoTrace* to convert edge pixels to vectorized representations
  - Represented by lines and curves
  - Short and trivial edges are screened
- Prediction for Partial Matching (PPM) to losslessly compress vectors

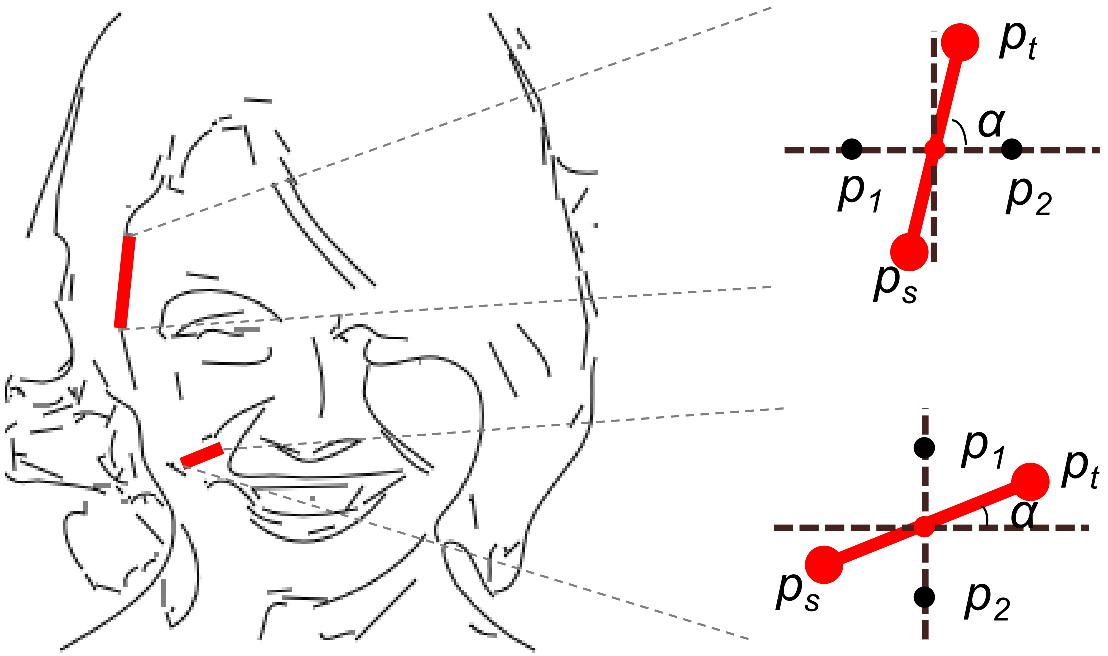


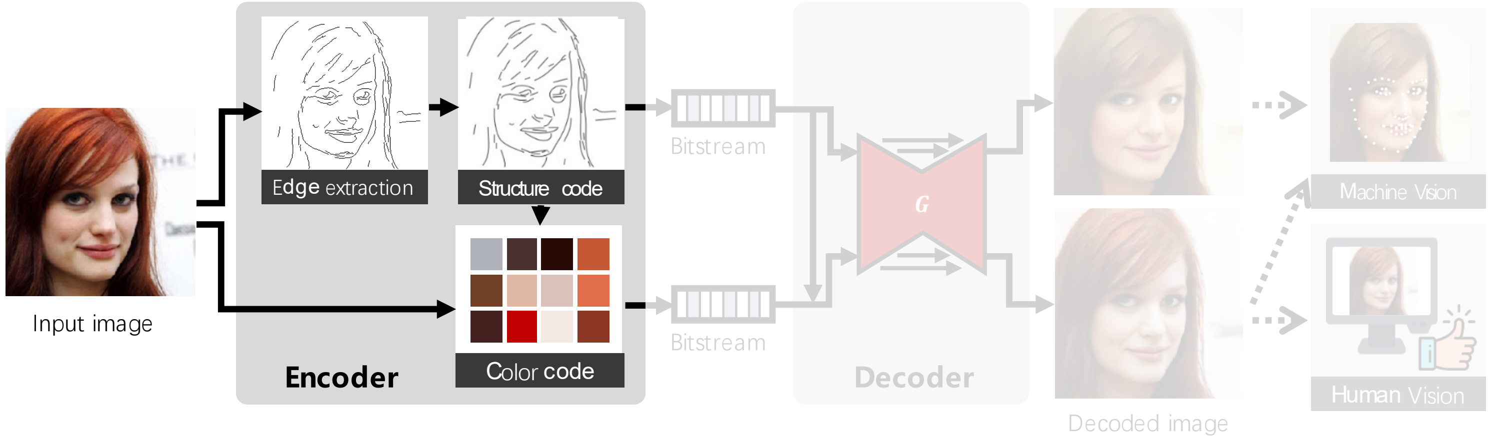
M.Weber. AutoTrace: a program for converting bitmap to vector graphic. 1998. http://autotrace.sourceforge.net/



ENCODER • COLOR

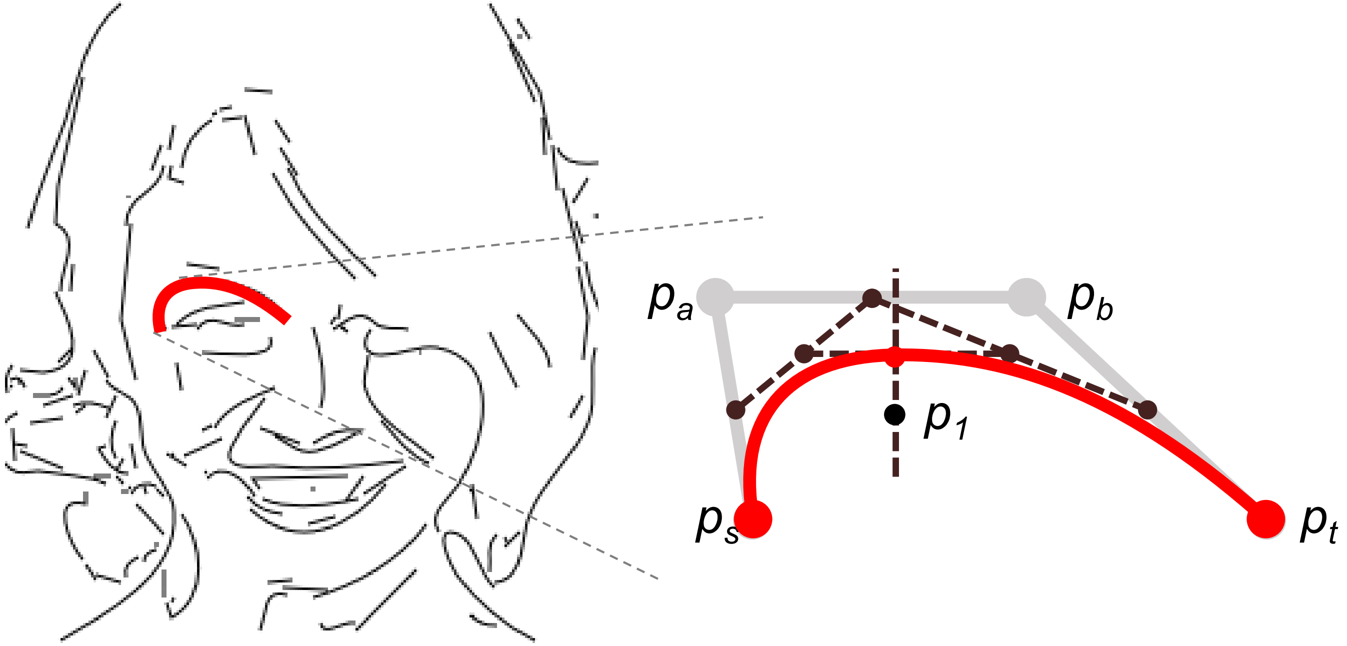
- Sparse pixels sampled according to edges
  - Segments: sample on both sides

