Video Coding for Machines

Yuan Zhang
China Telecom
January 2022

Contents

What is VCM?

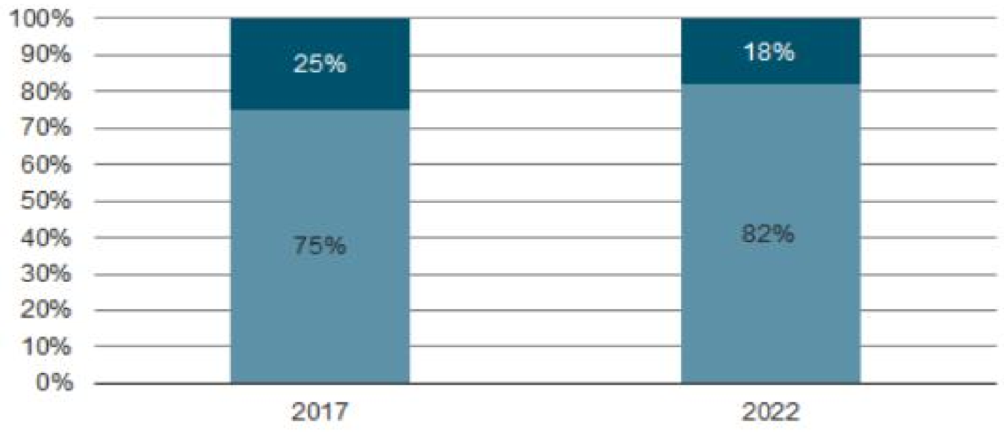
Video Coding for Machine

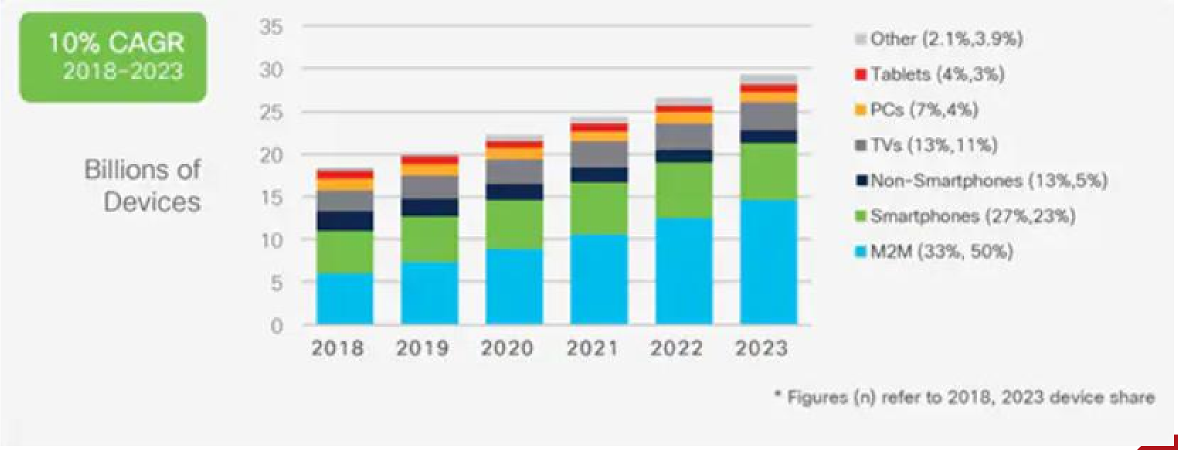
...01001110110001...

Why do we need VCM?

Video has occupied a very large portion of internet traffic.

- More and more video are consumed by machines.
- Automation, analysis and intelligence without or with human intervention -> machine vision or hybrid vision
- Machine-to-Machine (M2M) devices and connections are fast growing.
- Machine vision is different from human vision.
- Different purpose and evaluation metrics
- Video coding for machines becomes an important topic.



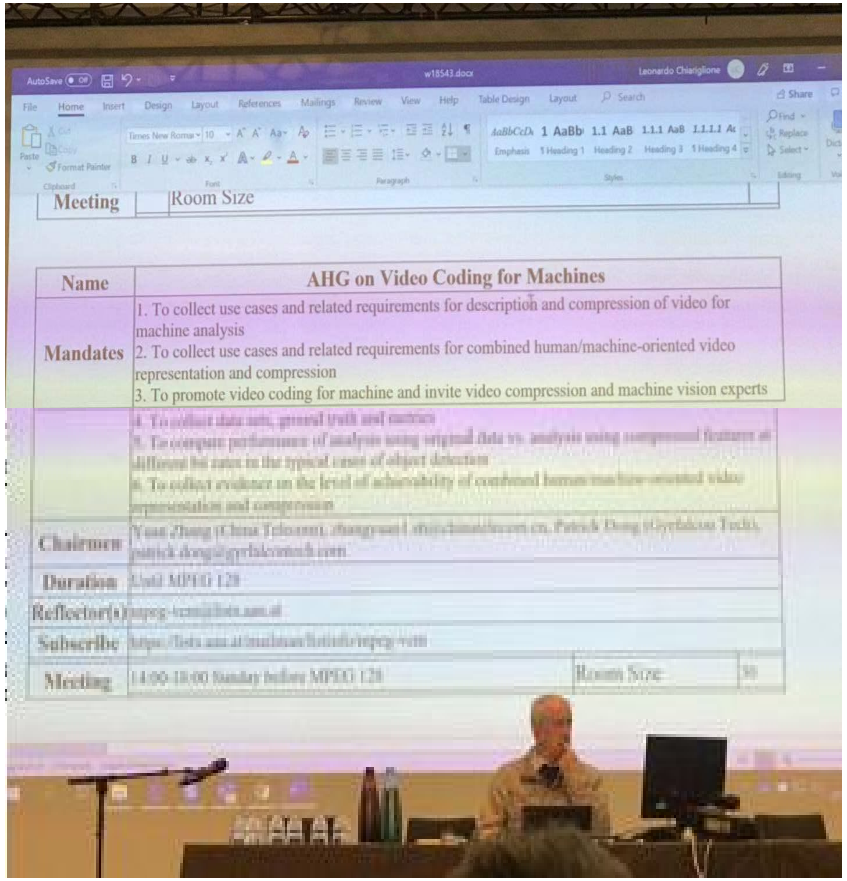


Source: Cisco Annual Internet Report (2018–2023) White Paper

MPEG VCM

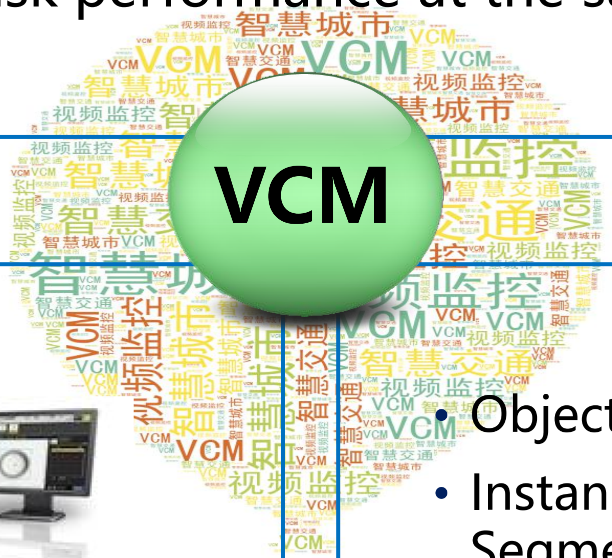
ISO/IEC JTC1/SC29 WG2 committee created the VCM Ad-Hoc Group in July 2019

