instituto de telecomunicações

TÉCNICO LISBOA

JPEG AI: The First International Standard for Image Coding Based on an End-to-End Learning-Based Approach

João Ascenso

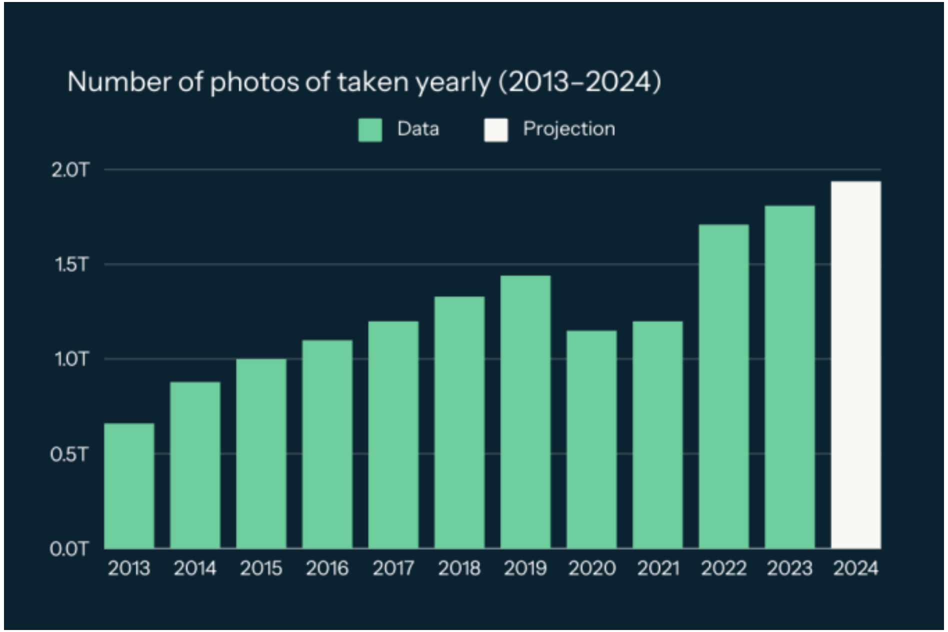
Instituto de Telecomunicações – Instituto Superior Técnico, University of Lisbon

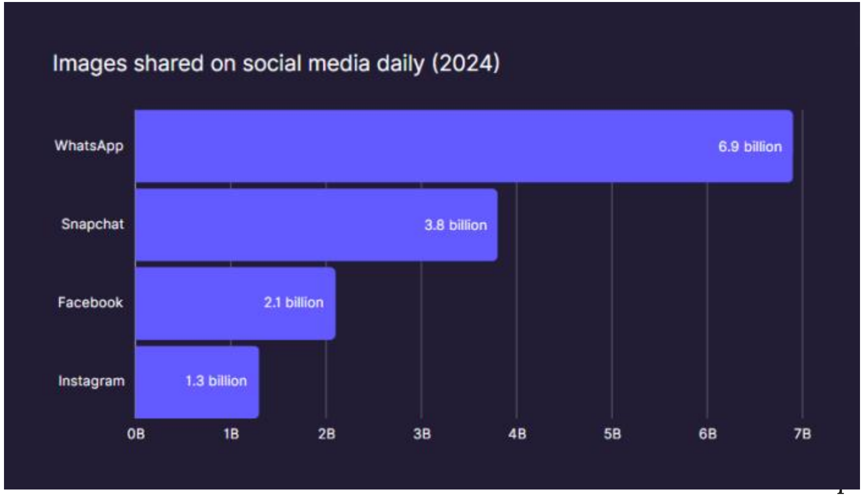
Outline

1. Context and Motivation
2. The JPEG AI Project
3. JPEG AI Verification Model
4. Performance Evaluation
5. Going Forward …

1 Context and Motivation

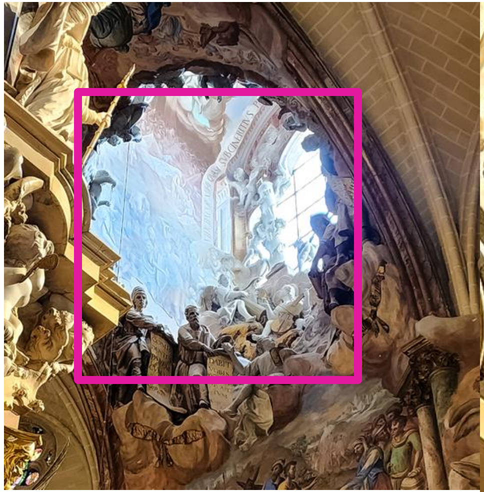
Many, Many Images….





Rich Ecosystem of Image Technologies

Image/Video Use Cases are Rather Different



Zooming in, go for details

Image Compression Landscape

JPEG

JPEG 2000

JPEG-XS

GIF

PNG

JPEG XL

AV1

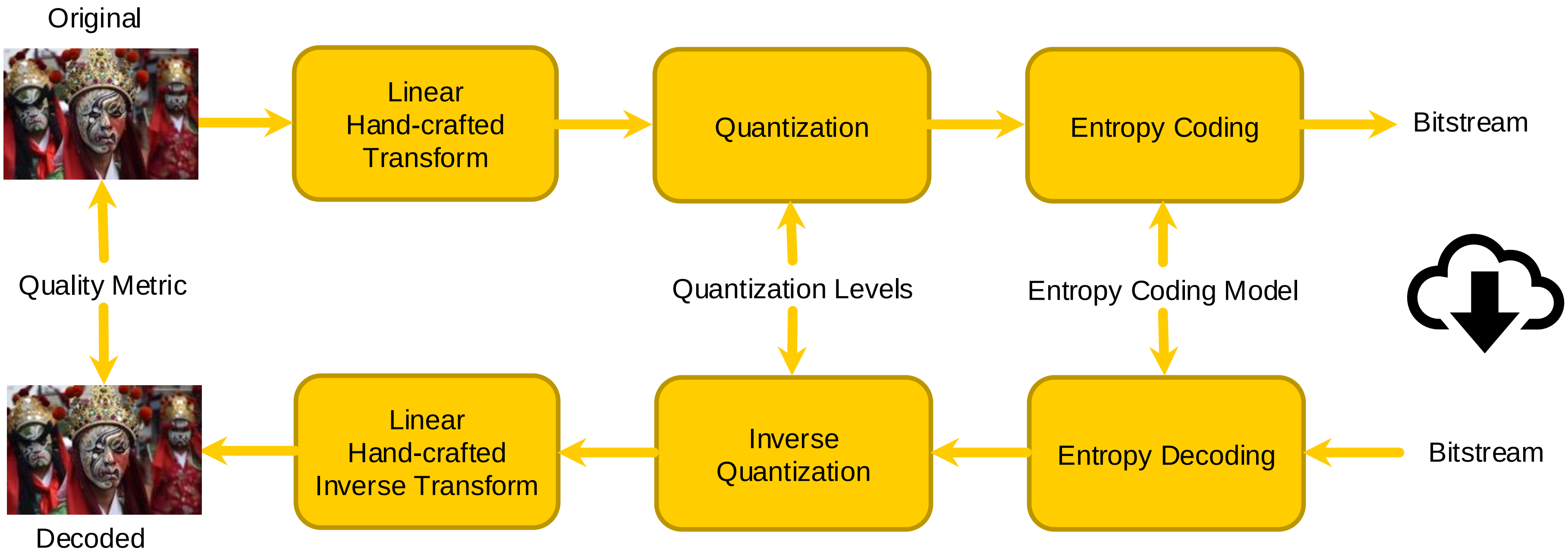
HEVC

webp

H.266 VVC

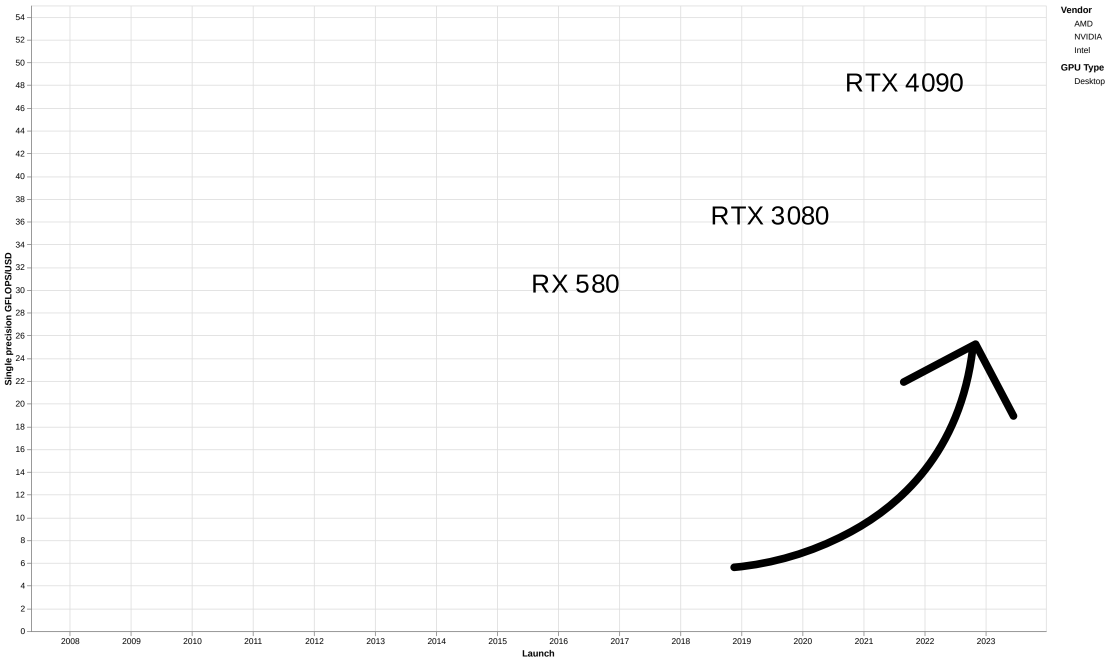
Classical Image Compression Pipeline

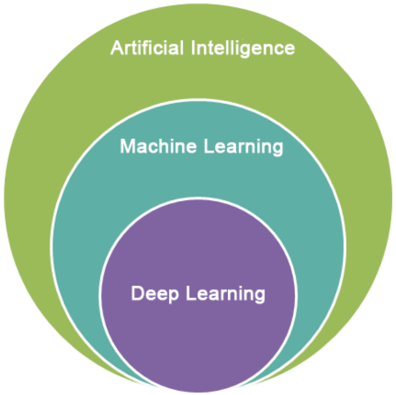
JPEG: simple, elegant, large ecosystem, interpretable, …



Deep Learning Explosion !

Giga FLoating-point Operations Per Second that you can buy with 1 USD





1. Big Data

- Larger Datasets
- Easier Collection & Storage

2. Hardware

- Graphics Processing Units (GPUs)
- Massively Parallelizable

3. Software

- Improved Techniques
- New Models
- Toolboxes

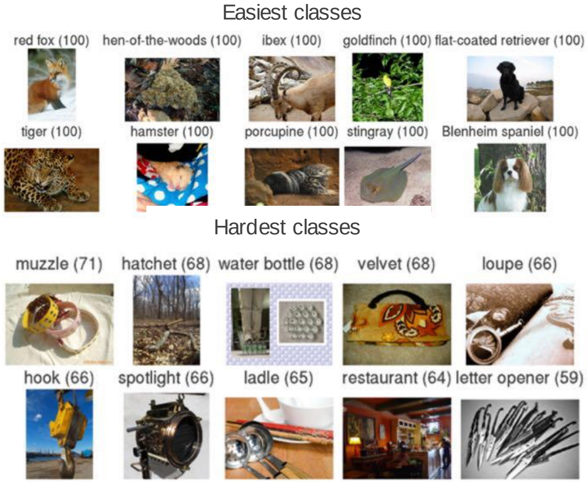
TensorFlow

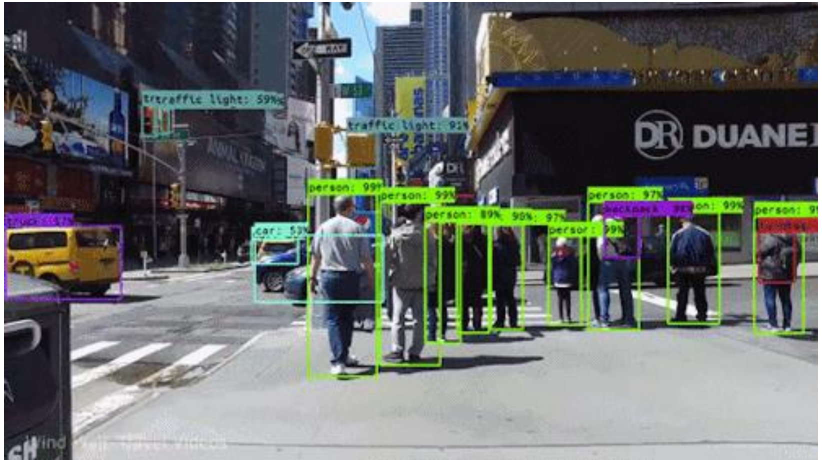
IMPROVE 2025

Deep Learning Achievements: Computer Vision

- Extremely successful in computer vision tasks:
  - ✓ Image classification, object detection, semantic segmentation, …
  - ✓ Face recognition, image generation, video understanding, …

Image classification



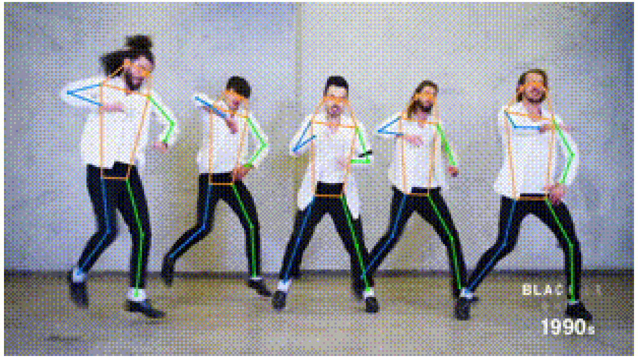


Deep Learning Achievements: Image Processing

- Extremely successful in image processing tasks:
  - ✓ Denoising, super-resolution, inpainting, style transfer, segmentation, …
  - ✓ Many other image restoration tasks (dehazing, deraining, etc.), …

Mordvinsev et al, 2015
Gatys et al, 2016

And Many Others ….