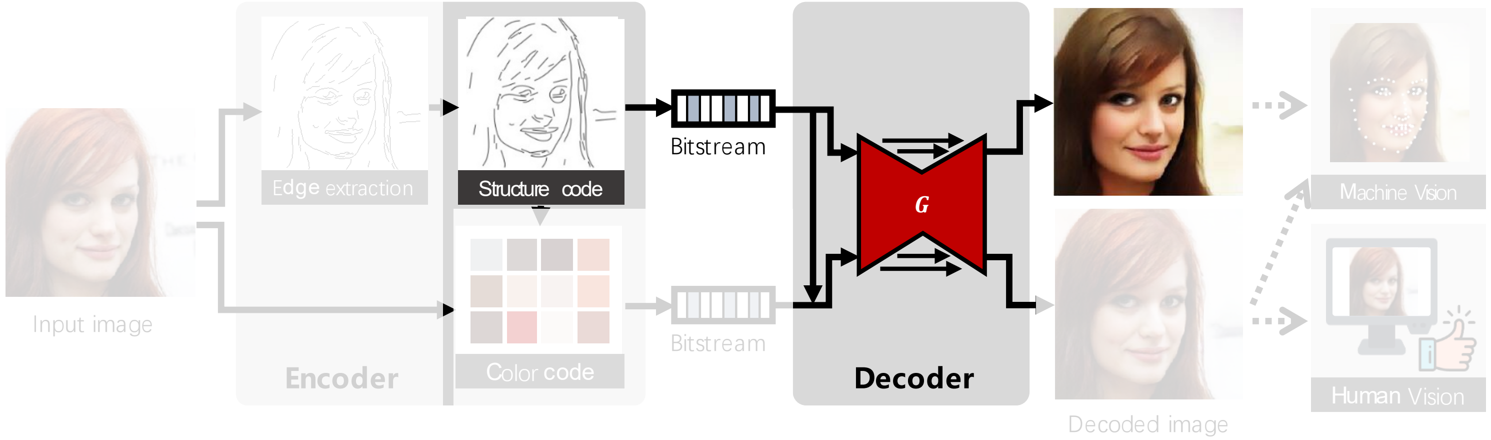




ENCODER • COLOR

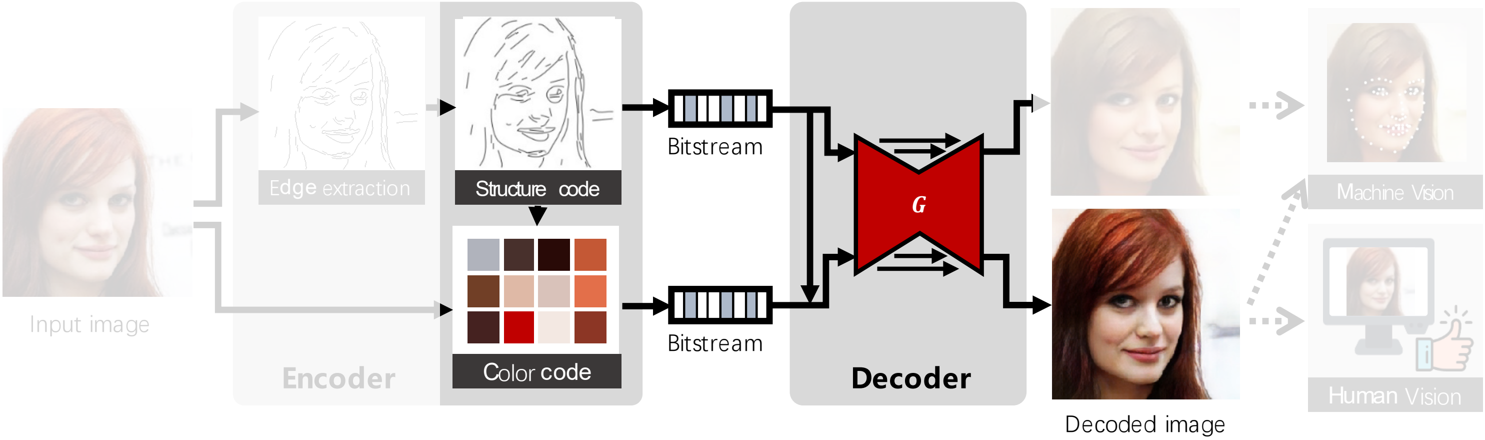
- Sparse pixels sampled according to edges
    - Segments: sample on both sizes
    - Curves: sample on areas with steepest gradients





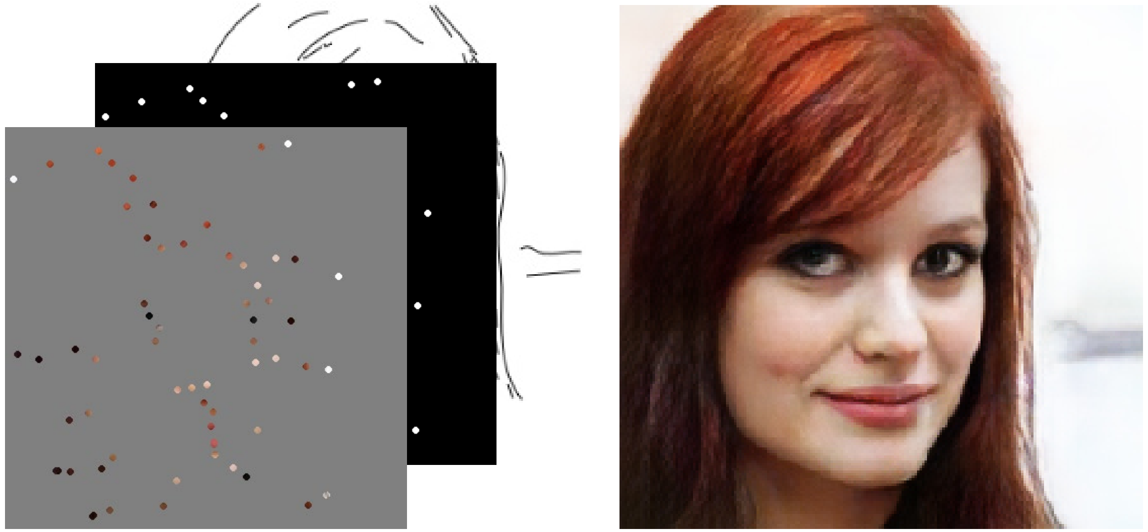
DECODER• MACHINE VISION

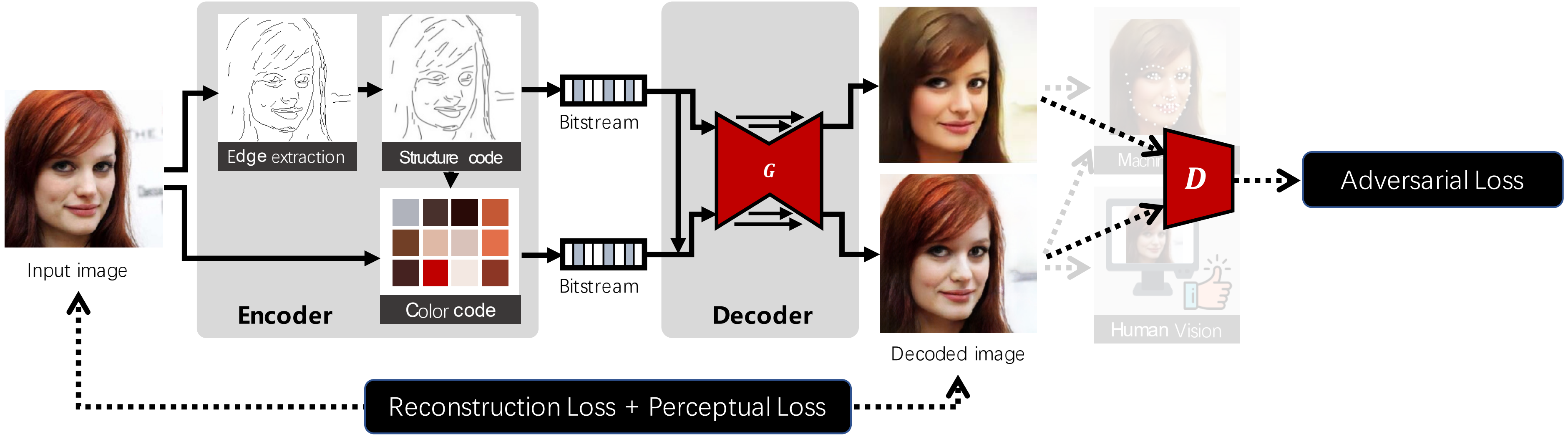
- Image-to-image translation
  - Render pixels with vectorized representations
  - Edge-to-RGB translation