Name	AHG on Video Coding for Machines		
Mandates	1. To collect use cases and related requirements for description and compression of video for machine analysis 2. To collect use cases and related requirements for combined human/machine-oriented video representation and compression 3. To promote video coding for machine and invite video compression and machine vision experts 4. To collect data sets, ground truth and metrics 5. To compare performance of analysis using original data vs. analysis using compressed features at different bit rates in the typical cases of object detection 6. To collect evidence on the level of achievability of combined human/machine-oriented video representation and compression		
Chairmen	Yuan Zhang (China Telecom), zhangyuan1.sh@chinatelecom.cn Patrick Dong (Gyrfalcon Tech), patrick.dong@gyrfalcontech.com		
Duration	Until MPEG 128		
Reflector(s)	mpeg-vcm@lists.aau.at		
Subscribe	https://lists.aau.at/mailman/listinfo/mpeg-vcm		
Meeting	14:00-18:00 Sunday before MPEG 128	Room Size	30



Scope:

- Define a bitstream from compressed video or extracted feature, which can be used for a variety of machine tasks, and ensuring high compression efficiency and machine intelligent task performance at the same time.



Use Cases:

- Video Surveillance
- Smart Traffic
- Smart City
- Smart Industry
- Smart Content
- Consumer Electronics

Machine Tasks:

- Object Detection
- Instance Segmentation
- Image Reconstruction
- Super Resolution
- ....
- Object Tracking
- Event identification
- Event Prediction
- Density Prediction
- ...

VCM empowers the machine vision industry

Video based and feature based compression experiments are carried out for each sub-tasks. Additional application scenarios need to be refined and researched, including: Smart glasses, unmanned store, unmanned warehouse, robots, smart fishery/agriculture, AR/VR gaming, etc.

	Description	Surveillance / Smart City	Intelligent Transportation	Intelligent Industry	Intelligent Content
Object Detection	Determine a bounding box for an object that may be in the input image / video along with object id	x	x	x	x
Object Segmentation	Determine which pixels belong to which objects by defining binary masks for each image	x	x	x	x
Image/Video Reconstruction	Given the compressed feature stream with an additional bit-stream return the reconstructed image/video	x		x	x
Image/Video Enhancement	With an additional bit-stream return the reconstructed image/video enhanced for human consumption such as super resolution, low light	x			x
Object Tracking	Determine the location of an object throughout video along with object id	x	x	x	
Event Recognition	Determine which event has occurred in the video	x	x	x	x
Event Prediction	Predict which event will occur	x	x		x
Anomaly Detection	Determine whether or not a nonstandard deviation has occurred such as malfunctions	x	x	x	x
Density Estimation	Estimation of population density within a certain bounding box	x			
Event Search	Provide a time stamp for when an event has occurred given an input image or video	x			x
Measurement	Measure the object parameters (size, orientation, curvature, angle)			x	
Object masking	Detect and conceal the certain object in video with a mask	x			x

Potential benefits

Coding video for machines

- ✓ Low bit-rate
- ✓ High precision

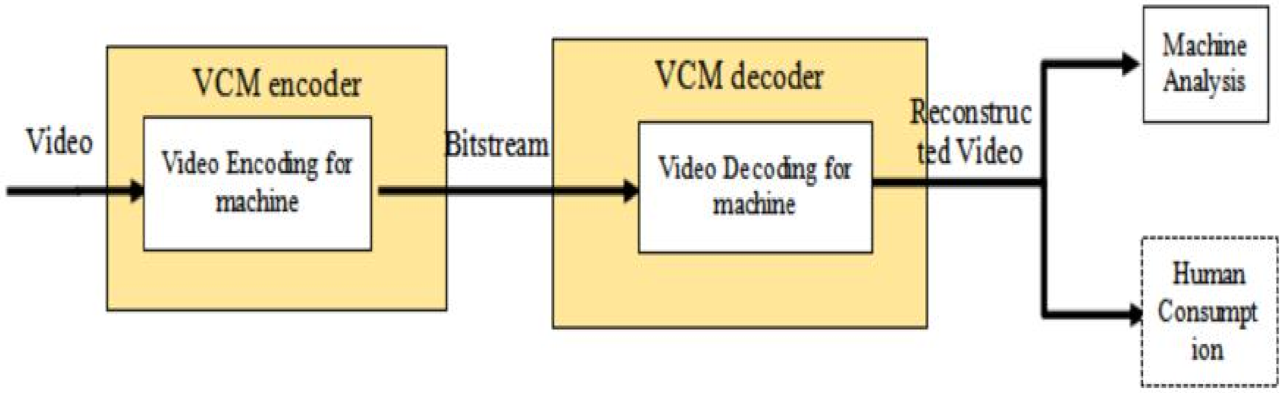
Coding video for human and machines

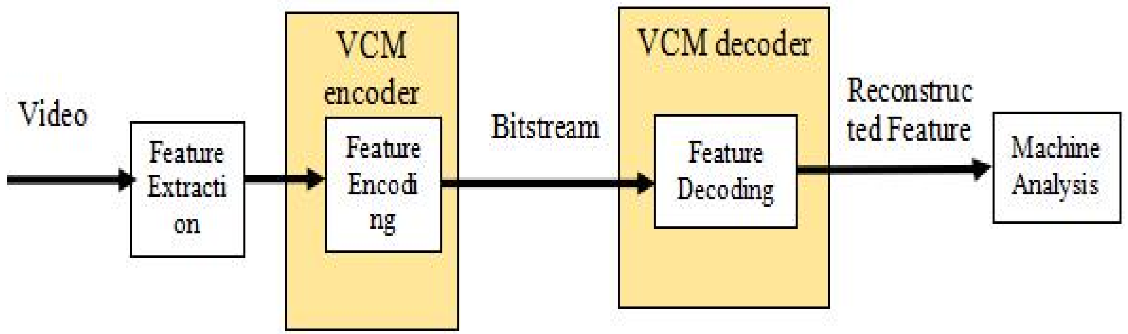
- ✓ Low bit-rate
- ✓ High precision
- ✓ High fidelity