Visual Coding vs Neural Networks

- Learning-based image compression
  - ✓ Non-linear transformations, entropy coding models, etc.
- Learning-based video compression
  - ✓ Optical flow, motion compensation, multi-frame fusion, etc.
- Models for typical image/video compression modules
  - ✓ Intra-prediction, in/out loop-filtering, entire encoder, etc.
- Learning-based point cloud compression
  - ✓ Geometry and attribute compression methods, etc.
- Learning-based light-field compression
  - ✓ Stereoscopic and multi-view representations, NeRF, etc.
- Neural networks models and activations compression
  - ✓ Enabling the efficient transmission of large models (or activations)

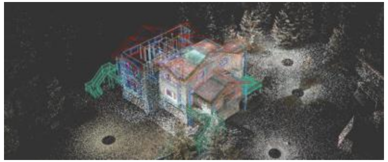
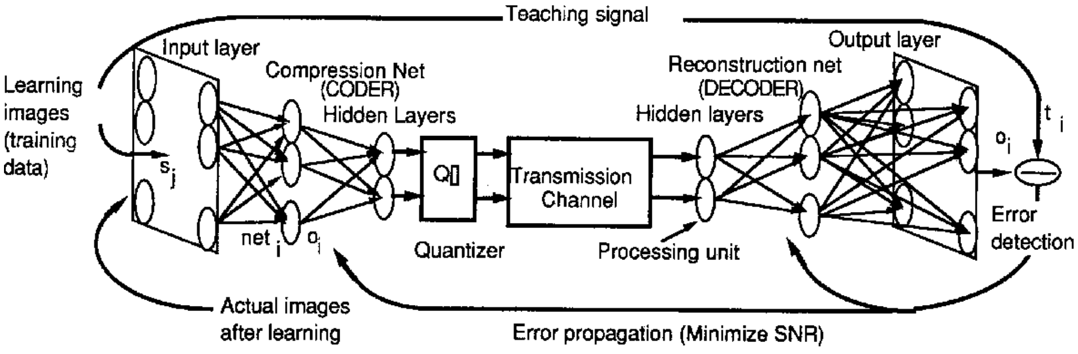


Image Compression with Neural Networks

- Very ~~recent~~ and promising field
  - ✓ N. Sonehara, M. Kawato, S. Miyake, K. Nakane, Image data compression using neural network model, Proceedings of the International Joint Conference On Neural Networks, Washington DC, 1989, pp. 35–41.
  - ✓ G.L. Sicurana, G. Ramponi, Artificial neural network for image compression, Electron. Lett. 26, (7) (1990) 477–479.

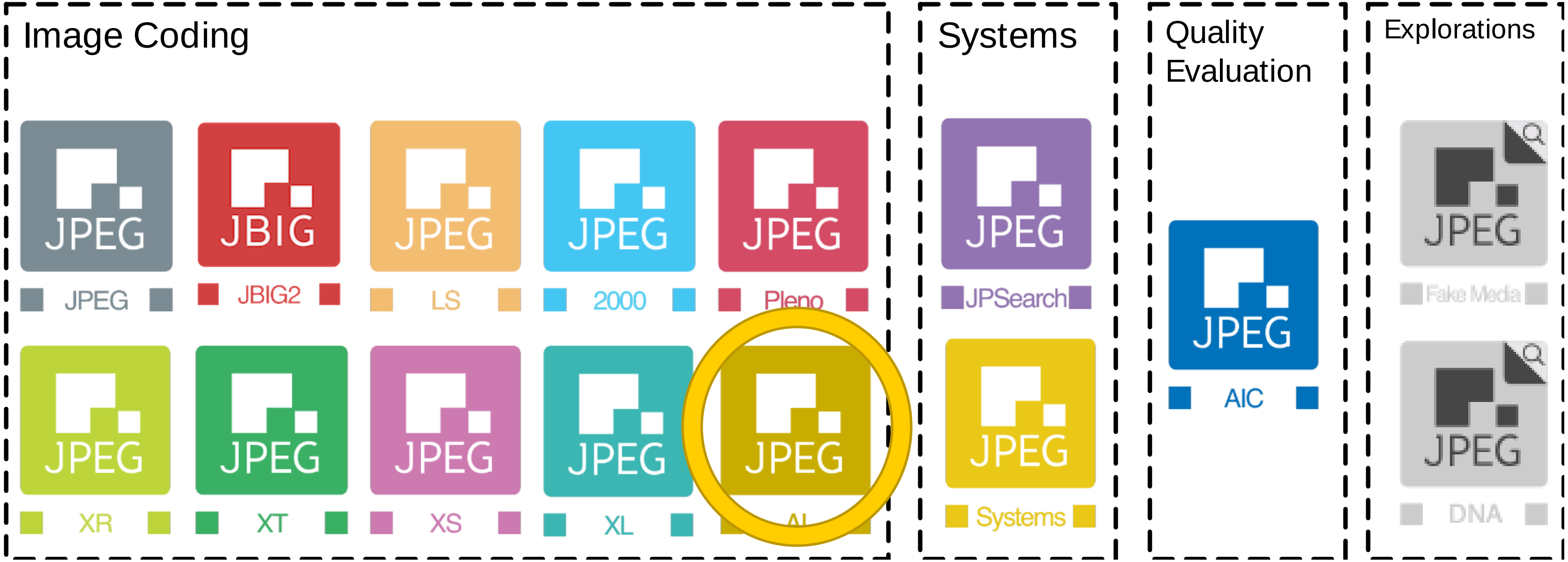


As old as JPEG !!!

2

The JPEG AI Project

JPEG Family of Standards

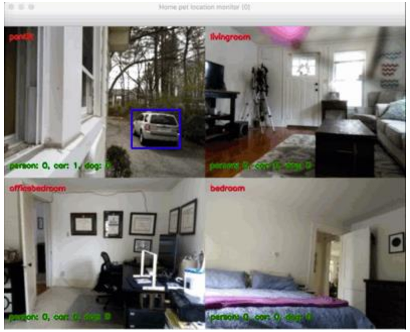


JPEG AI Achievements

- JPEG AI Project (ISO/IEC 6048) within the JPEG standardization group develops and standardize learning-based image compression
  - ✓ Joint standardization effort between SC29/WG1 and ITU-T SG16
  - ✓ Active since 2019, International Standard in publication phase
- Call for Evidence combined with MMSP Workshop Grand Challenge
  - ✓ 6 codecs submitted (out of 8 registered)
- Some relevant public documents:
  - ✓ White Paper on JPEG AI Scope and Framework
  - ✓ JPEG AI Uses Cases and Requirements
  - ✓ JPEG AI Training and Test Conditions
  - ✓ JPEG AI Call for Proposals
  - ✓ And many more …
- Check for more information: https://jpeg.org/jpegai/

JPEG AI Use Cases

- Cloud storage
- Visual surveillance
- Autonomous vehicles and devices
- Image collection storage and management
- Live monitoring of visual data
- Media distribution

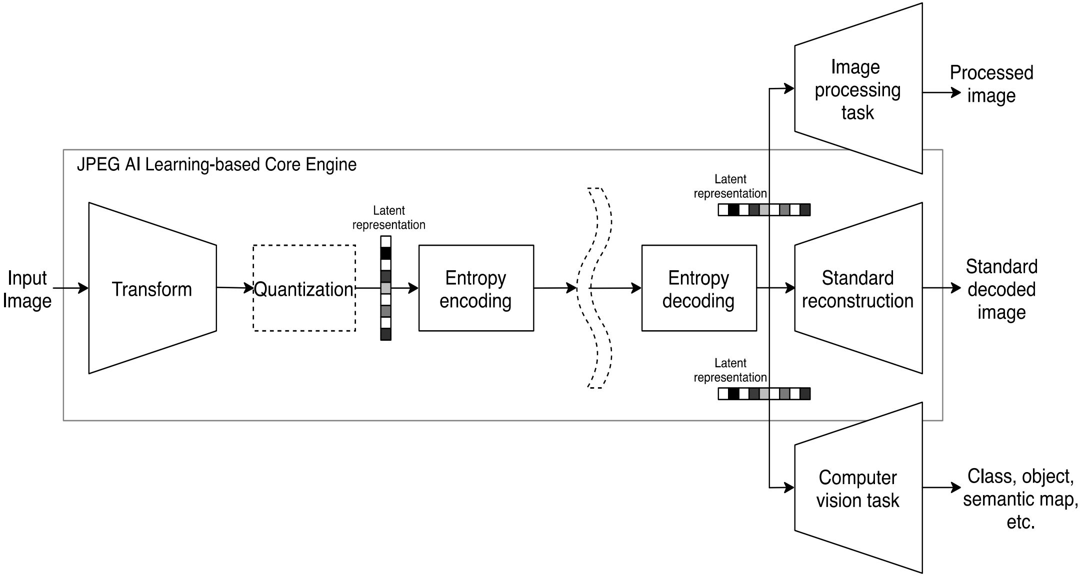


JPEG AI Scope

The JPEG AI scope is the creation of a learning-based image coding standard offering a single-stream, compact, compressed domain representation, targeting both human visualization, with significant compression efficiency improvement over image coding standards in common use at equivalent subjective quality, as well as effective performance for image processing and computer vision tasks, with the goal of supporting a royalty-free baseline

Image processing tasks	Computer vision tasks
Super-resolution	Image retrieval and classification
Low-light enhancement	Object detection and recognition
Color correction	semantic segmentation
Exposure compensation	Event detection and action recognition
Inpainting	Face detection and recognition

JPEG AI Framework



- Advantages for image processing and computer vision task:
  - ✓ *Single-stream representation:* same compressed stream is also useful for decoding
  - ✓ *Energy efficient:* reduces the resources needed to perform these tasks
  - ✓ *High accuracy:* allows performing these tasks using features extracted from the original instead of the lossy decoded images

Application-driven Requirements

- High coding efficiency is important for many applications such as cloud storage or media distribution

- Content understanding is vital for many applications such as visual surveillance, autonomous vehicles, image collection management, etc
  - ✓ Objects may need to be recognized
  - ✓ Images may need to be classified for organization purposes
  - ✓ Actions or events may need to be recognized

- Content is not consumed by humans in the same way as the original reference in many applications such as in media distribution
  - ✓ Noise can be reduced
  - ✓ Resolution can be increased
  - ✓ Colors can be corrected