DECODER• HUMAN VISION

- Image-to-image translation
  - Render pixels with vectorized representations
  - Generate masks for completion synthesis
  - Image inpainting





LOSS FUNCTIONS

- Reconstruction Loss

$$\mathcal{L}_r = \mathbb{E}[\lambda_1 \| I_G - I \| + \lambda_2 \mathrm{SSIM}(I_G, I)]$$

- Perceptual Loss

$$\mathcal{L}_p = \mathbb{E}[\lambda_3 \mathrm{PERC}(I_G, I)]$$

- Adversarial Objective

$$\mathcal{L}_G = -\mathbb{E}[D(I_G, E, M)]$$

$$\mathcal{L}_D = \mathbb{E}[\mathrm{ReLU}(\tau + D(I_G, E, M))] + \mathbb{E}[\mathrm{ReLU}(\tau - D(I, E, M))]$$

EXPERIMENTAL RESULTS

HUMAN VISION

Subjective preference survey.

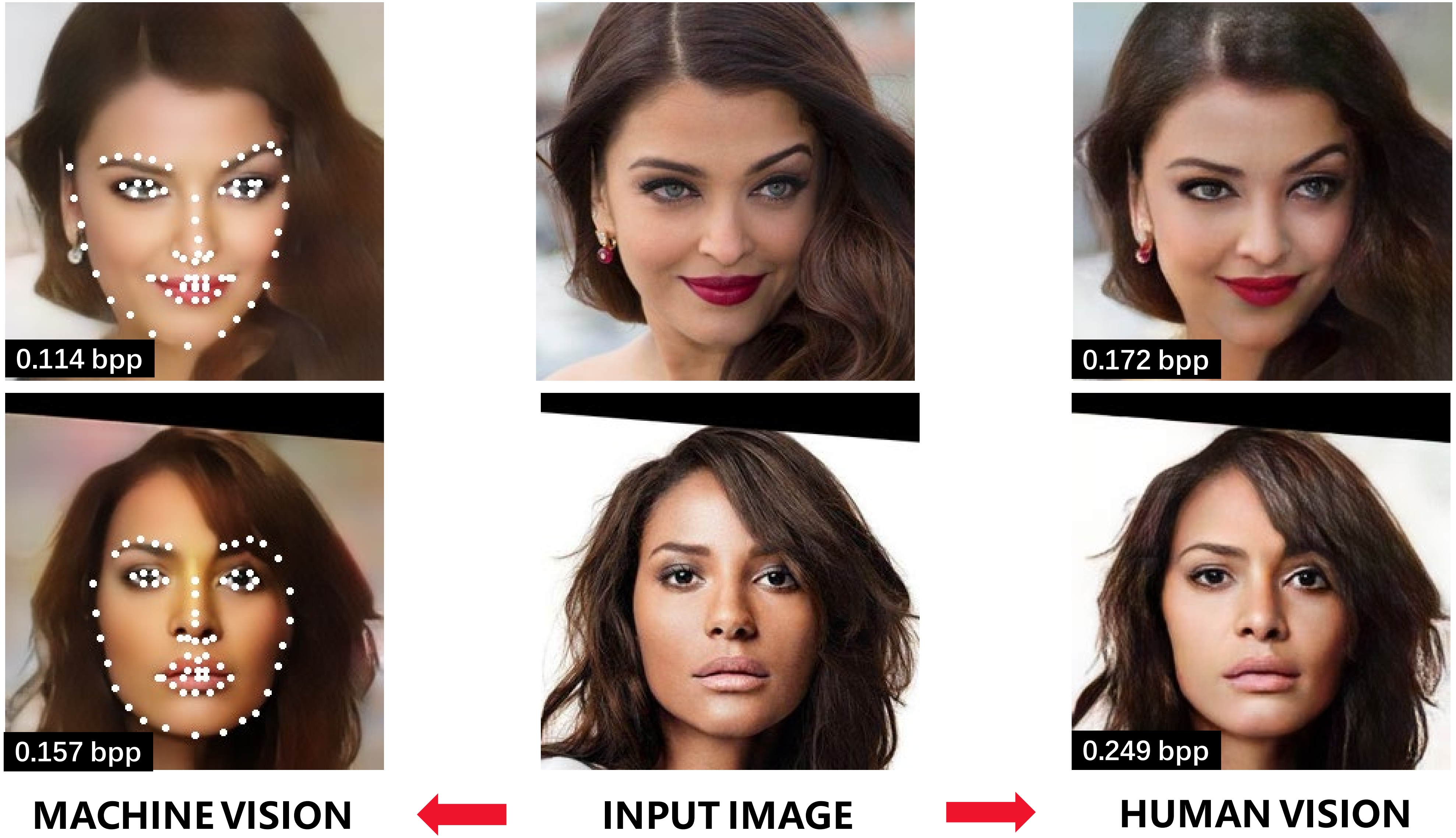
Measuring fidelity and Aesthetics.

MACHINE VISION

Evaluate facial landmark detection.

Measuring information preservation.