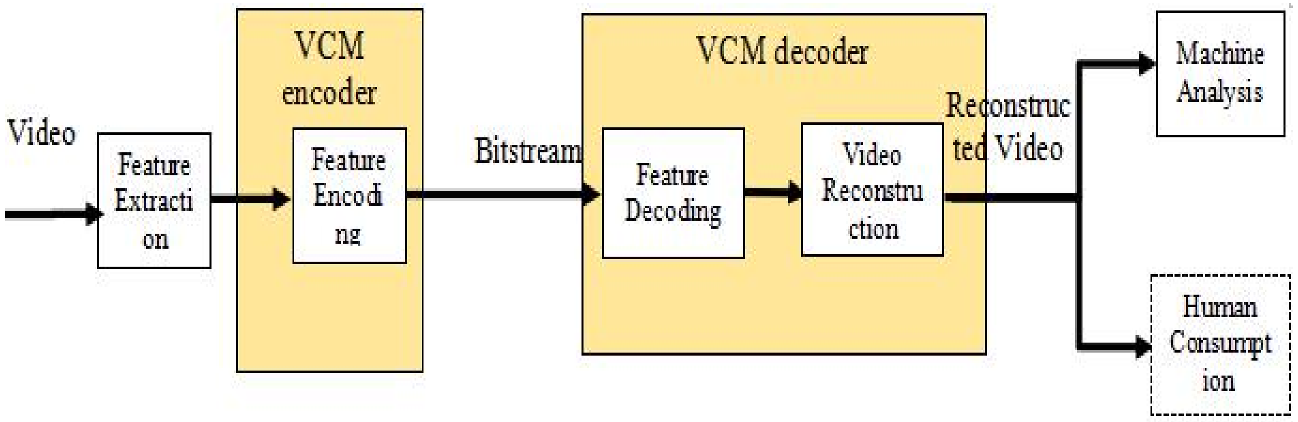
Coding feature for machines

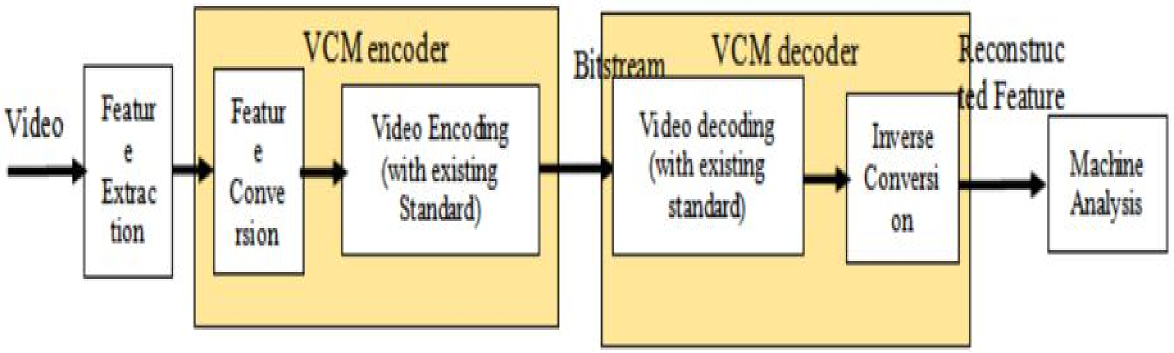
- ✓ Balancing computation load
- ✓ Privacy protection

VCM architectures









Requirements

Video coding

- Coding efficiency shall be significantly improved compared to that of state-of-the-art
- standards.
- Support various intelligent task accuracy, human vision quality and bitrate.
- Either machine only or hybrid machine and human consumption shall be supported.

Feature extraction

- Computational offloading shall be supported.
- Privacy protection shall be supported.

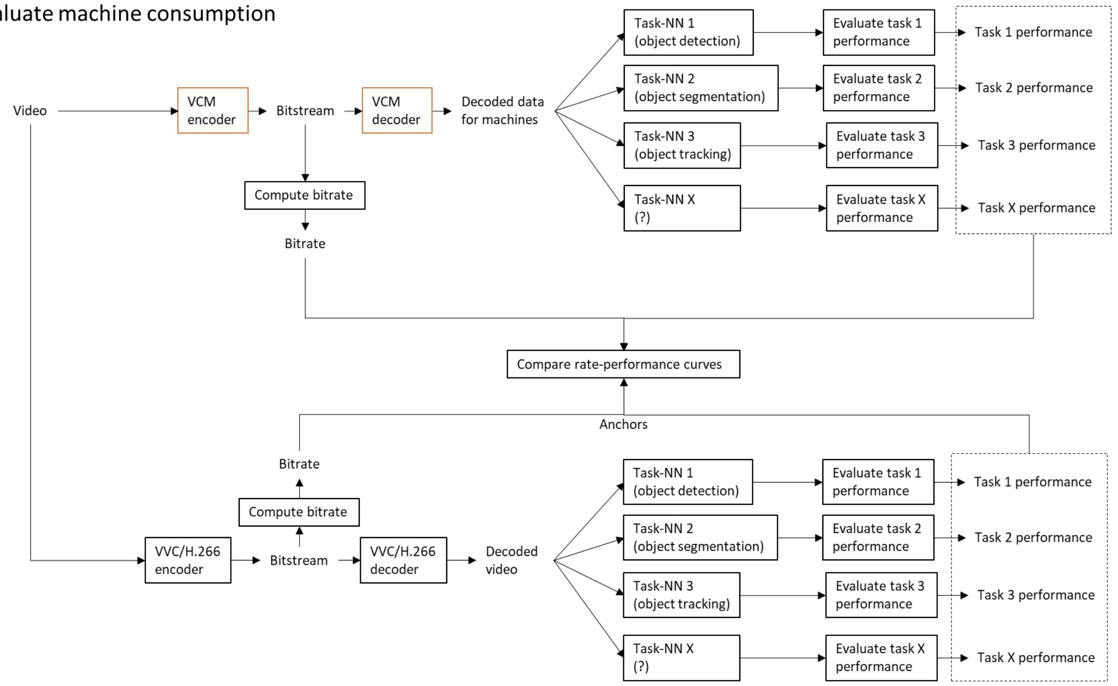
Feature coding

- Coding efficiency shall be competitive compared to the state-of-art video coding solution.
- Support various intelligent task accuracy and bitrate.
- The coding technology shall support machine consumption and support multiple tasks.