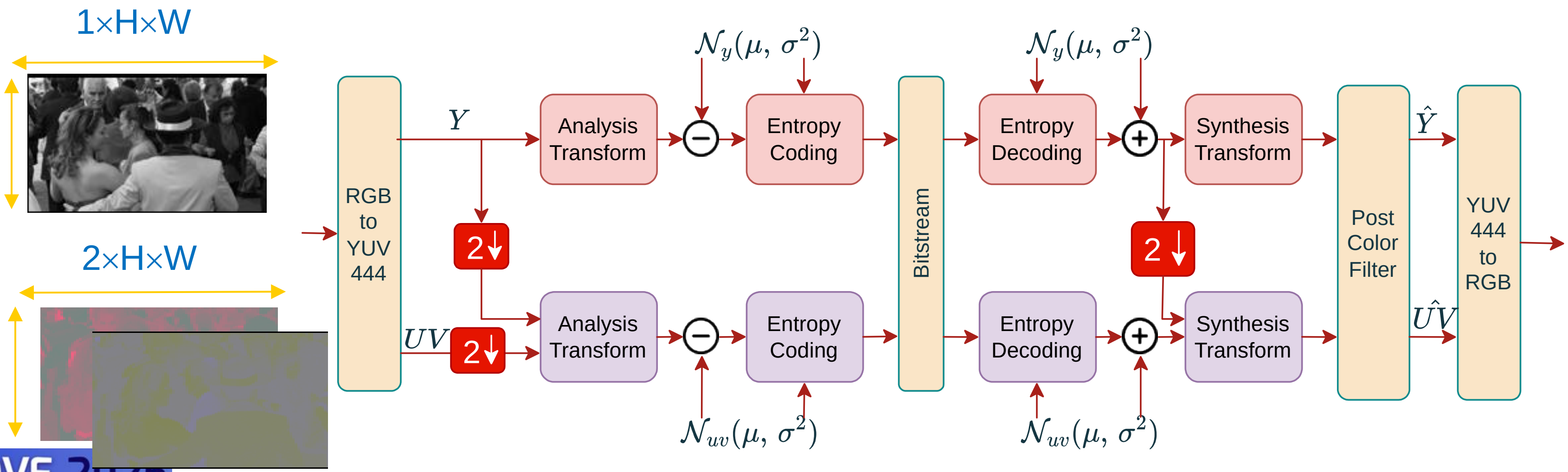
3

JPEG AI Verification Model

IMPROVE 2025

JPEG AI VM High Level Architecture

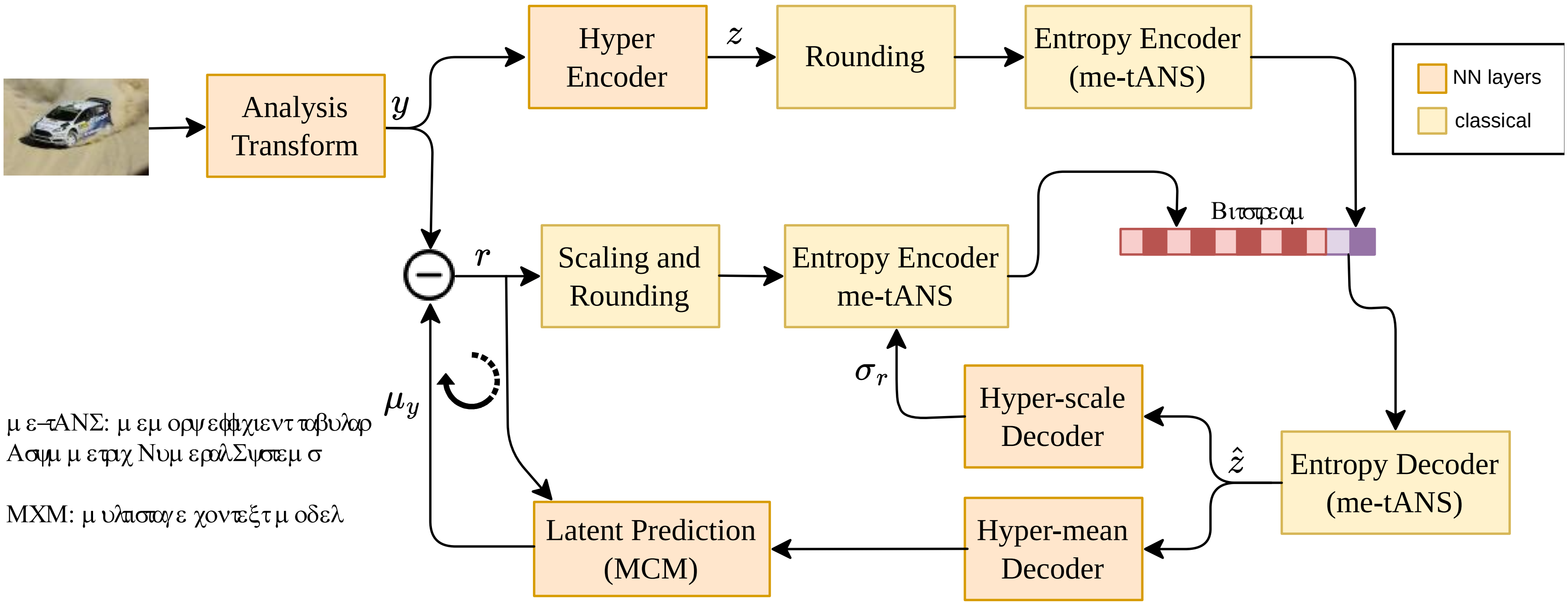
- New architecture never proposed before
  - ✓ Works with YUV BT.709 colour space and supports 4:4:4 and 4:2:0
  - ✓ Exploits spatial correlation with the analysis and synthesis transforms
  - ✓ Probabilistic latent model is obtained from side information (hyper-prior)
- Two encoding pipelines are present, one for luma and another for chroma
  - ✓ Chroma pipeline encodes UV in half of the resolution of Y (and has less depth)
  - ✓ Independent pipelines using networks with same architecture, but different number of channels



JPEG AI Key Characteristics

- Probability table for entropy coding is modelled with $\mathbb{N}(0,\sigma)$ for every latent element
- Latents are predicted and only the residual is coded and transmitted
  - ✓ Exploits spatial correlation at the latent domain
- Entropy decoding is decoupled of latent prediction and reconstruction
  - ✓ Entropy decoding of a latent doesn't depend on previously decoded latents
- Hyper scale decoder
  - ✓ Provides estimation of the variance of the entropy coding model distribution
- Hyper "mean" decoder
  - ✓ Provides estimation of the mean (explicit prediction) of the latent

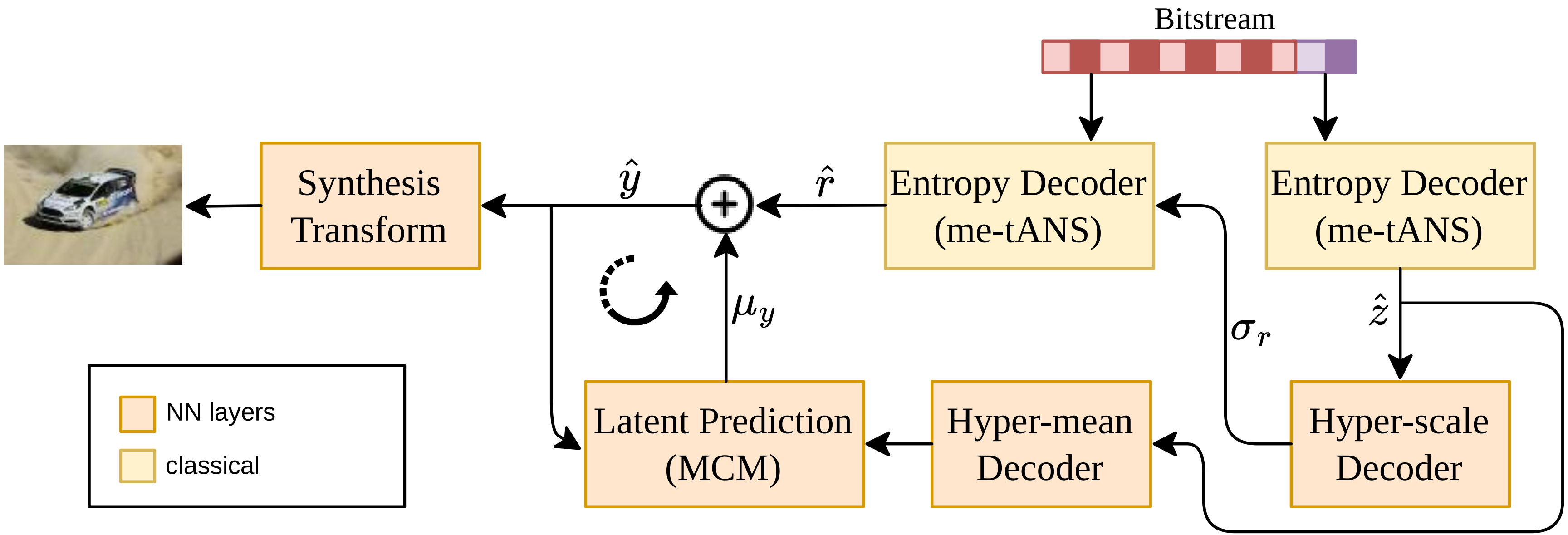
JPEG AI VM Encoder Architecture



μ ε–τΑΝΣ: μ εμ ορψ εφφιχιεντ ταβυλαρ
Ασψμ μ ετριχ Νυμ εραλ Σψστεμ σ

ΜΧΜ: μ υλτισταγε χοντεξτ μ οδελ

JPEG AI VM Decoder Architecture

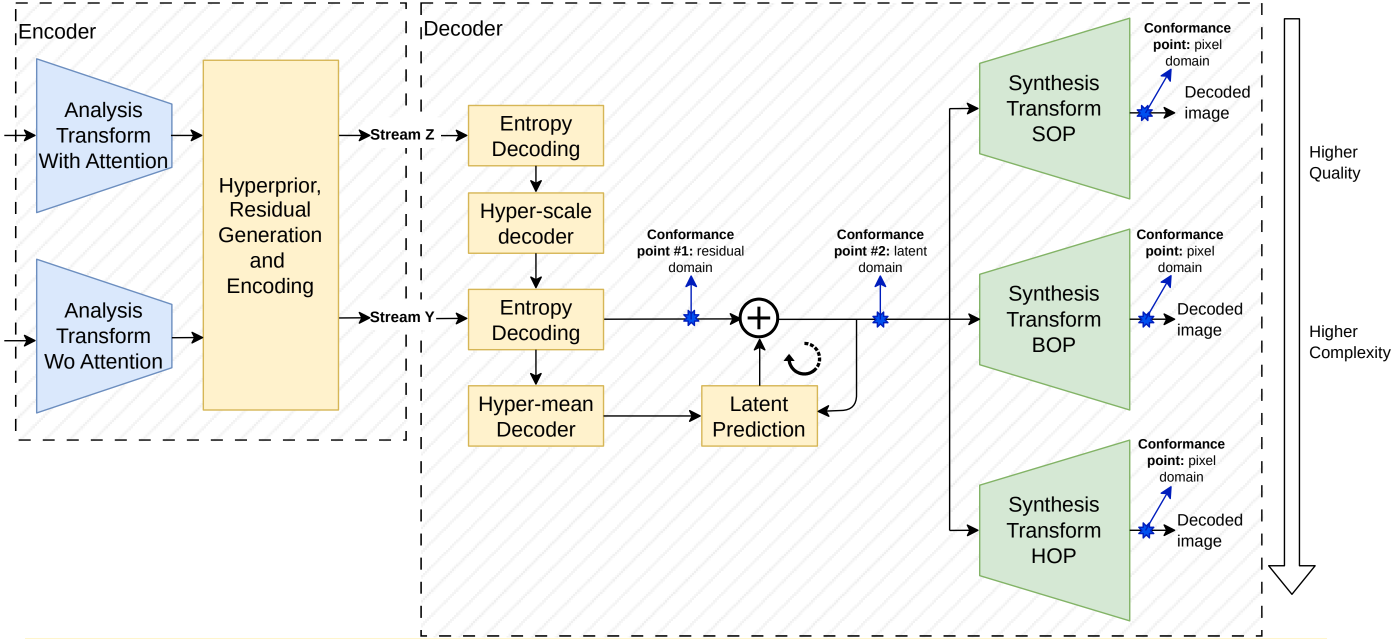


Addressing Complexity Issues

- Three operating points are supported:
  - ✓ CPU operating point targeting legacy devices
  - ✓ Base operating point targeting mobile devices
  - ✓ High operating point for more hardware-capable devices with powerful GPUs and no energy constraints
- Base operating point should provide 10-15% compression efficiency gains over VVC Intra with approx. 22 kMAC/px
- High operating point should provide 25-30% compression efficiency gains over VVC Intra with approx. 220 kMAC/pxl

JPEG AI Multi-branch Decoding

Receiver can support just one decoder (operating point) to decode any stream

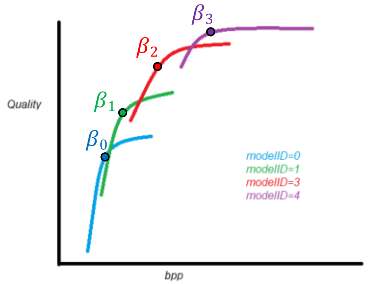


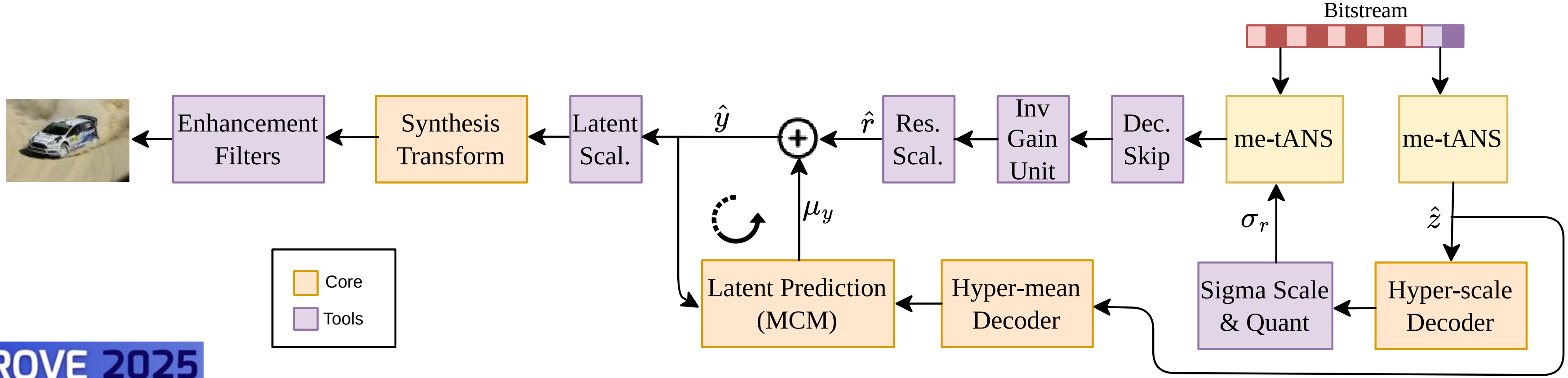
JPEG AI VM supports

2 Encoder × 3 Decoder = 6 possible combinations compatible to each other

JPEG AI has Several Flag-enabled Tools

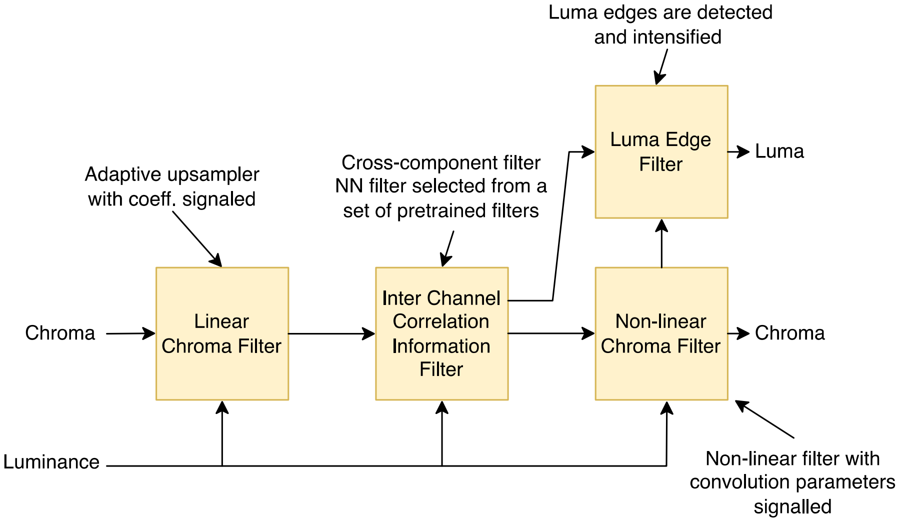
- Skip mode allows skip writing/parsing from the bitstream residual latent elements which can be identified by encoder and decoder to be zero
- Variable rate coding with Gain Units
  - ✓ Model parameters defined by ModelID
  - ✓ “Gain” factor for residual & variance defined by $\Delta\boldsymbol{\beta}$ (signalled)
- Residual and the standard deviation parameter scaling
- Enhancement filters increase mostly the chroma quality