SCALABLE OUTPUT



HUMAN PERCEPTION

Input image

JPEG 0.266 bpp

HUMAN PERCEPTION

Input image

Ours 0.249 bpp

HUMAN PERCEPTION

Input image

JPEG 0.208 bpp

HUMAN PERCEPTION

Input image

Ours 0.198 bpp

HUMAN PERCEPTION

Input image

JPEG 0.178 bpp

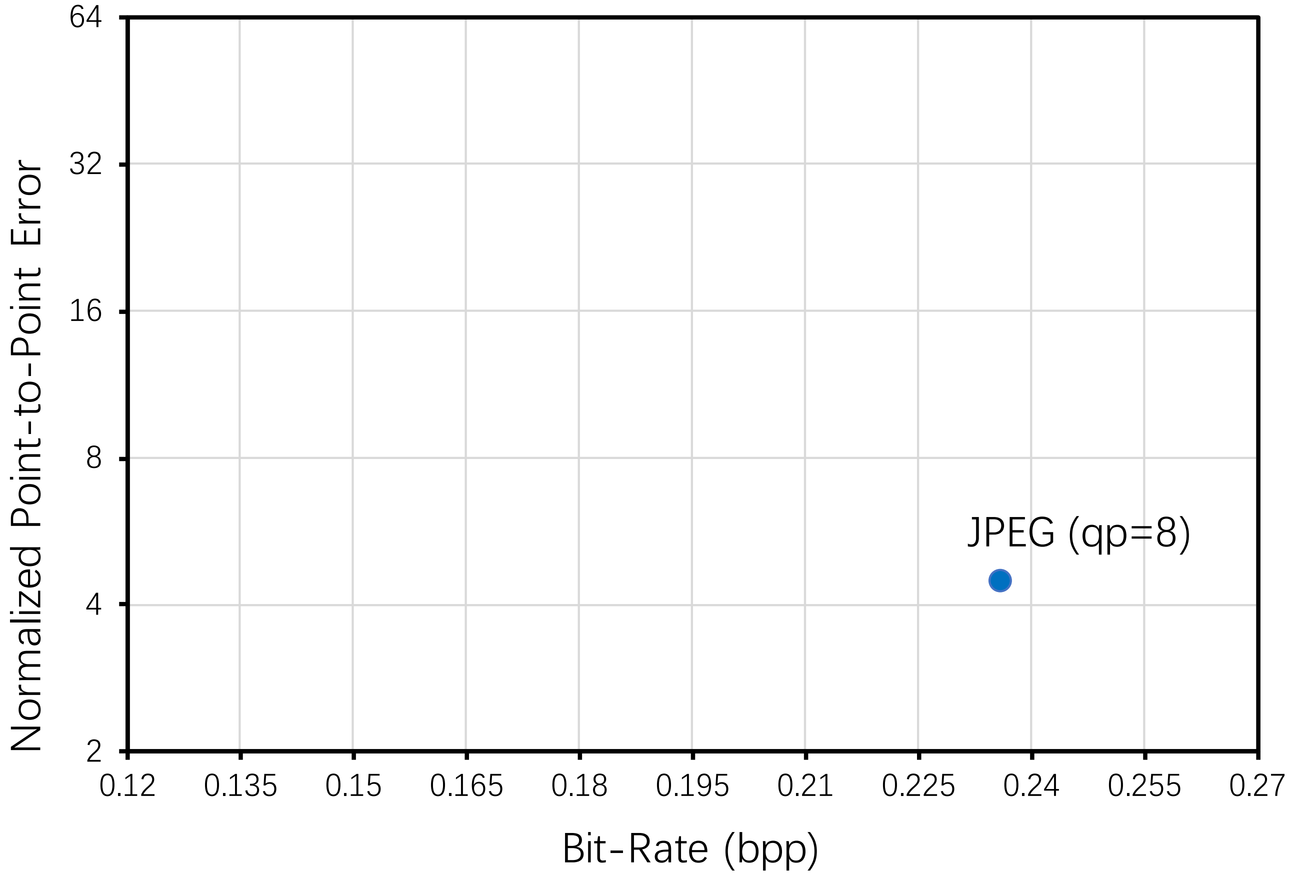
HUMAN PERCEPTION

Input image

Ours 0.177 bpp

MACHINE VISION

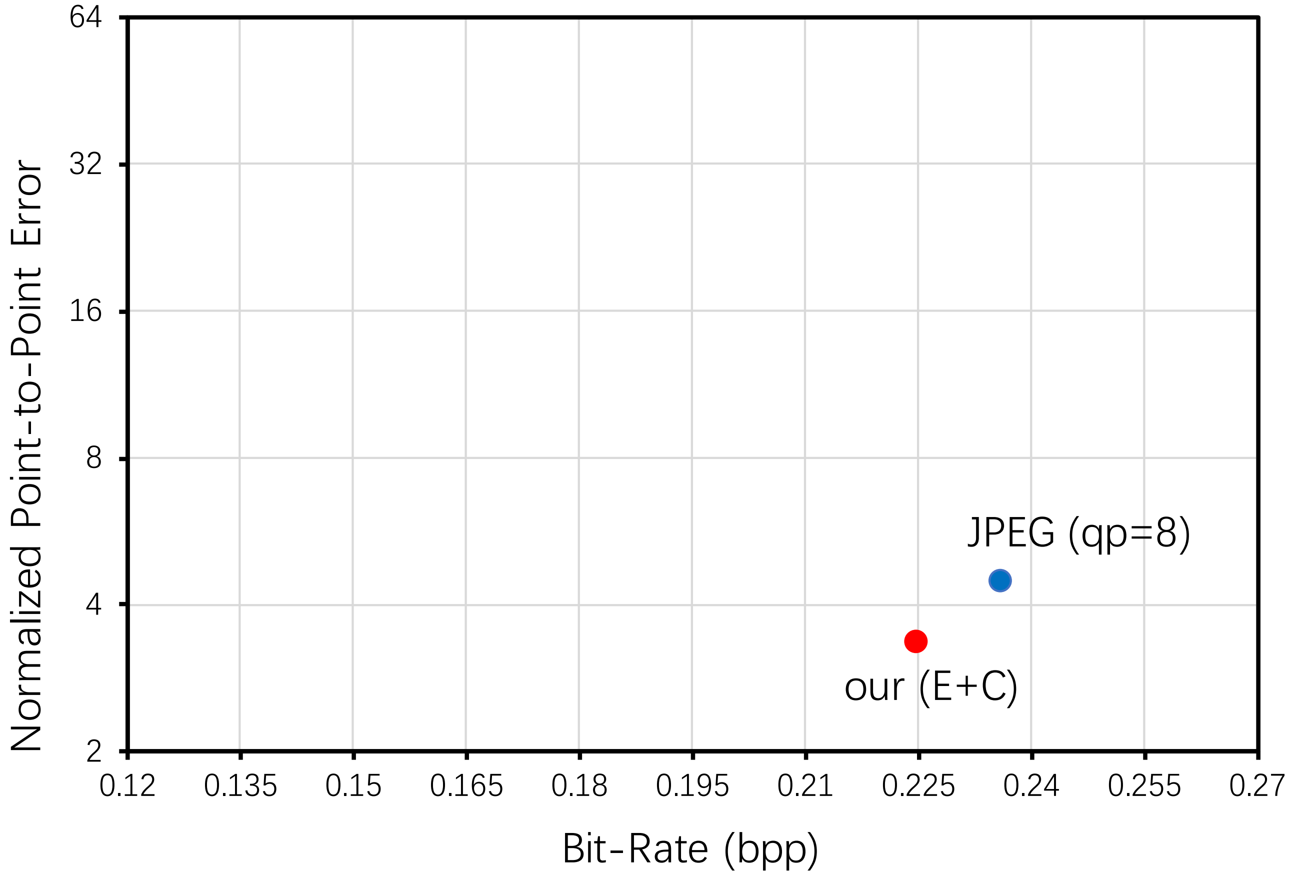
LANDMARK DETECTION ACCURACY



JPEG (qp=8)