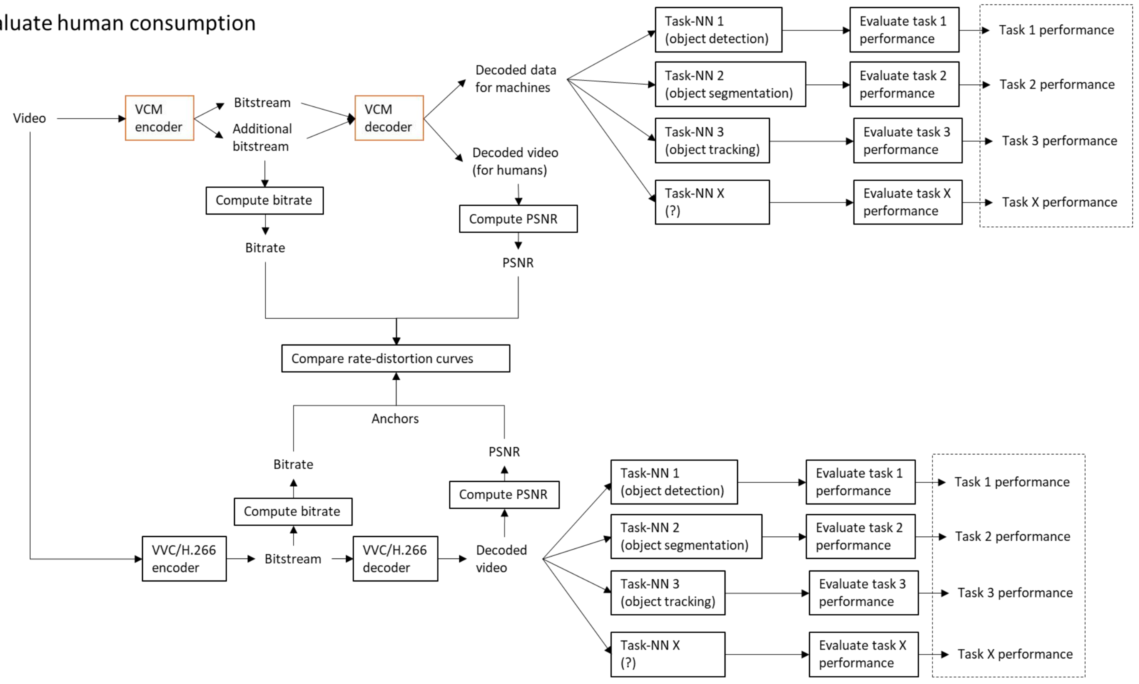
Evaluation Methodology

Anchors are generated using current SOTA technologies, and received technical proposals are compared to anchors according to two aspects: Coding performance and machine task performance.

Evaluate machine consumption



Evaluate human consumption



Evaluation Methodology

Five machine vision tasks are selected to cover the main tasks identified in the use cases.

Five Datasets with suitable license terms are adopted for evaluation.

Machine Task	Network Architecture	Evaluation Dataset	Evaluation Metric
Object Detection	Faster R-CNN with ResNeXt-101 backbone	OpenImageV6 TVD FLIR SFU-HW-object-v1	mAP@0.5 mAP@[0.5:0.95]
Instance Segmentation	Mask R-CNN with ResNeXt-101 backbone	OpenImageV6 TVD	mAP@0.5
Object Tracking	JDE-1088x608	TVD HiEve-10*	MOTA
Action Recognition	SlowFast	HiEve-10*	frame mAP (fmAP)
Pose Estimation	HRNet	HiEve-10*	mAP@0.5

赋 能 未 来

Evaluation Metrics

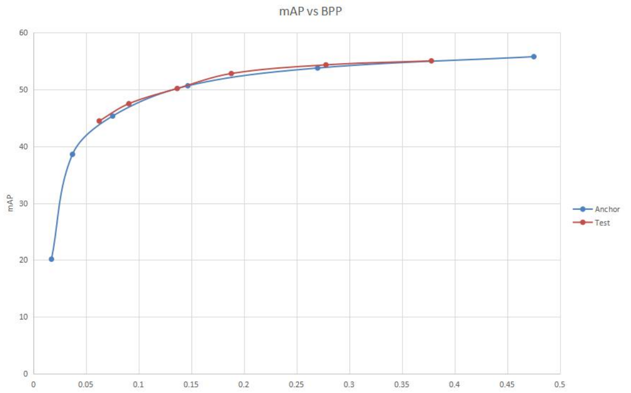
Bits per pixel (BPP) is used to measure bitstream cost for image dataset.

Bitrate in kbps is used to measure bitstream cost for video dataset.

BD-rate and BD-mAP/BD-MOTA/BD-fmAP are used to compare a proposed solution to the anchor solution for a single task.

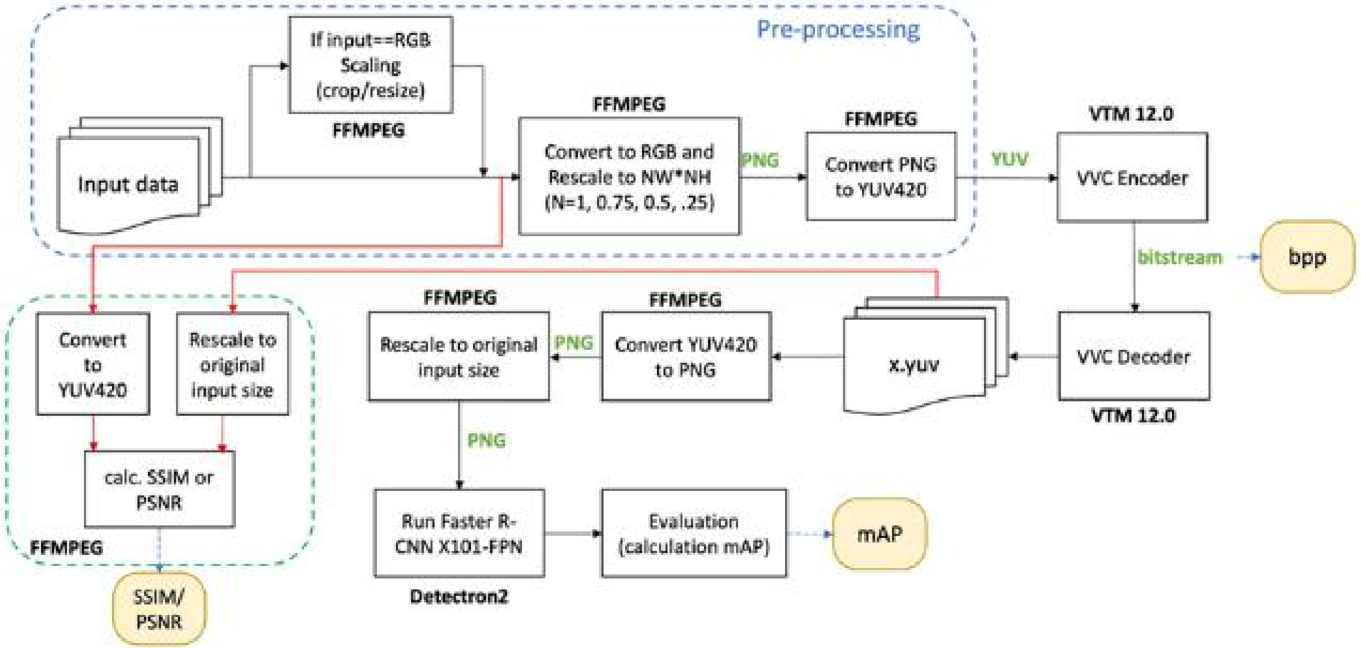
Excel template is used to compute metrics.