Tool Example: Enhancement Filter Technologies

- Enhancement filters bring 26% gain in Chroma PSNR
- Linear chroma filter and non-linear chroma filter use signalled parameters and perform upsampling/color correction
- Inter channel correlation information filter provides enhancement of colour information exploiting correlation with luminance
- Luma edge filters adaptively enhances (scale) edges to improve decoded quality



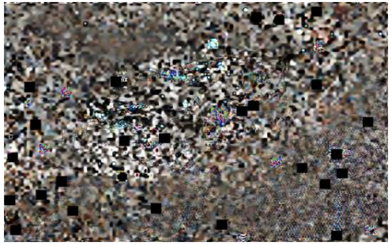
Device Reproducibility

Due to the use of floating-point arithmetic and different orders for the operations the result depends on platform heavily.

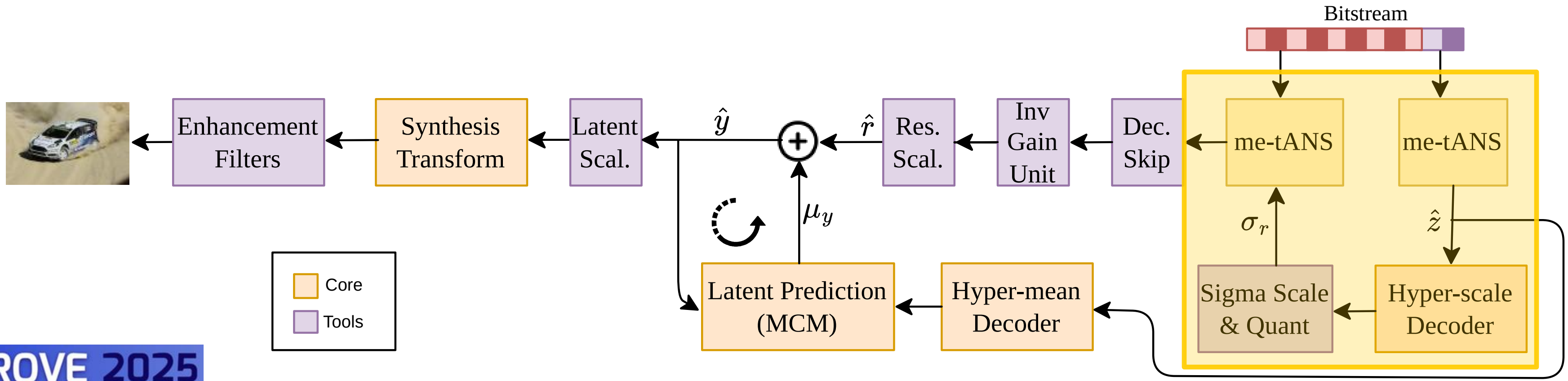
Leads to wrong interpretation of the parsed symbols in arithmetic coder

How does effect look like?

Encoded and decoded on same device



Encoded and decoded on different devices



Hyper Scale Decoder

Bit-exact behavior **in entropy part must be guaranteed!**

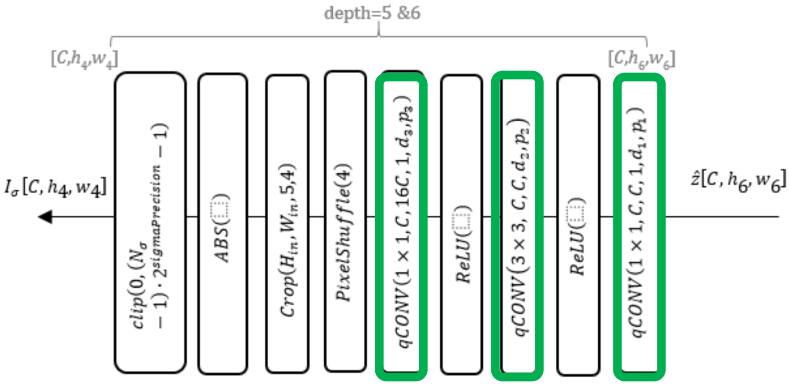
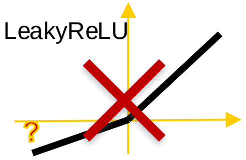
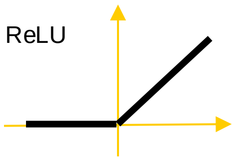


Figure 10.3-1 - Hyper Scale Decoder





convolution layer *CONV*

...

$$out[c_{out}, i, j] = bias[c_{out}] + \sum_{c_{in}=0}^{C_{in}} weigth[c_{in}, c_{out}] \star input[c_{in}, s \cdot i, s \cdot j];$$

$$i = 0, \dots, h_{out} - 1; j = 0, \dots, w_{out} - 1; c_{out} = 0, \dots, C_{out} - 1$$

where "⋆" is 2D **cross-correlation operator** with kernel size $K_{ver} \times K_{hor}$

....

quantized convolution layer *qCONV*

... three-steps operation:

$$temp[c_{in}, i, j] = clip(-d, d-1, input[c_{in}, i, j]),$$
$$i = 0, \dots, h_{in} - 1; j = 0, \dots, w_{in} - 1; c_{in} = 0, \dots, C_{in} - 1;$$

$$R[c_{out}, i, j] = bias[c_{out}] + \sum_{c_{in}=0}^{C_{in}-1} weigth[c_{in}, c_{out}] \star temp[c_{in}, s \cdot i, s \cdot j];$$

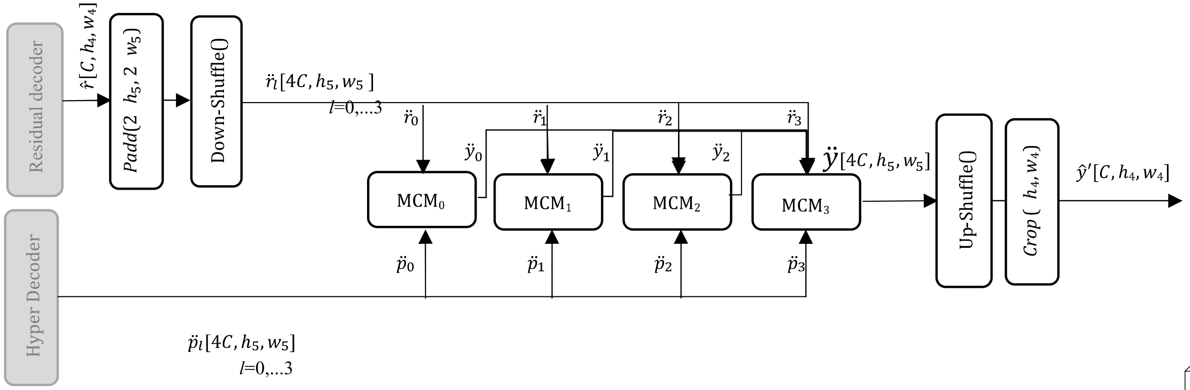
where "⋆" is 2D **cross-correlation operator** with kernel size $K_{ver} \times K_{hor}$.

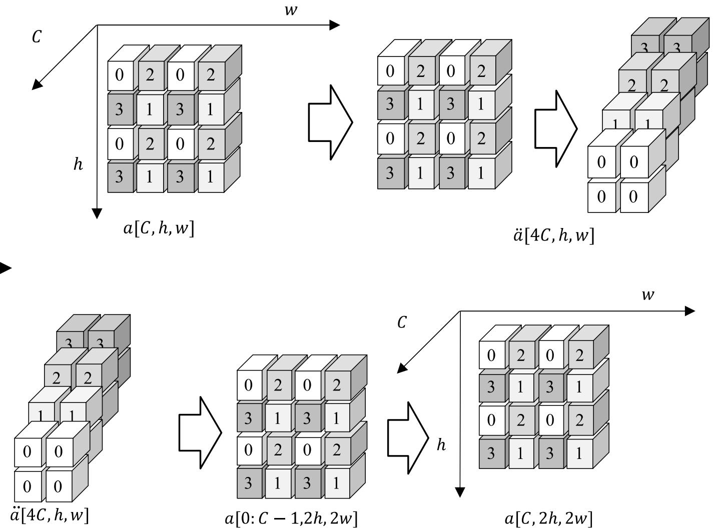
$$out[c_{out}, i, j] = (R[c_{out}, i, j]) \gg p[c_{out}];$$
$$i = 0, \dots, h_{out} - 1; j = 0, \dots, w_{out} - 1; c_{out} = 0, \dots, C_{out} - 1.$$

The tensor $weigth$ of shape $[C_{in}, C_{out}, K_{ver}, K_{hor}]$ contains learnable ***8-bit integer*** weights, the tensor $bias$ of shape $[C_{out},]$ contains learnable ***31-bit integer*** biases. All parameters $weigth$ and $bias$ are part of learnable quantized model.
The combination of clipping value *d*, de-scaling shifts $p[c_{out}]$ and magnitude for the quantized model parameters allows control over bit depth of register $R[c_{out}, i, j]$ (guaranteed to be within ***32 bits***).
...

Spatial Prediction @ Latent Domain

- Aims to predict the mean of $\tilde{y}$ using the explicit prediction and residual decoded data
  - ✓ 3D chess-board split of the tensor
- Significant complexity reduction (minimizes serial processing) in comparison to previous approaches such as wavefront parallelizable models with masked convolutions

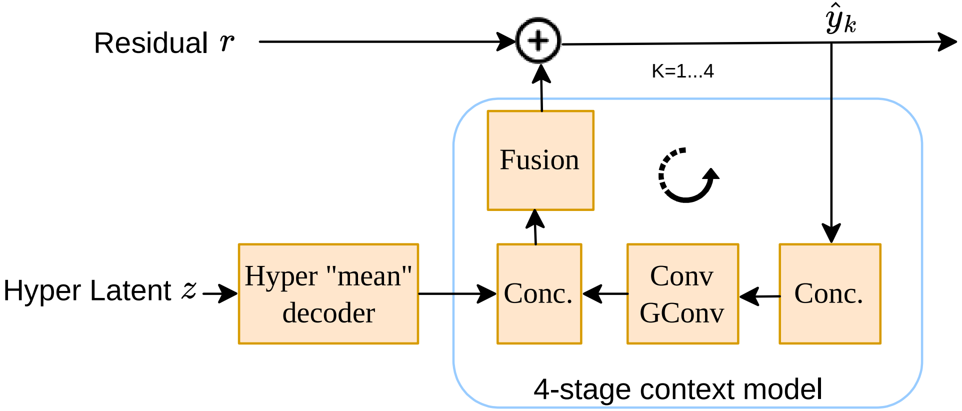


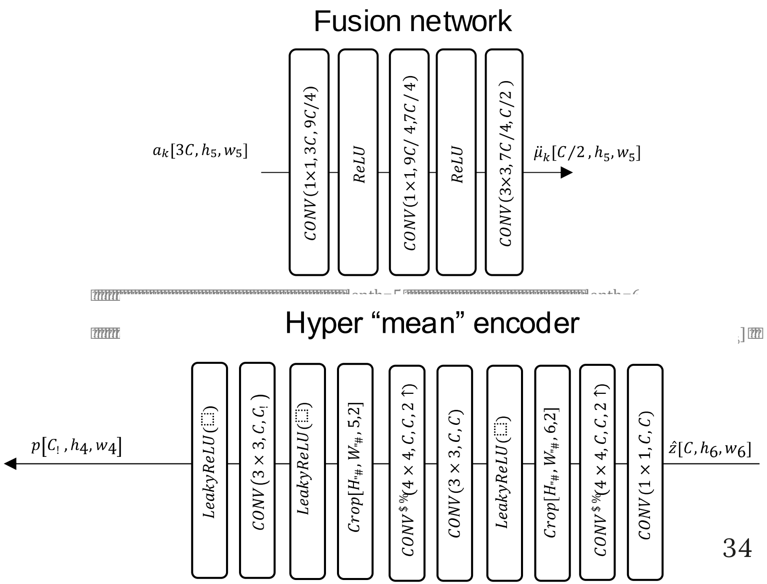


“3D chess-board” split of tensor

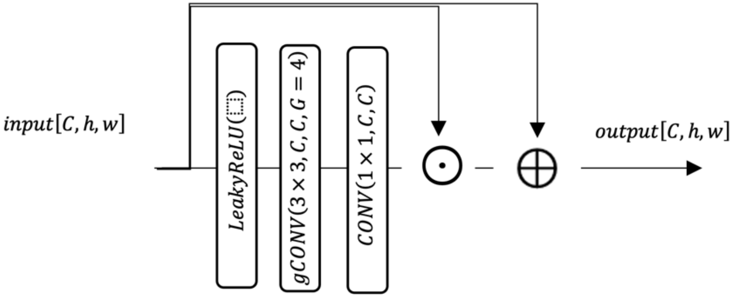
Multistage (4-stage) Context Model

- Hyper-mean encoder provides an explicit prediction derived from the hyper latent tensor
- 4-stage context model: concatenates and process already reconstructed latent sample groups which are fused together with the explicit prediction of the hyper mean decoder