MACHINE VISION

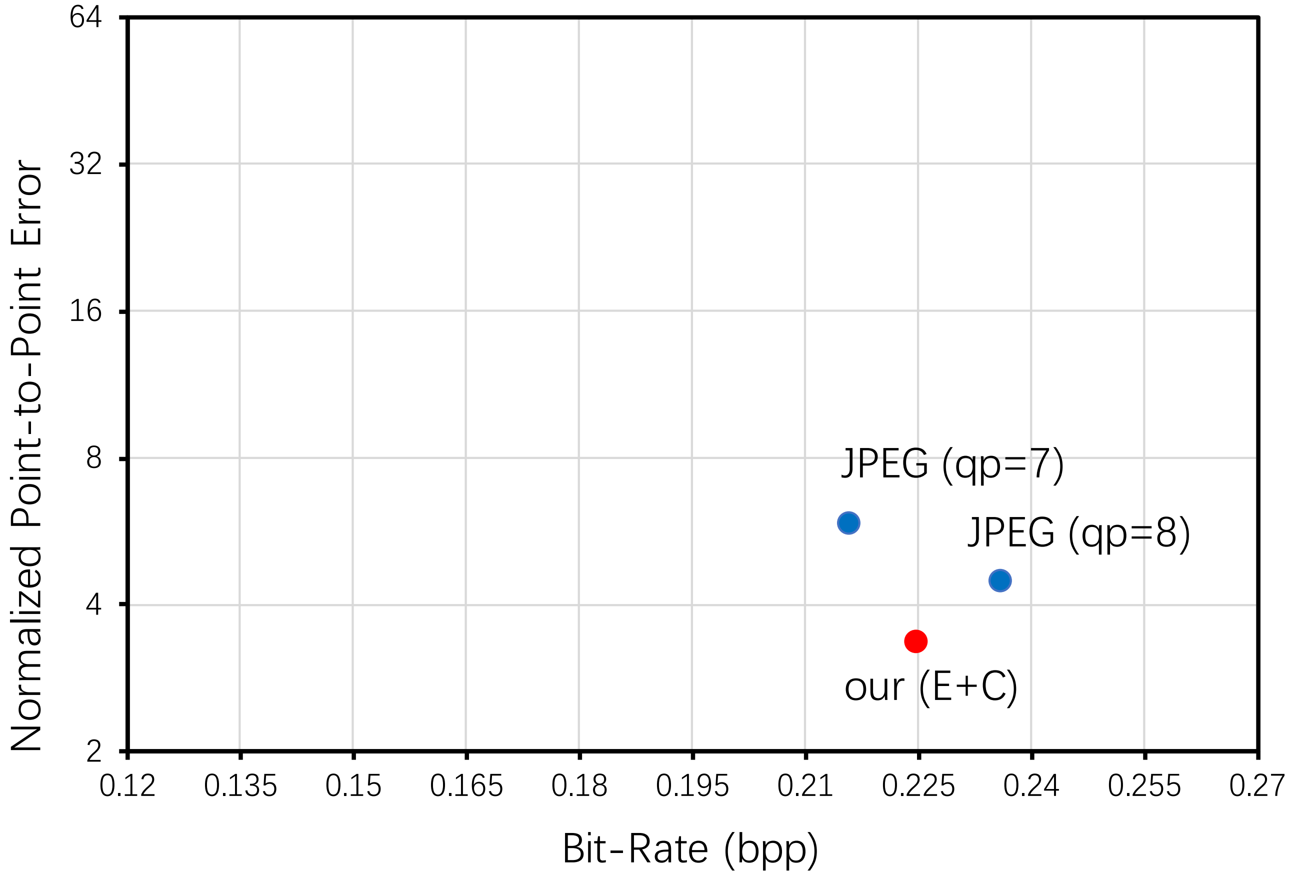
LANDMARK DETECTION ACCURACY



OUR (edge+color)

MACHINE VISION

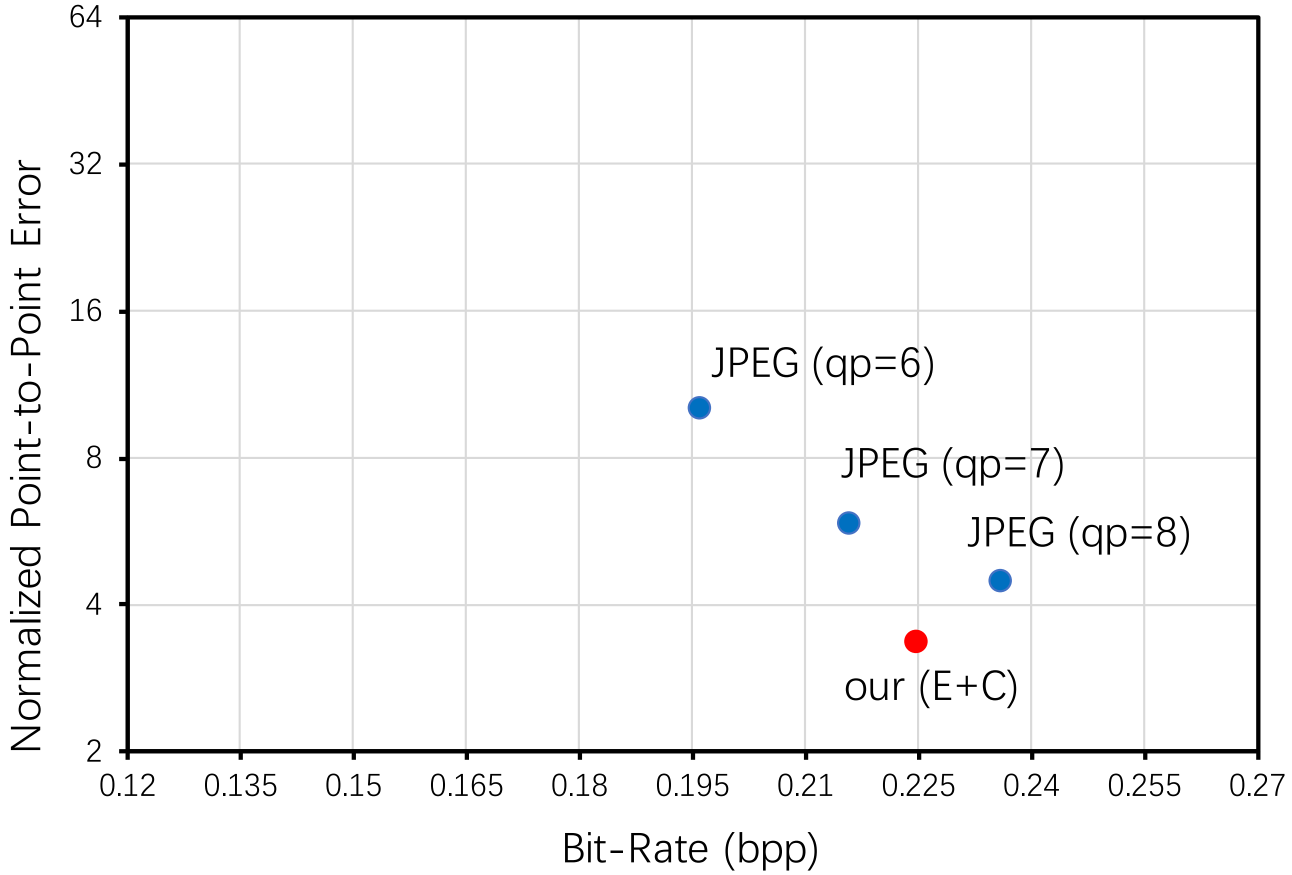
LANDMARK DETECTION ACCURACY



JPEG (qp=7)

MACHINE VISION

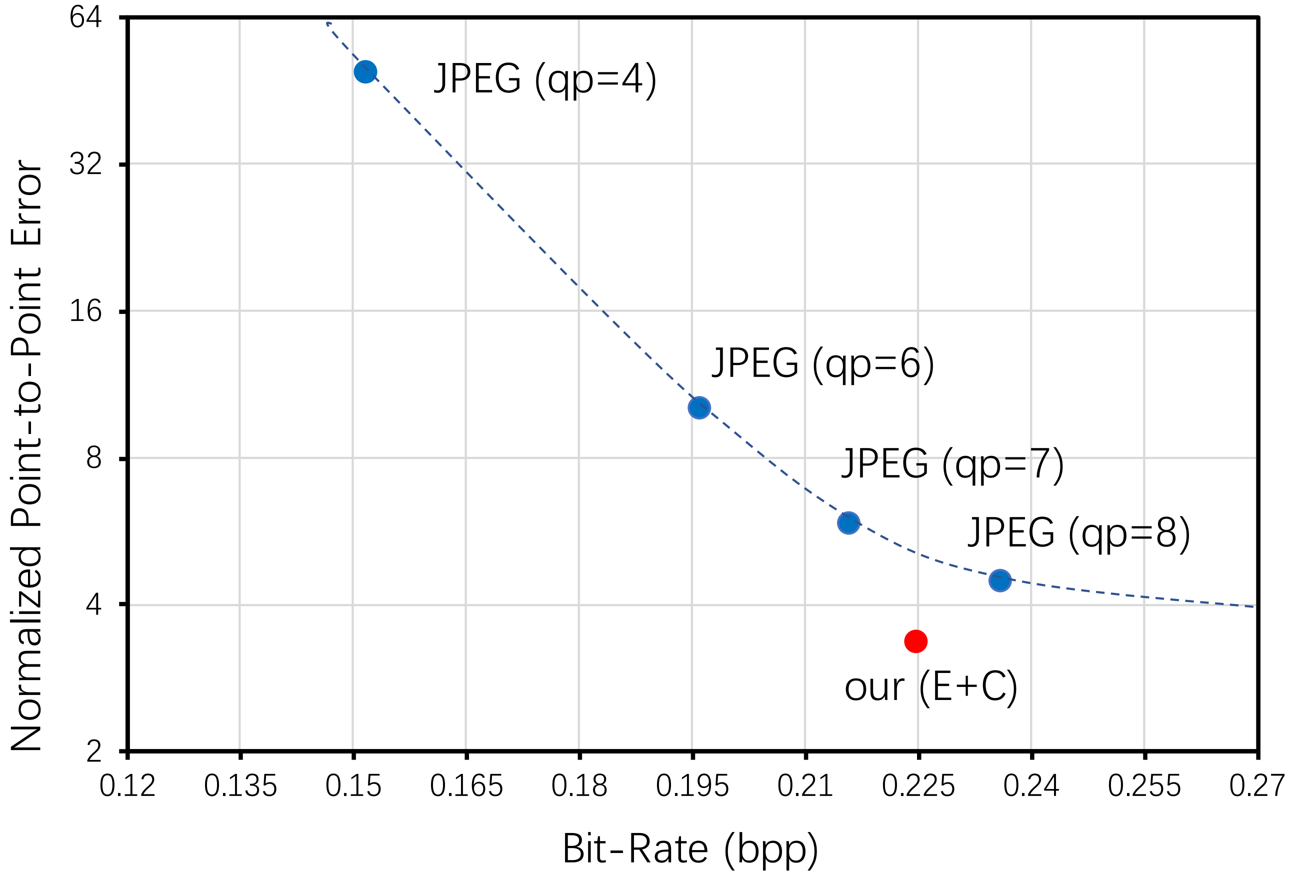
LANDMARK DETECTION ACCURACY



JPEG (qp=6)