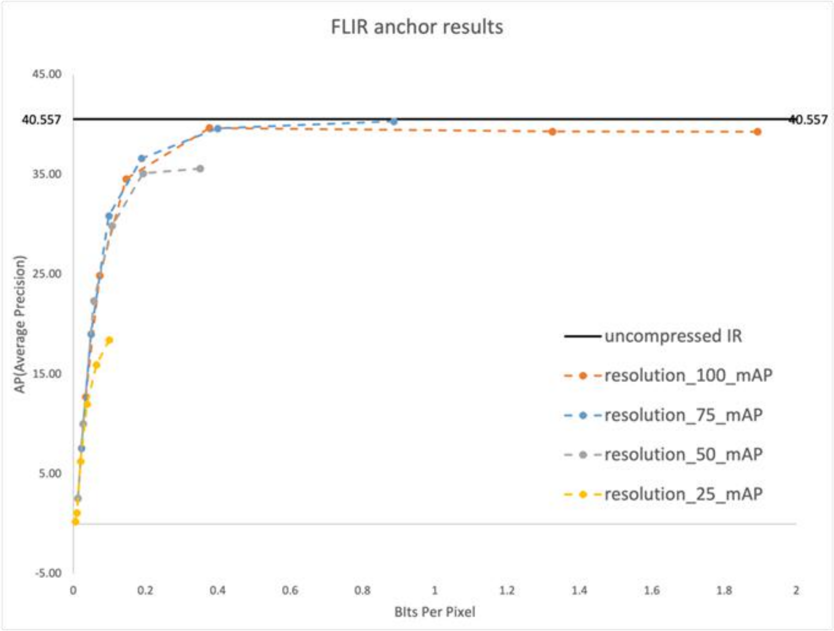
			Reference: VCM Anchor (VTM-12.0)							Test: tested						
Scale	Dataset	QP/Slice	BPP	mAP	Y psnr	U psnr	V psnr	Enc T [h]	Dec T [h]	BPP	mAP	Y psnr	U psnr	V psnr	Enc T [h]	Dec T [h]
100%	OpenImageV6	22	0.863	78.929						0.481	78.890				29.957	32.322
		27	0.509	77.989						0.361	78.453				29.951	32.442
		32	0.287	77.263						0.246	77.787				29.144	31.820
		37	0.153	73.963						0.172	76.418				29.118	31.543
		42	0.078	68.842						0.115	74.242				29.119	31.650
		47	0.037	58.021						0.079	71.488				29.108	31.508
	FLIR	22	1.892	39.317	43.079											
		27	1.325	39.323	38.038											
		32	0.376	39.685	31.483											
		37	0.146	34.578	29.758											
		42	0.072	24.888	28.319											
		47	0.034	12.746	26.605											
	TVD	22	0.475	55.748						0.378	55.011				2.605	2.866
		27	0.270	53.752						0.278	54.307				2.605	2.880
		32	0.147	50.632						0.188	52.785				2.588	2.876
		37	0.075	45.311						0.137	50.152				2.585	2.878
		42	0.037	38.586						0.091	47.480				2.589	2.873
		47	0.017	20.155						0.063	44.449				2.539	2.908



Anchor Generation

For each intelligent task (like object detection, object segmentation, object tracking, etc.), the anchors are generated following a fixed pipeline: Preprocessed using ffmpeg4.2.2, Coded using VTM 12.0 with 4 different resolutions (100%, 75%, 50%, 25%) and 6 different QPs (22, 27, 32, 37, 42, 47).



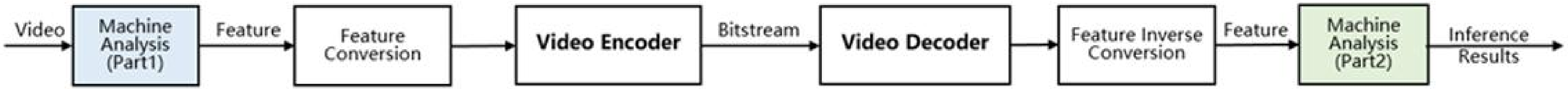


Received Technologies

The Received Technologies can be classified into two categories:

Category 1 (Track 1): Feature Coding

The input to the codec is usually feature map from a neural network.

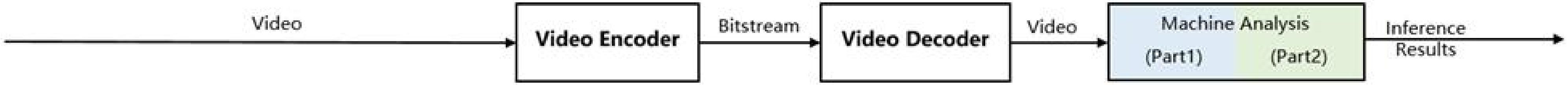


(a) Coding features directly

(b) Coding features as images using existing codec

Category 2 (Track 2): Image/Video Coding:

The Codec module typically follows a video-in-video-out manner.