Synthesis Transform



High Operation Point
~180 kMAC/pxl

Base Operation Point
~20 kMAC/pxl

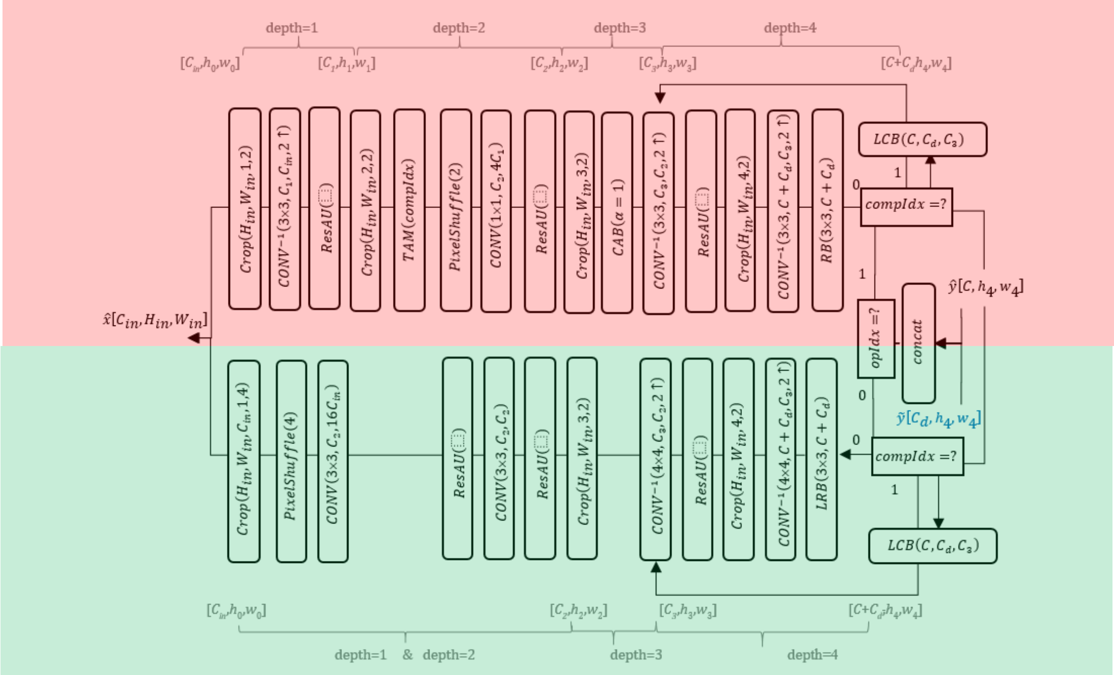


Figure 8.3-1 – Synthesis transform Net

Network is deeper for primary component (Luma)

IMPROVE 2025

Bring the Attention !

Attention modules <u>only for High profile</u>

TAM

***T**ransformers-based **A**ttention **M**odule*

CAB

***C**onvolutions-based **A**ttention **B**lock*

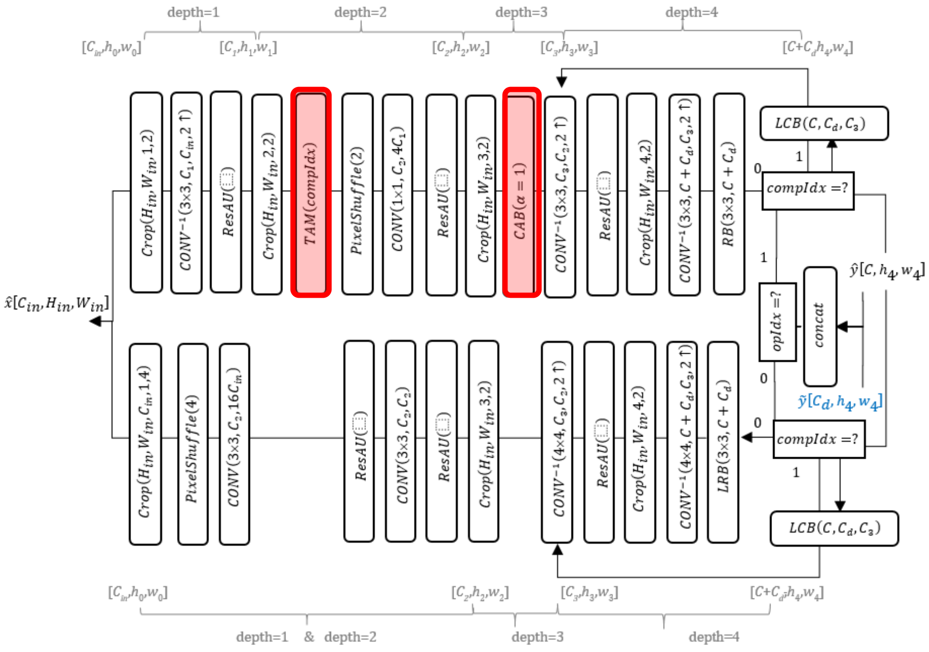


Figure 8.3-1 – Synthesis transform Net

Attention Blocks: Convolutional vs Transformer

Three branches to represent skip, feature and mask (to improve receptive field)

Three branches to represent query, key and value
Transposed-attention map A of size C×C is computed

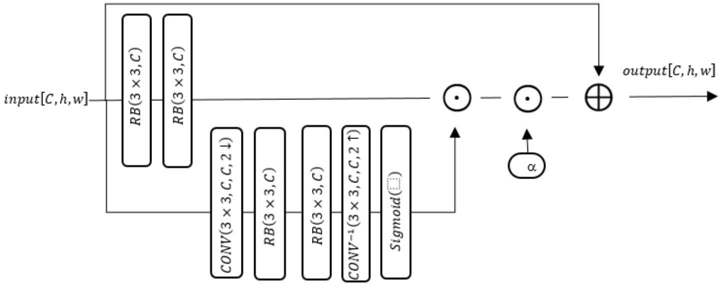


Figure 3.5.30-1 - Convolution-base attention block

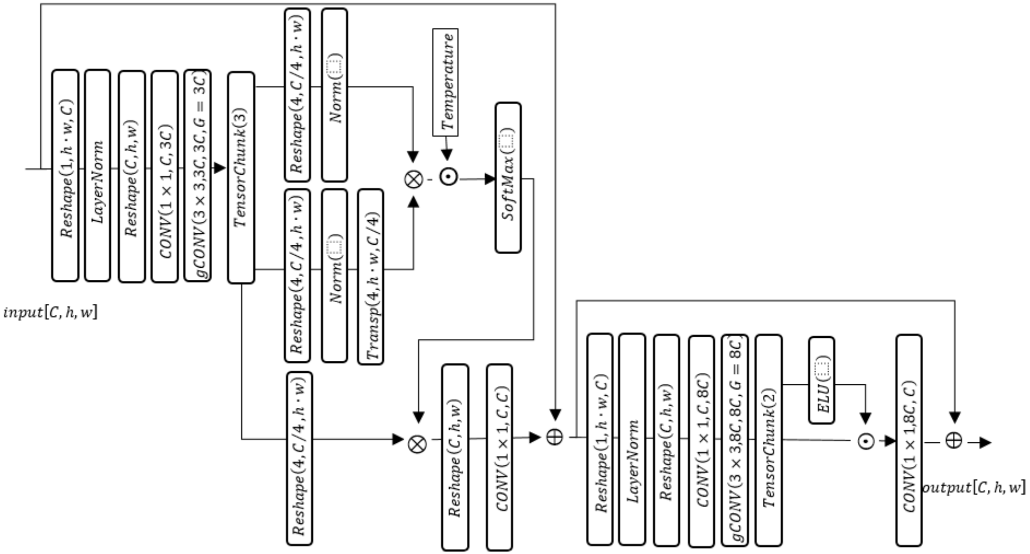


Figure 3.5.39-1 Transformer-based attention block.

JPEG AI Region of Interest Decoding

The residual is multiplied by a gain tensor for local quality control.
Quality index map is predicted, coded and inserted into the codestream

JPEG AI VM3.4 - 0.12 bpp

JPEG AI VM3.4 + ROI coding - 0.10 bpp

Original image

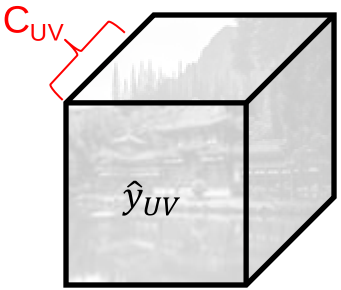
ROI mask (white)

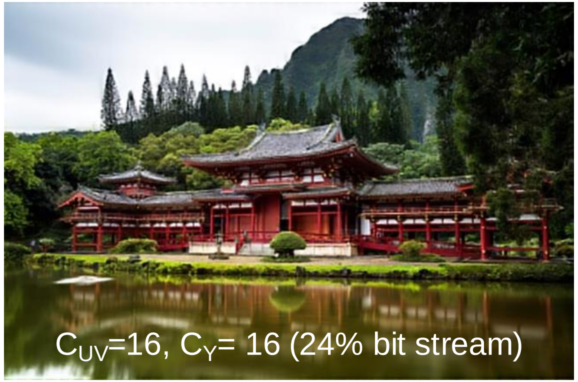
Allocating more bits on the ROI and fewer bits on the background

JPEG AI Progressive Decoding

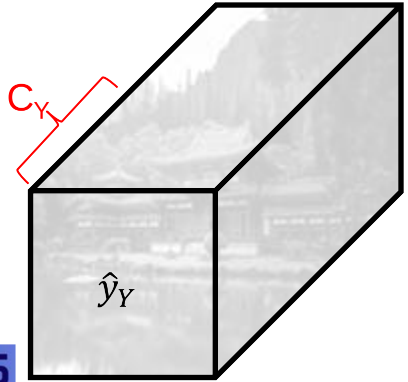
Partial decode part of the 160 channels of residual can reduce the time used for decoding.

Chroma residual
(total C_{UV}= 96)

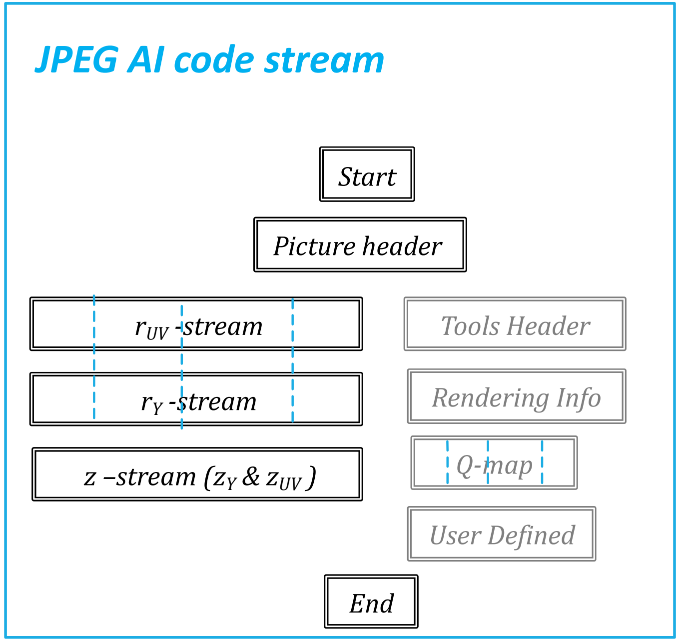


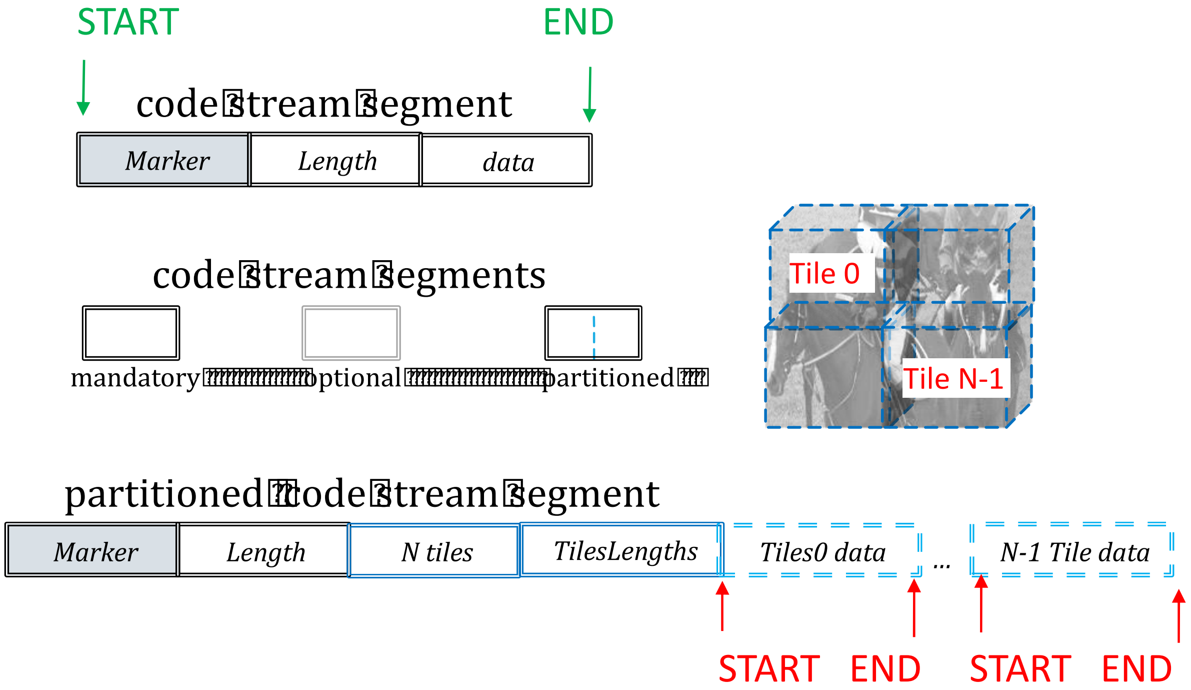


Luma residual
(total C_Y= 160)