MACHINE VISION

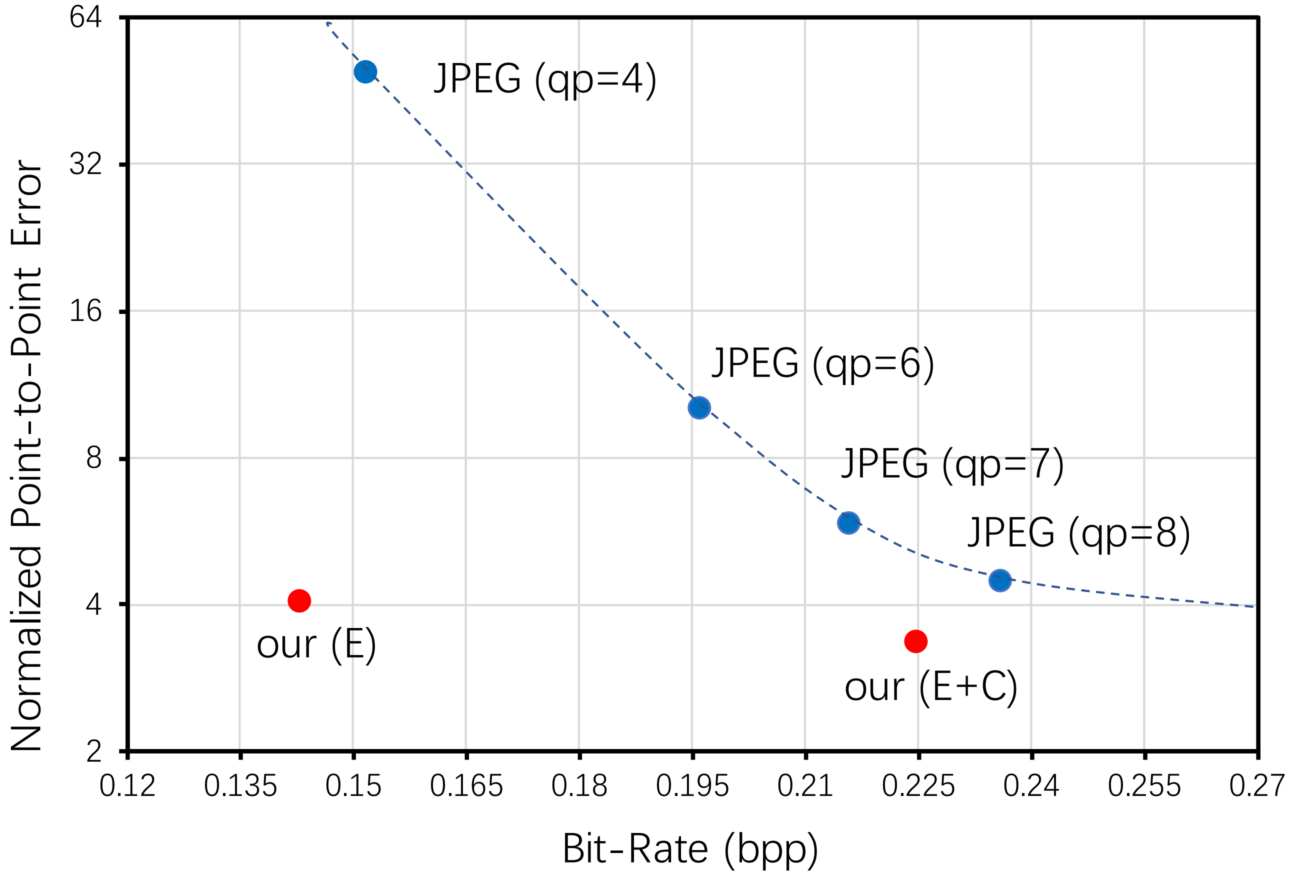
LANDMARK DETECTION ACCURACY



JPEG (qp=4)

MACHINE VISION

LANDMARK DETECTION ACCURACY

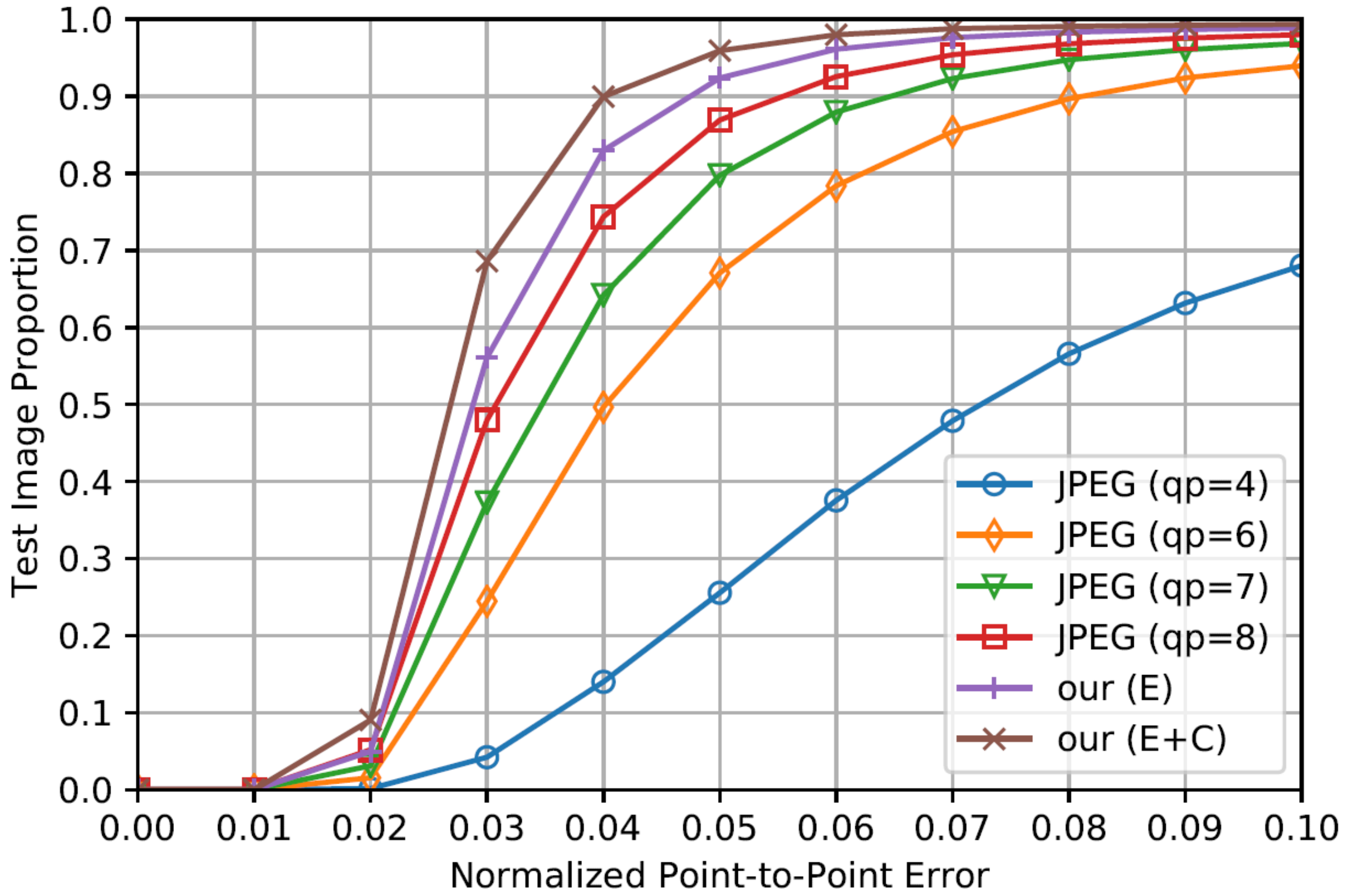


OUR (edge)