(a) End-to-End Coding

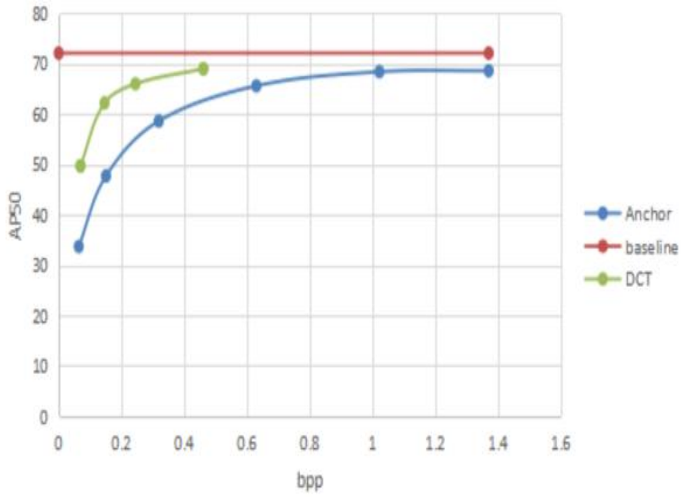
(b) Descriptor based Coding

(c) Enhancing Image Coding for Machines with Compressed Feature Residuals

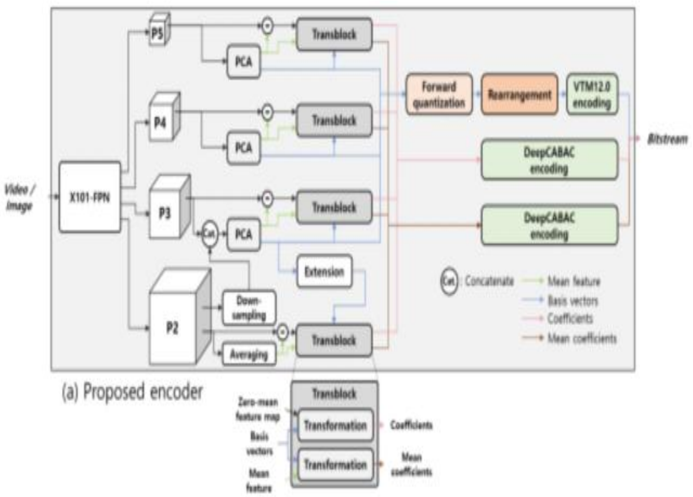
Received Technologies: Feature Coding

Category 1(a): Packed features are coded directly

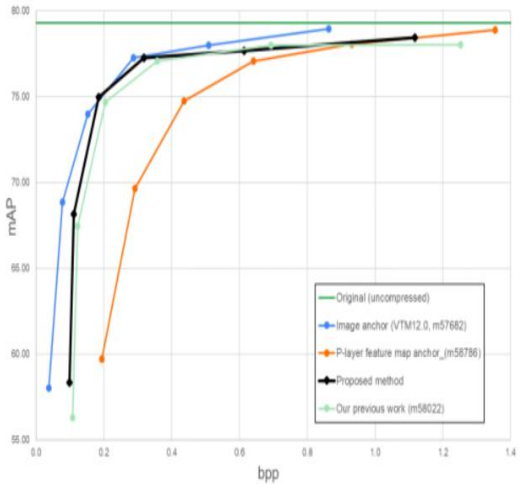
Features are directly coded with new coding kernels, typically follows a Quantization + Entropy Coding manner which achieves close performance as coding images using VTM codec



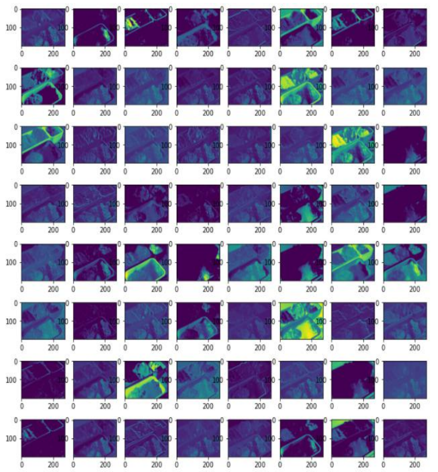
DCT/DWT + CABAC (m58000)



PCA + deepCABAC (m58787)



kmeans + BAC (m56749)

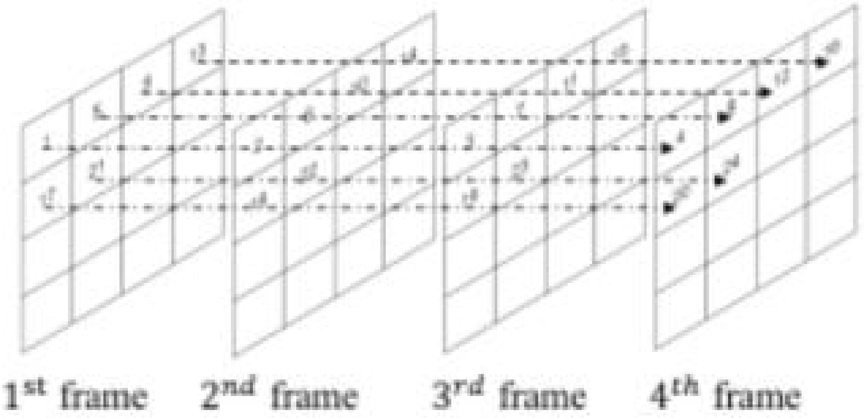


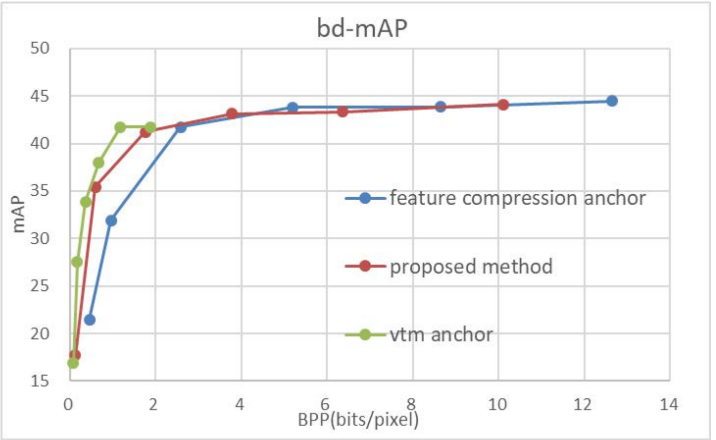
Channel Correlation measure based on KL-Divergence(m57394)

Received Technologies: Feature Coding

Category 1(b): Packed features are coded using video codec

Features are packed as images or videos and coded using VVC. The order of channels are enhanced so that the prediction module in VVC can perform at it best. Resulted bitstreams are much larger than those from VCM anchor solution





(m58081)

赋 能 未 来

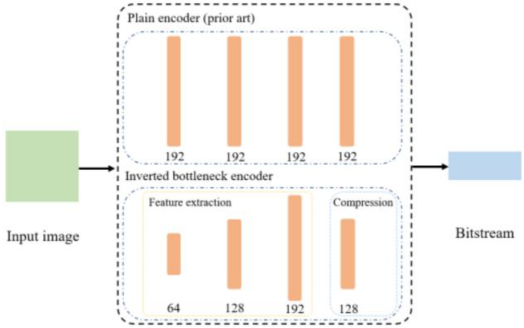
Received Technologies: Image/Video Coding

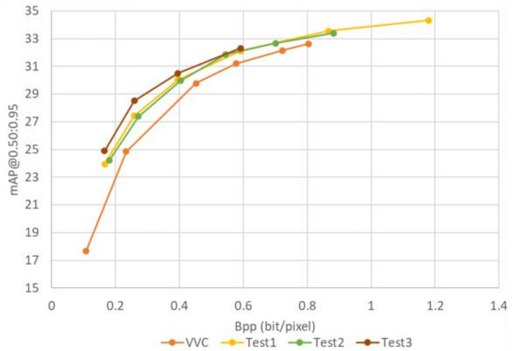
Category 2(a): End-to-end Learning Based Codec

Image compression network: Cheng2020, bmshj2018_hyperprior, or modified mbt2018-mean network

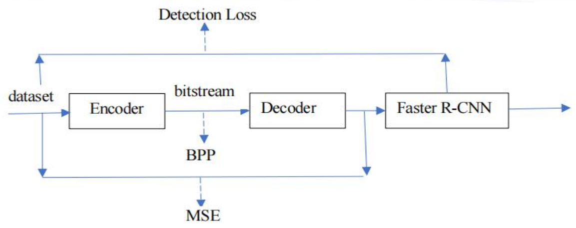
Jointly trained with VCM object detection network in which its parameters are fixed.