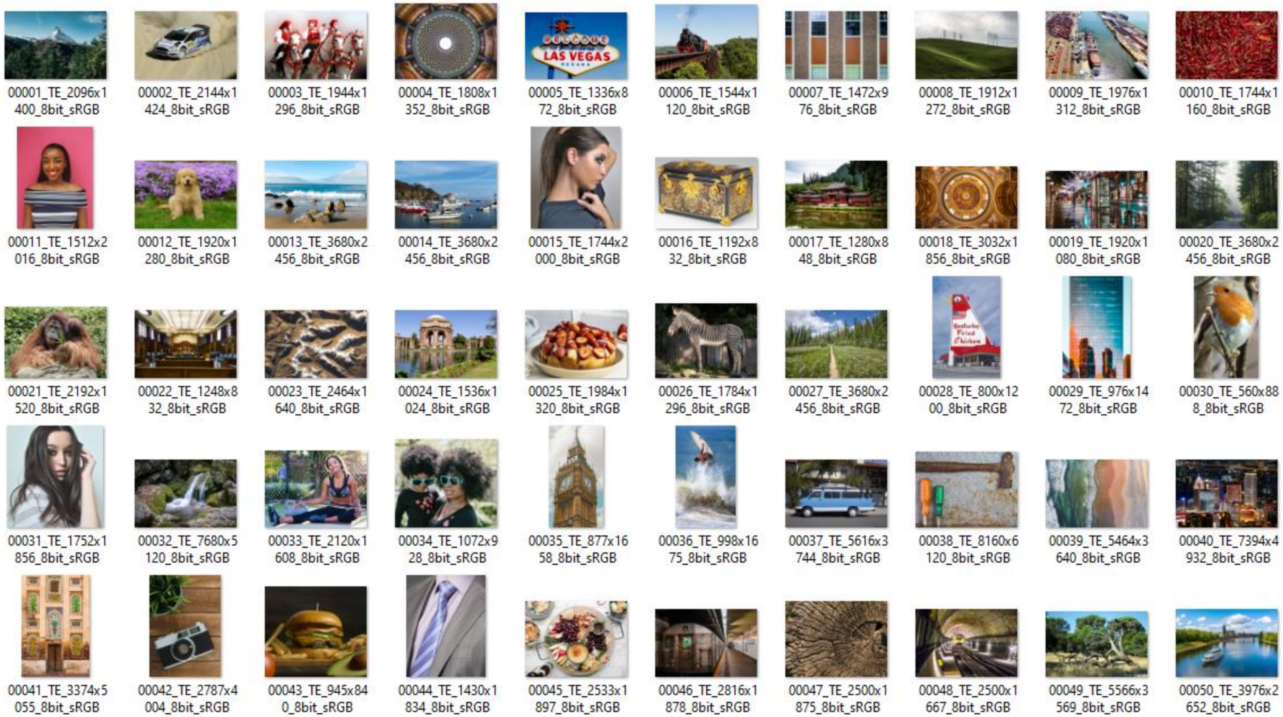
JPEG AI Bitstream Structure





4 Performance Evaluation

JPEG AI Natural Images Dataset

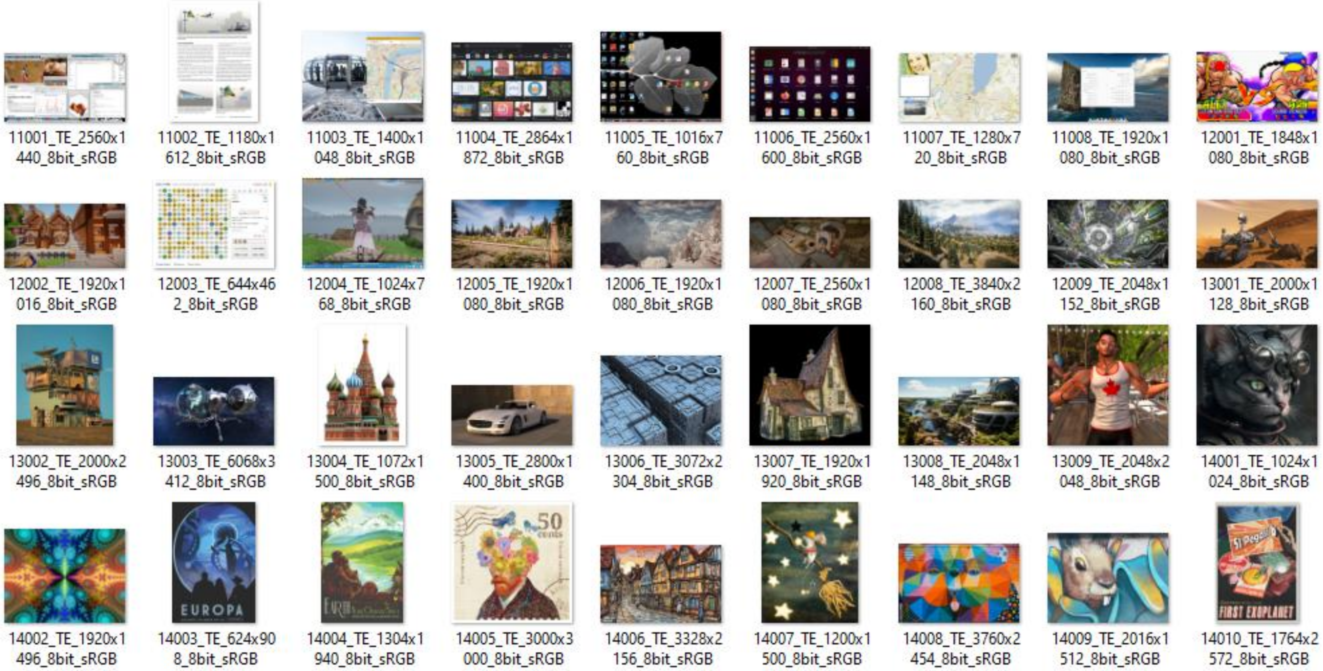


JPEG AI Test Set:
50 camera captured images

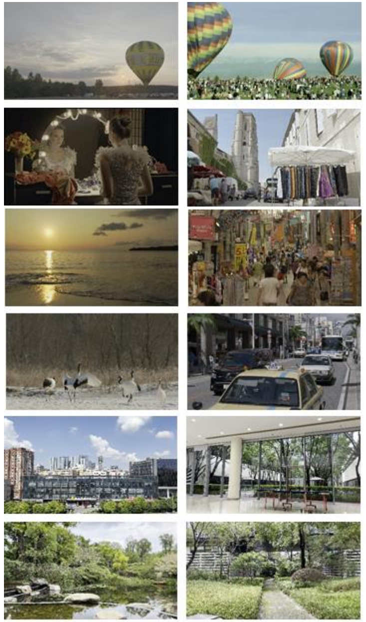
Training Set:
5000+ images
Validation Set:
350+ images

JPEG AI Additional Datasets

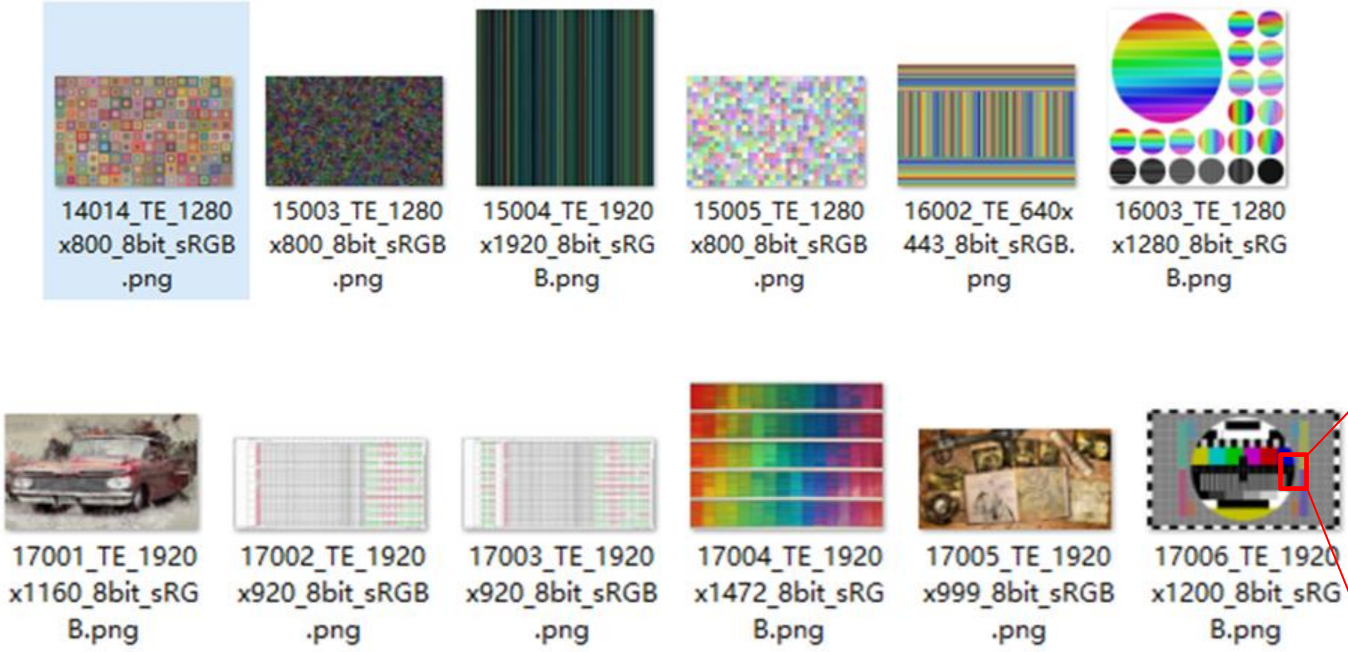
36 synthetic images



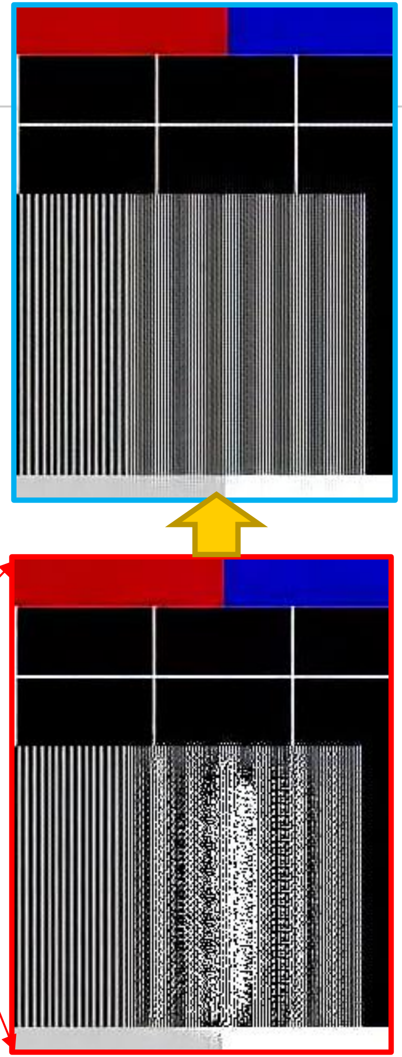
12 HDR images



JPEG AI Crash Dataset

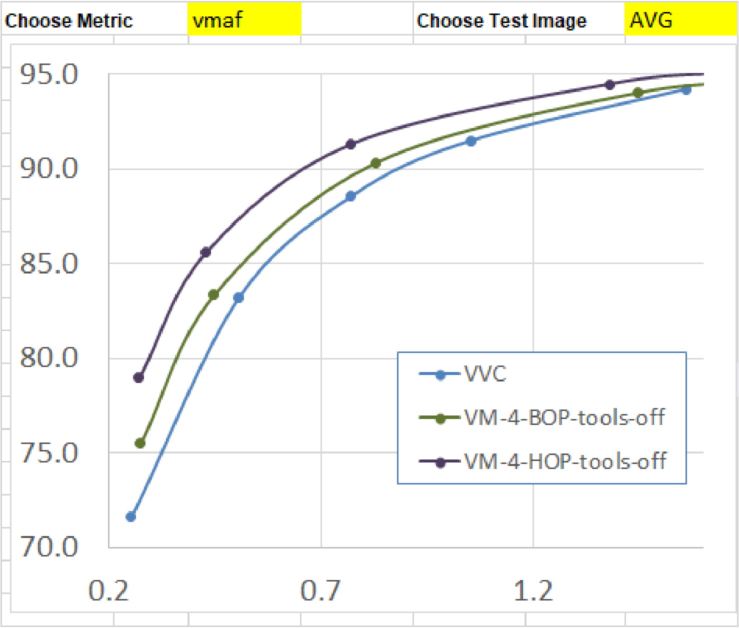


JPEG AI DIS



JPEG AI RD Performance

tools-off: only “off-line trained”, no content adaptation, no encoder search,



JPEG AI VM4 RD Performance

RVS – Residual and Variance Scale
Filters – Adaptive re-sampler, ICCI (cross-color filter), LEF (luma edge filter) and non-linear chroma filter
LSBS – Latent Scale Before Synthesis
CWG – Channel-Wise Gain

Base operating point !