MACHINE VISION

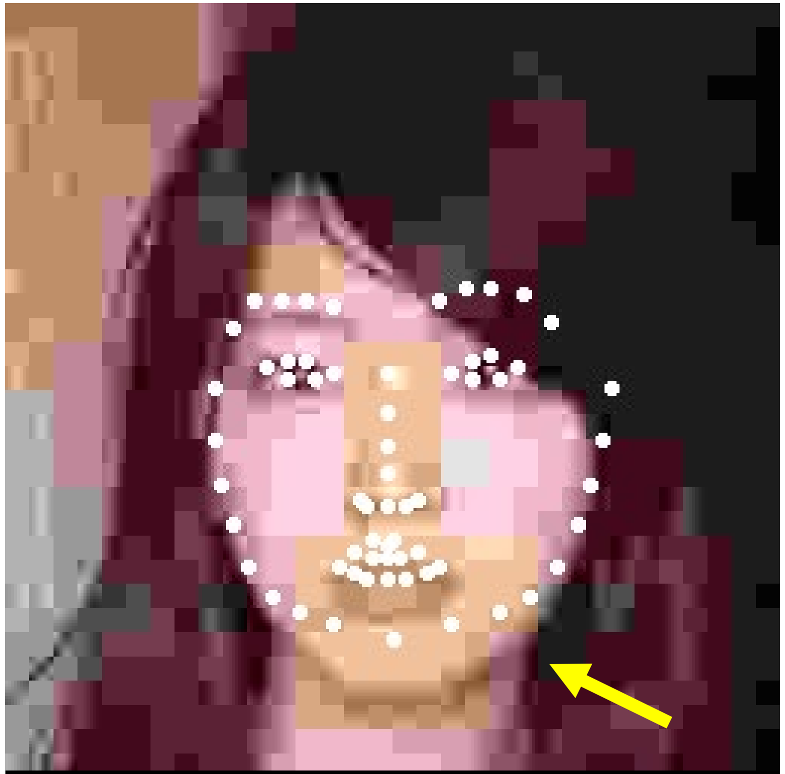
LANDMARK DETECTION ACCURACY

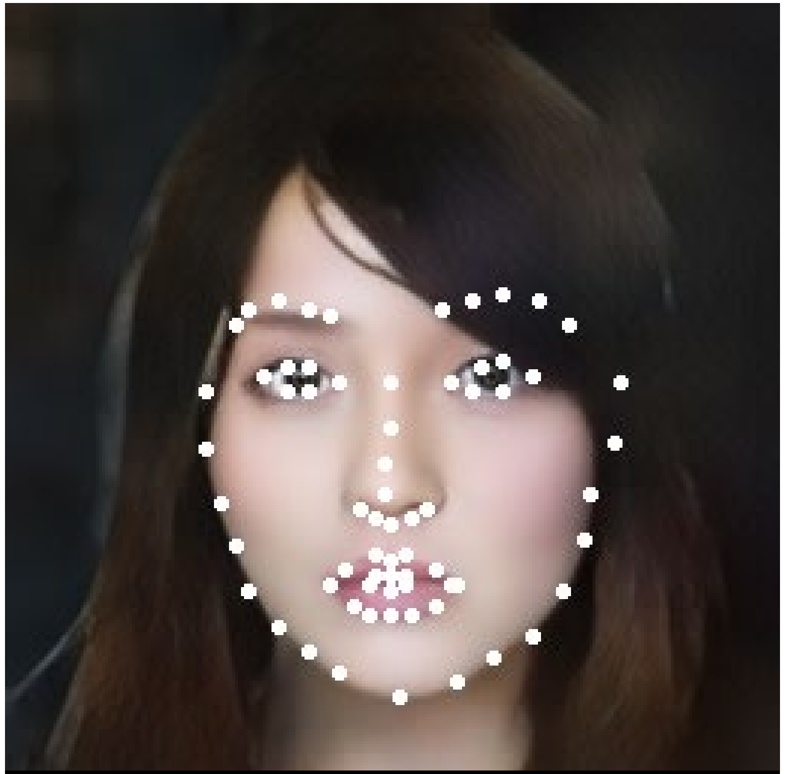
- Quantitatively evaluate the accuracy of facial landmark detection on the reconstructed images.
- Results show statistically improved accuracy at a lower bit-rate.
- While the basic layer maintain a high accuracy, the enhancing layer provide more fidelity.