- Inverted Bottleneck Structure + Joint Optimization of MSE, bitrate, and task accuracy(m56416), a maximum BD rate gain of 28.09% is achieved.

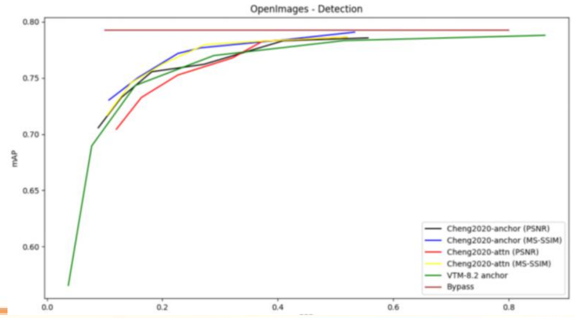




- MS-SSIM optimized Cheng2020 network(m58050), a BD-rate gain of 23.56% is achieved.



$$L_{overall} = R + \lambda_{mse} L_{mse} + \lambda_{detect} L_{detect}$$

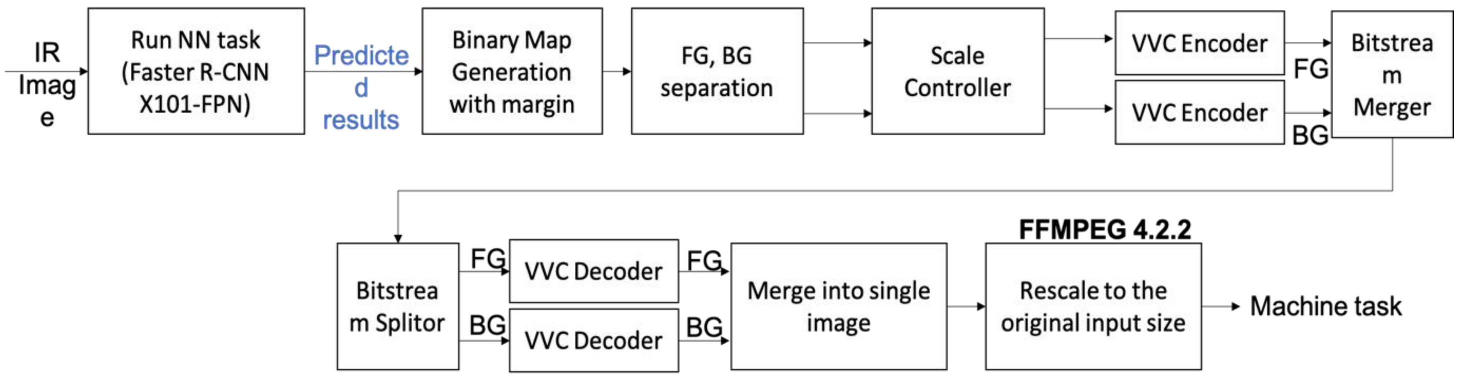


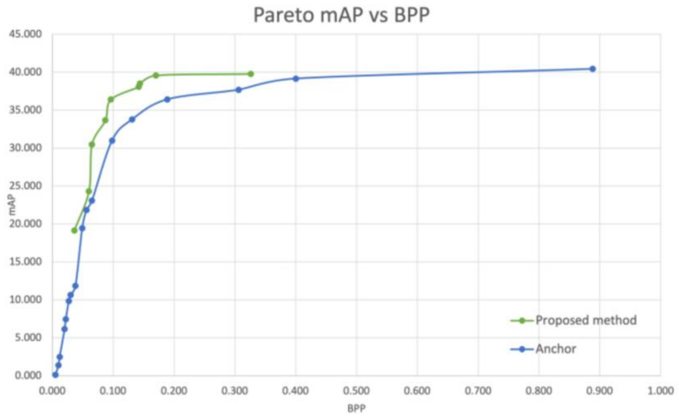
Received Technologies: Image/Video Coding

Category 2(b): Descriptor based Codec

Images are seperated into foreground and background using a pre-detection, and the foreground is coded using a lower QP while background is coded with a higher QP.

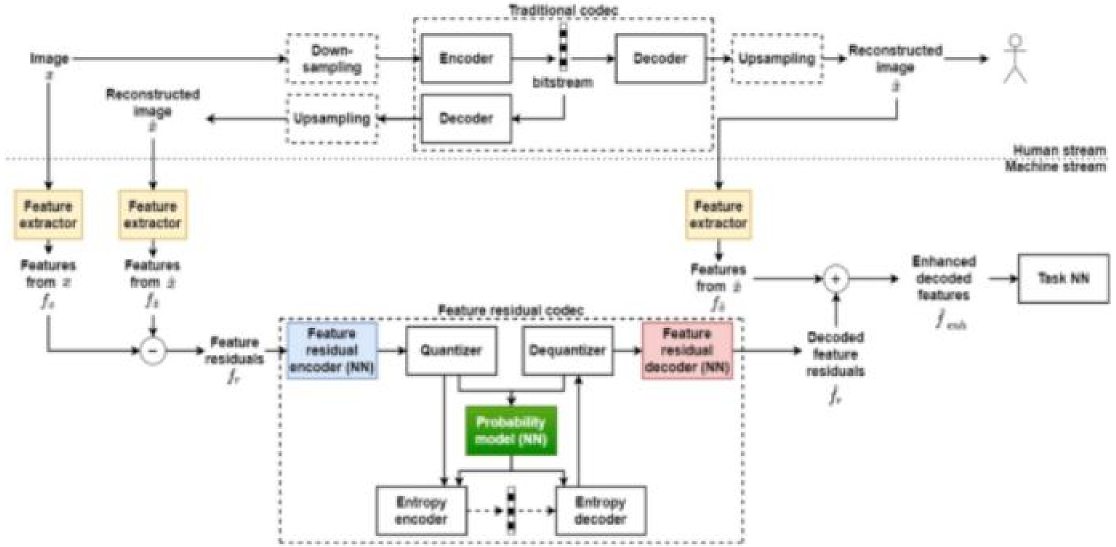
Region Based Coding with Machine Attention(m56572), a BD-rate gain of 30.76% is achieved