		5 points BD-rate (0.06, 0.12, 0.25, 0.5, 0.75)								10%						
		BD rate vs VVC									Dec. complexity					c. comple:
Test	AVG	msssim Torch	vif	fsim	nlpd	iw-ssim	vmaf	psnrHVS	Monotonicity	Max Bit Dev.	MAX kMAC/pxl	AVG kMAC/pxl	Time GPU, x	Model	ModelS	Time GPU
v4.4-tools-off-GPU	-10.6%	-28.6%	-1.2%	-13.0%	-9.8%	-24.7%	-0.7%	3.9%	TRUE	317%	22	22	0.10	2.93E+06	1.17E+07	0.001
v4.4-tools-on-GPU	-16.2%	-27.3%	1.8%	-28.6%	-13.4%	-24.7%	-26.4%	5.5%	TRUE	393%	29	26	0.18	3.38E+06	1.32E+07	0.002
v4.4-tools-off-GPU-LH	-11.4%	-29.3%	-2.0%	-13.8%	-10.6%	-25.3%	-1.6%	3.0%	TRUE	314%	0	#DIV/0!	#VALUE!	2.93E+06	1.17E+07	0.001
v4.4-only-RDLR	-12.4%	-30.6%	-3.1%	-14.5%	-11.3%	-26.0%	-3.4%	1.8%	TRUE	317%	22	22	0.10	2.93E+06	1.17E+07	0.001
v4.4-only-ResVarScale0	-13.6%	-29.1%	-1.5%	-19.6%	-13.2%	-25.4%	-8.6%	1.9%	TRUE	343%	22	22	0.12	2.93E+06	1.17E+07	0.001
v4.4-only-ResVarScale1	-14.2%	-28.6%	-0.2%	-22.5%	-14.4%	-25.1%	-10.5%	1.8%	FALSE	380%	22	22	0.12	2.93E+06	1.17E+07	0.001
v4.4-only-EnhancementFilters	-11.2%	-28.4%	-0.9%	-14.3%	-9.0%	-24.6%	-5.8%	4.7%	TRUE	318%	28	25	0.14	3.38E+06	1.32E+07	0.002
v4.4-only-LSBS	-11.5%	-28.7%	-1.6%	-12.1%	-9.4%	-24.7%	-8.4%	4.6%	TRUE	317%	22	22	0.11	2.93E+06	1.17E+07	0.001
v4.4-only-ECThread8	-10.6%	-28.6%	-1.2%	-13.0%	-9.8%	-24.7%	-0.7%	3.9%	TRUE	317%	22	22	0.10	2.93E+06	1.17E+07	0.001
v4.4-only-CWG	-12.9%	-28.9%	-0.7%	-20.9%	-12.0%	-25.6%	-5.6%	3.4%	TRUE	328%	22	22	0.10	2.93E+06	1.17E+07	0.001

High operating point !

		5 points BD-rate (0.06, 0.12, 0.25, 0.5, 0.75)								10%						
		BD rate vs VVC									Dec. complexity					c. comple:
Test	AVG	msssim Torch	vif	fsim	nlpd	iw-ssim	vmaf	psnrHVS	Monotonicity	Max Bit Dev.	MAX kMAC/pxl	AVG kMAC/pxl	Time GPU, x	Model	ModelS	Time GPU
v4.4-tools-off-GPU	-25.2%	-38.7%	-16.3%	-26.6%	-24.1%	-35.9%	-22.8%	-11.7%	TRUE	368%	212	207	0.37	9.97E+06	3.99E+07	0.002
v4.4-tools-on-GPU	-28.6%	-36.4%	-13.4%	-38.1%	-25.6%	-34.6%	-43.0%	-9.0%	TRUE	445%	230	221	0.49	1.04E+07	4.14E+07	0.003
v4.4-tools-off-GPU-LH	-25.9%	-39.3%	-17.0%	-27.4%	-24.9%	-36.5%	-23.5%	-12.4%	TRUE	364%	0	#DIV/0!	#VALUE!	9.97E+06	3.99E+07	0.002
v4.4-only-RDLR	-25.7%	-39.5%	-17.2%	-26.8%	-24.4%	-36.3%	-23.3%	-12.3%	TRUE	368%	212	207	0.37	9.97E+06	3.99E+07	0.009
v4.4-only-ResVarScale0	-27.3%	-38.8%	-16.3%	-31.6%	-26.6%	-36.2%	-28.9%	-12.9%	TRUE	392%	212	207	0.38	9.97E+06	3.99E+07	0.002
v4.4-only-ResVarScale1	-27.6%	-38.3%	-15.4%	-32.4%	-27.3%	-35.9%	-30.5%	-13.1%	FALSE	435%	212	207	0.39	9.97E+06	3.99E+07	0.002
v4.4-only-EnhancementFilters	-25.6%	-38.4%	-16.0%	-28.6%	-23.4%	-35.7%	-26.7%	-10.7%	TRUE	369%	218	209	0.40	1.04E+07	4.14E+07	0.003
v4.4-only-LSBS	-25.7%	-38.7%	-16.6%	-25.8%	-23.8%	-35.9%	-28.4%	-11.0%	TRUE	368%	212	207	0.38	9.97E+06	3.99E+07	0.002
v4.4-only-ECThread8	-25.2%	-38.7%	-16.3%	-26.6%	-24.1%	-35.9%	-22.8%	-11.7%	TRUE	368%	212	207	0.36	9.97E+06	3.99E+07	0.002
v4.4-only-CWG	-26.9%	-38.4%	-15.7%	-34.2%	-25.5%	-36.1%	-27.0%	-11.7%	TRUE	376%	212	207	0.35	9.97E+06	3.99E+07	0.002

Performance with Multi-branch Decoding

Only differ in the analysis and synthesis transforms

- Enc0 – Synthesis Transform without attention
- Enc1 – Synthesis Transform with attention
- SOP – Simple operating point
- BOP – Base operating point
- HOP – High operating point

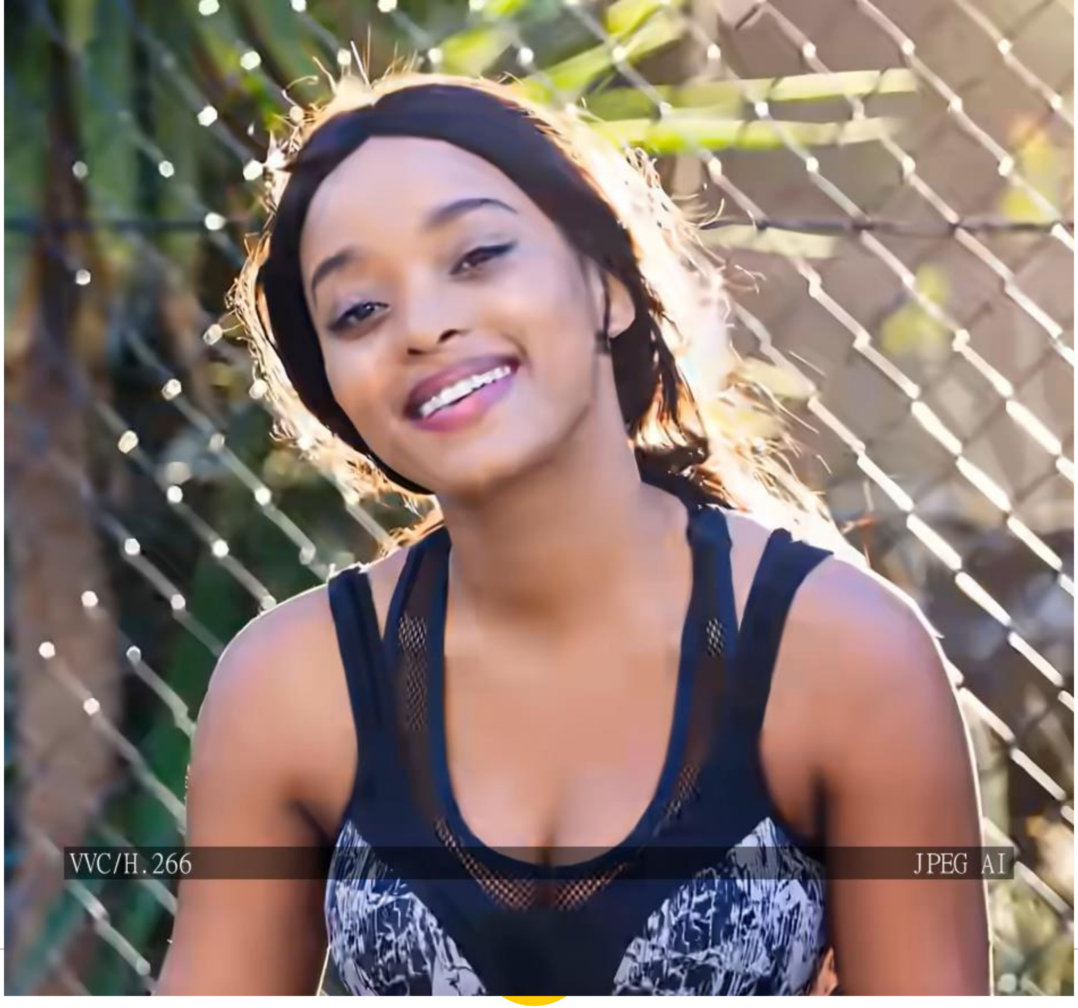
		5 points BD-rate (0.12, 0.25, 0.5, 0.75, 1.0)									
		BD rate vs VVC-012-025-050-075-100							Dec. complexity		Enc. Comp.
Test	AVG	msssim Torch	vif	fsim	nlpd	iw-ssim	vmaf	psnrHVS	kMAC/pxl	Time GPU, x	Time GPU
v5.1-Enc0-SOPDec-tools-off-GPU	-12.4%	-31%	2.8%	-15%	-13%	-27%	-5%	0.9%	8	0.1	0.0004
v5.1-Enc0-SOPDec-tools-on-GPU	-17.5%	-32%	4%	-24%	-15%	-28%	-28%	0.4%	13	0.2	0.0017
v5.1-Enc0-BOPDec-tools-off-GPU	-16.3%	-33%	-2.2%	-20%	-16%	-29%	-11%	-3%	22	0.1	0.0004
v5.1-Enc0-BOPDec-tools-on-GPU	-21.0%	-33%	-1.2%	-28%	-18%	-30%	-32%	-4%	26	0.2	0.0017
v5.1-Enc1-HOPDec-tools-off-GPU	-24.0%	-38%	-12%	-30%	-22%	-34%	-21%	-11%	214	0.4	0.0010
v5.1-Enc1-HOPDec-tools-on-GPU	-28.0%	-38%	-11%	-38%	-24%	-34%	-40%	-11%	216	0.4	0.0023

For the CPU platform, the decoder complexity is 1.6x/3.1x times higher compared to VVC Intra (reference implementation) for the simplest/base operating point.