MACHINE VISION

LANDMARK DETECTION RESULTS

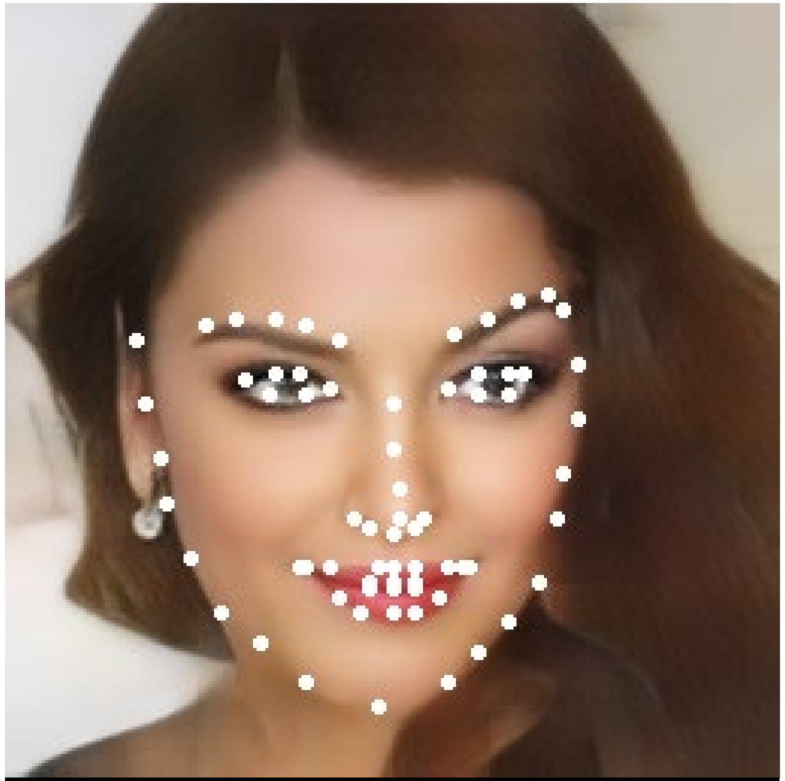

JPEG 0.131 bpp


Ours 0.115 bpp

MACHINE VISION

LANDMARK DETECTION RESULTS

JPEG 0.138 bpp

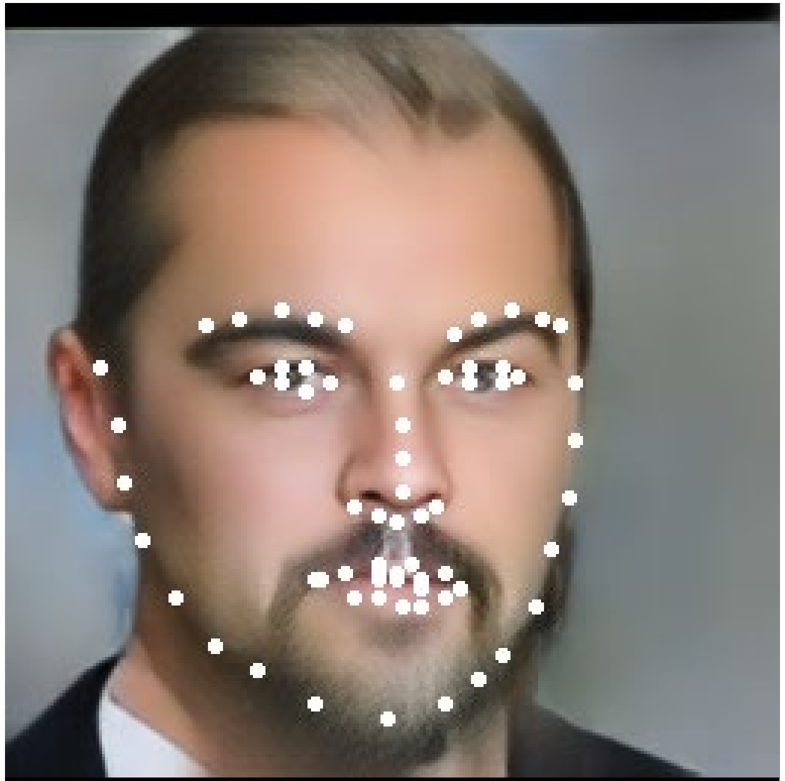


Ours 0.114 bpp

MACHINE VISION

LANDMARK DETECTION RESULTS

JPEG 0.145 bpp

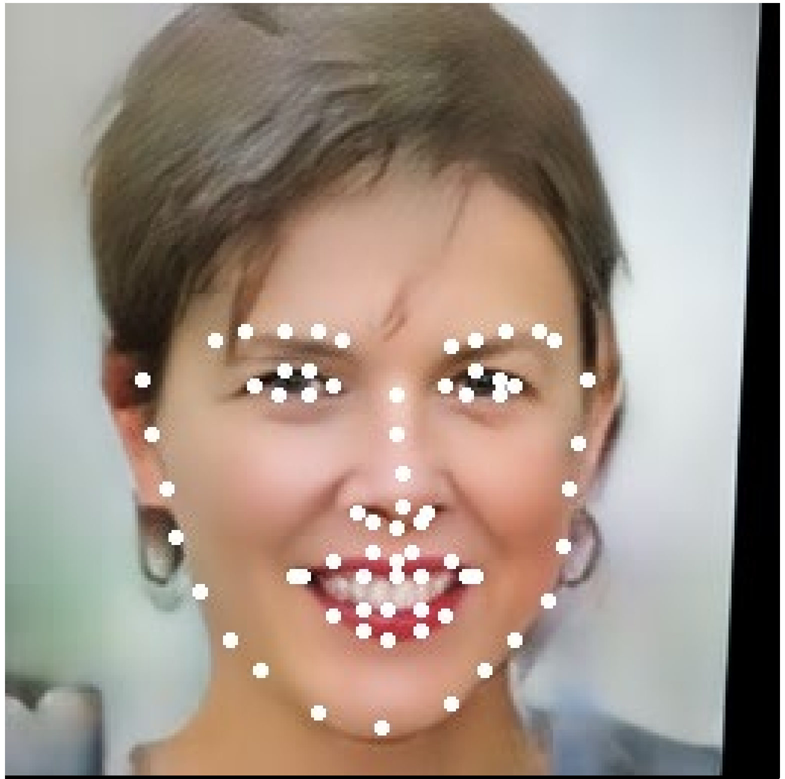


Ours 0.108 bpp

MACHINE VISION

LANDMARK DETECTION RESULTS

JPEG 0.158 bpp



Ours 0.143 bpp

CONCLUSION

APPROACH TO COLLABORATIVE CODING

- Edge + sparse pixels, vectorized representation
- Generative adversarial reconstruction
- Human-machine collaborative feature extraction

INFORMATION SCALABLE FRAMEWORK

- Base layer → Semantically accurate
- Enhanced layer → Visually faithful
- Efficient feature adaptation

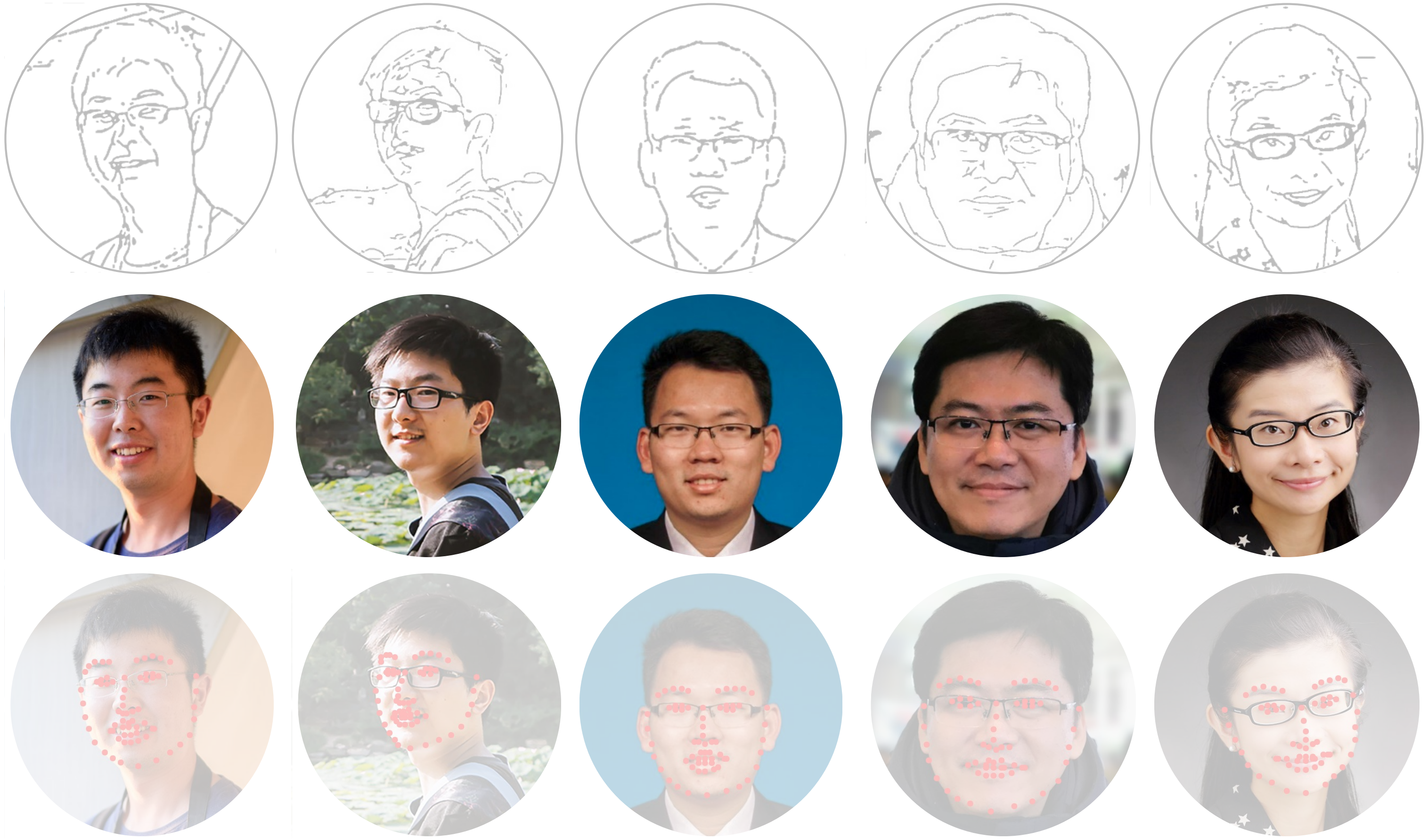
FUTURE DIRECTIONS

- Self-learned feature adaptation
- Multi-task collaborative inference
- Theoretical analysis on collaborative coding

PAPER ID 818

PROJECT

CODE



Thank You!