VVC 0.50 bpp
VMAF = 80.3 PSNR-Y 31.4 MS_SSIM = 0.987

VM3.4-HOP-tools-on 0.44 bpp
VMAF=88.07 PSNR-Y=30.6 MS_SSIM = 0.992

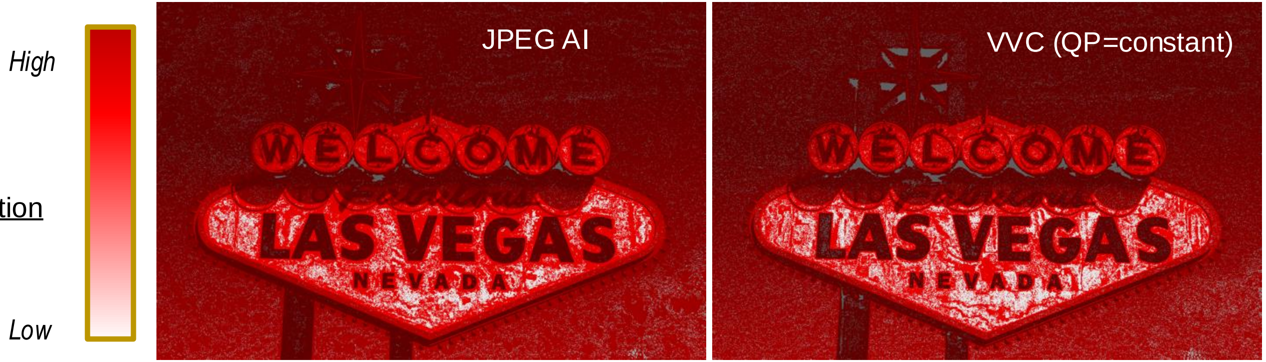
Original


VVC/H.266
JPEG AI

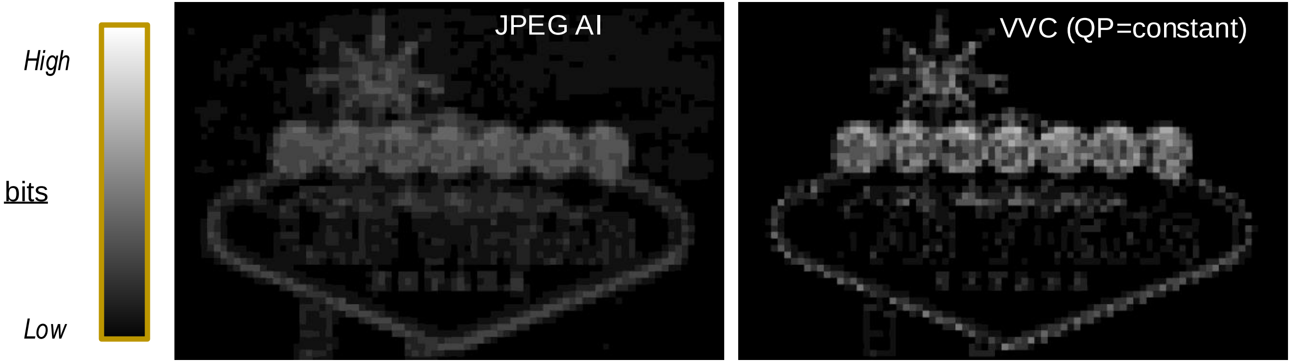
Bits Allocation by JPEG AI and VVC

Original image
1336x872

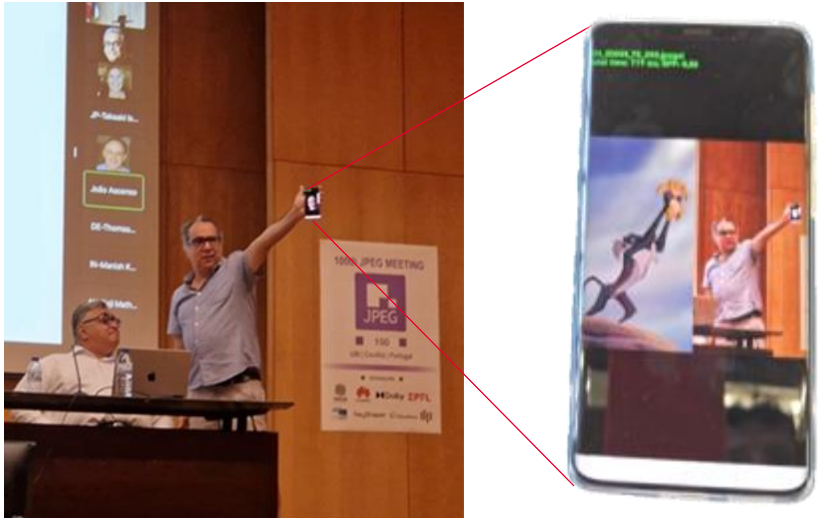
distortion



bits



JPEG AI Decoder on Smartphones



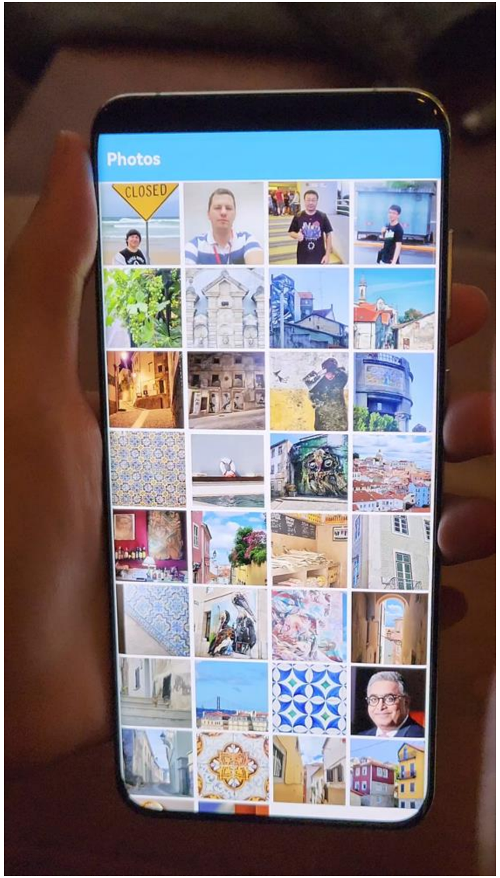
Main targets:

- Demonstrate to the world that JPEG AI can fly on smartphone right now even without dedicated chip
- Identify JPEG AI design issues preventing deployment on mobile platform as early as possible
- Verify device interoperability of JPEG AI standard

- Configuration: JPEG AI CE6.1/VM3.4 base operating point
- Device #1: Huawei Mate50 Pro with Qualcomm Snapdragon 8+ Gen1
- Device #2: iPhone 14/15 Pro Max, 1K patch images

JPEG AI Smartphone Demos

Huawei Mate50 Pro



Iphone 14 Pro Max

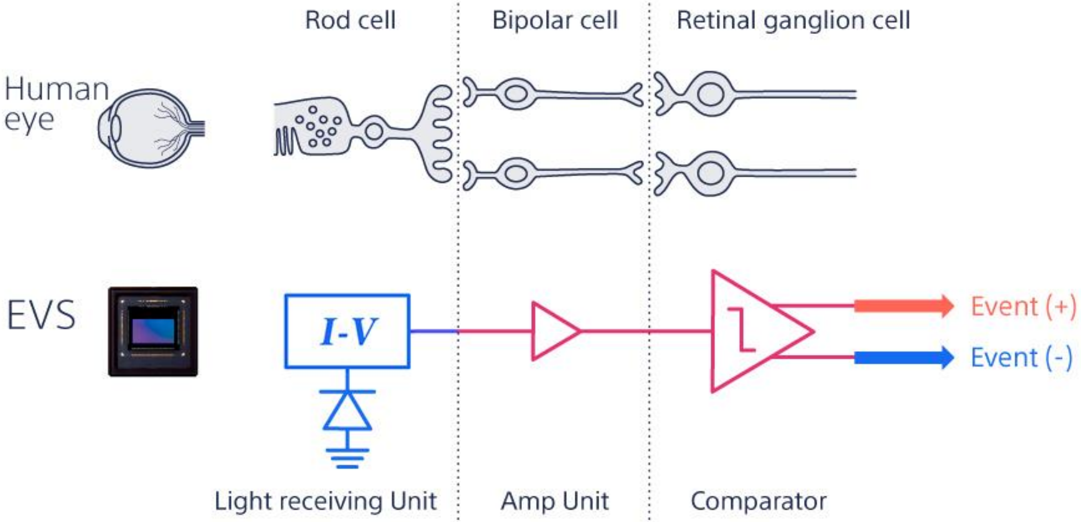
IMPROVE 2025

5 Going Forward …

Biological Inspired Acquisition

Deep learning already disrupted compression! What about sensing?

Differential visual sampling model in which time-domain changes in the incoming light intensity are pixel-wise detected and compared to a threshold, triggering an event if it exceeds the threshold.



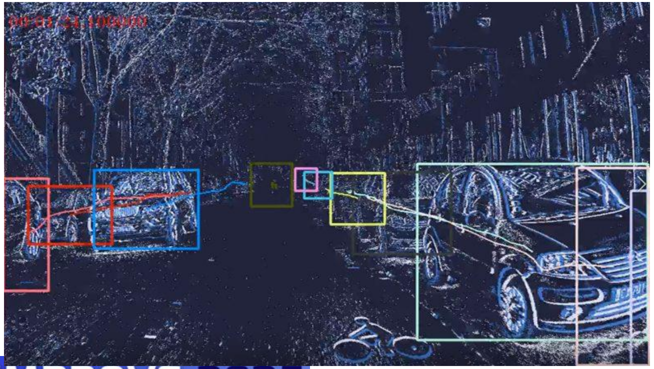
Event-based or Neurmorphic Imaging

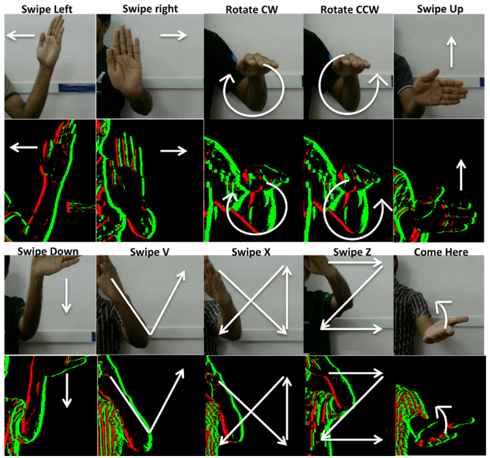
- Event cameras each sensor pixel is in charge of controlling the light acquisition process in an asynchronous and independent way
  - ✓ According to the dynamics of the visual scene
  - ✓ Producing a variable data rate output

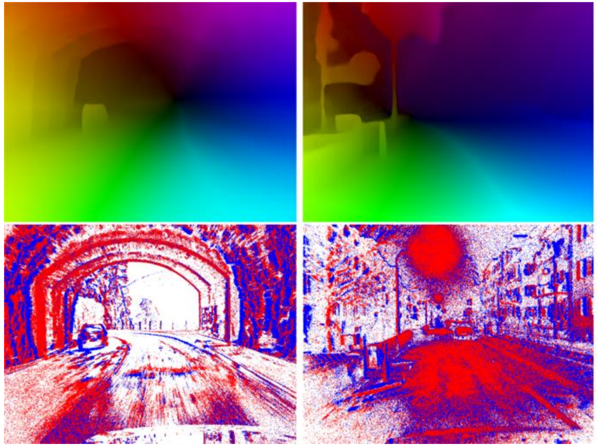
- Relevant advantages:
  - ✓ High temporal resolution
  - ✓ Very high dynamic range
  - ✓ Low latency
  - ✓ Low power consumption
  - ✓ No fixed frame rate

Machine Consumption

- Eye tracking, localization and mapping and gesture recognition for AR/VR headsets
- Obstacle avoidance, localization and mapping,
- Motion capture, gesture recognition and visual interfaces
- Presence detection, object detection and action recognition
- Industrial machine vision

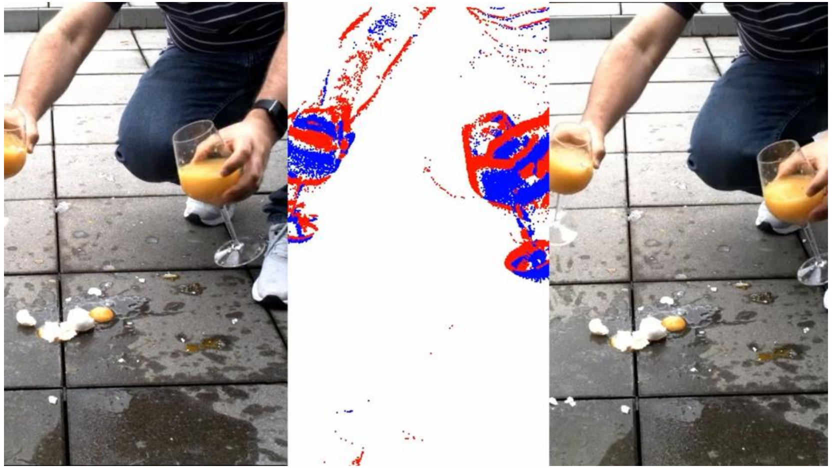






Human Consumption

- Still image deblur
- Motion deblur
- Super slow-motion
- Very high-dynamic range



High dynamic range video reconstruction

New Exploration Activity !

The scope of JPEG XE is the creation and development of a standard to represent Events in an efficient way allowing interoperability between sensing, storage, and processing, targeting machine vision applications.

JPEG AI Next Steps

- Profile/level and conformance discussion is ongoing
- Version 1 addresses several (but not all) JPEG AI 'core' and 'desirable' requirements with emphasis on compression efficiency for standard reconstruction
- Version 2 will address/include:
  - ✓ JPEG AI requirements not yet addressed in version 1, e.g. related to processing and computer vision tasks
  - ✓ Significantly improved solutions for JPEG AI requirements already addressed in Version 1, e.g. compression efficiency

Part	Title	WD	CD	DIS	FDIS	IS
1	JPEG AI: Core coding system	-	23/11	24/04	-	24/10
2	JPEG AI: Profiling	-	24/07	25/01	-	**25/07**
3	JPEG AI: Reference software	-	24/07	25/01	-	**25/07**
4	JPEG AI: Conformance	-	24/10	25/04	-	**25/10**
5	JPEG AI: File format	-	24/07	25/01	-	**25/07**

Final Remarks

- The first learning-based image compression international standard is under active development!
  - ✓ Significant higher compression efficiency compared to the best performing conventional image coding solutions, notably H.266/VVC and H.265/HEVC
  - ✓ Can be efficiently deployed in resource-constrained mobile devices
  - ✓ Much less encoding complexity, online encoder search is now done offline

- Main challenge is to have a multi-purpose bitstream (THE visual language) that is good for a multitude of visual tasks!
  - ✓ Not only image compression but for content understanding and image enhancement!

- “Artificial Intelligence” can be brought to the sensing process to have an even more rich visual data representation!