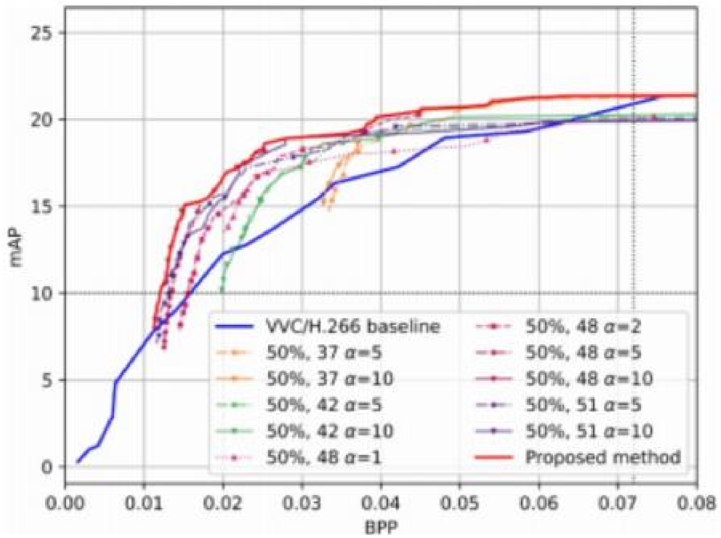




Category 2(c). Enhancing Image Coding for Machines with Compressed Feature Residuals

- CityScapes dataset is used. Fast R-CNN as the object detection task network
- Compared to VVC/H.266, achieve BD-rate gain 40.5%





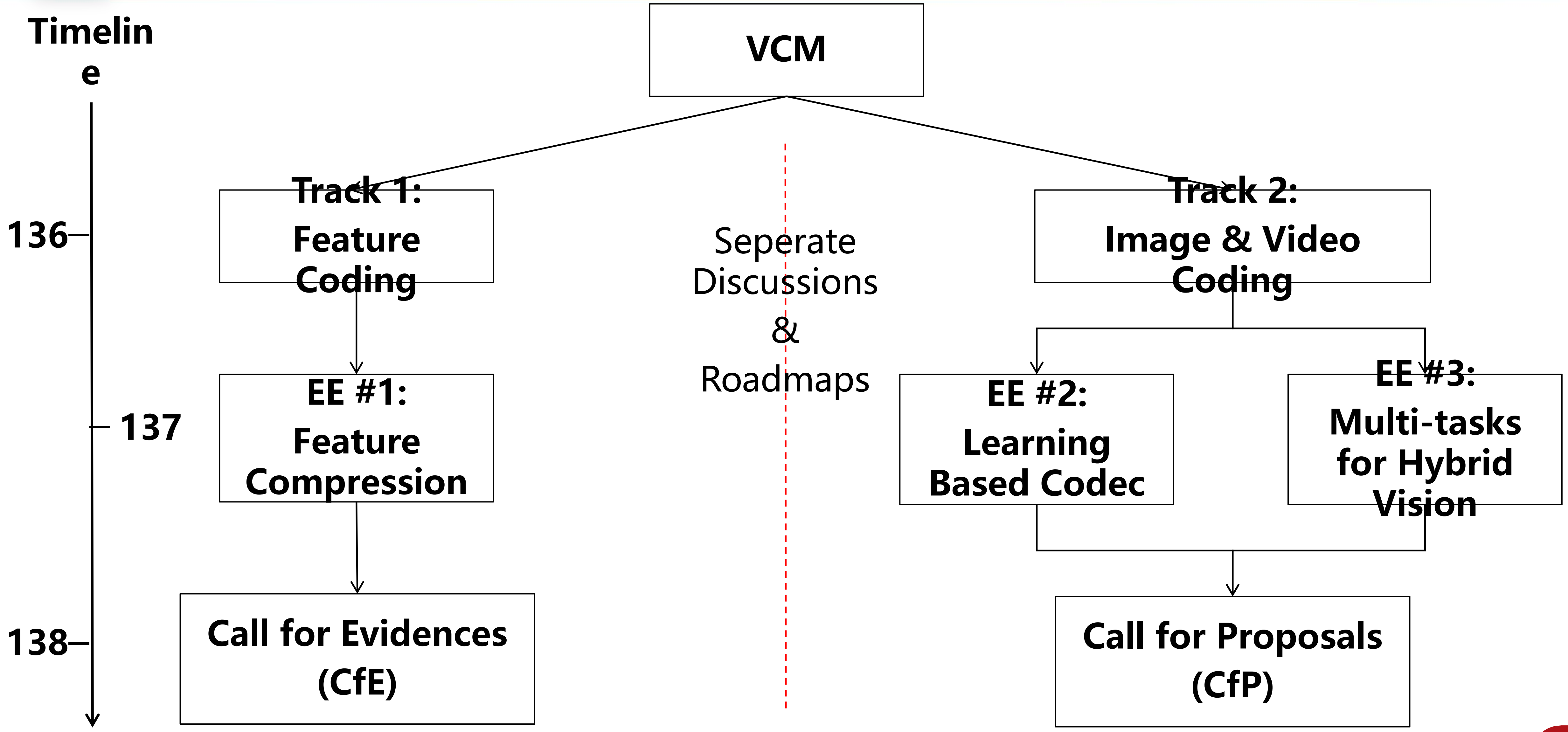
(m58072)

Two-Track Work Plan

In October 2021 MPEG meeting, it was decided to split MPEG VCM work into two tracks:

- Track 1 – Feature extraction and compression
  - ✓ Draft CfE: April 2022
  - ✓ CfE: July 2022
- Track 2 – Images and video compression
  - ✓ Draft CfP: January 2022
  - ✓ CfP: April 2022

Exploration Experiments(EEs)



EE1: Feature Compression

EE1 was launced with the target of better understanding contributed technologies related with feature compression for VCM.

Anchors

1. Codec:
- VTM 12.0

2. Network Split point:
- Faster R-CNN X101-FPN Stem layer
- Faster R-CNN X101-FPN p2-p5
- YOLOv3 layers 75, 90, 105
- DarkNet-53 Split Point

3. Tensor to YUV:
- Each channel packed in a raster-scan order filling the frame area. And form a YUV400 file.

4. Machine Task:
- Object Detection

5. Dataset:
- Open Images

Compared to

Technologies

1. Quantization:
- kmeans
- PCA
- Normalization + z-score
- DCT, DWT
- …

2. Channel Correlation
- Feature prediction using KL-Divergence, Levene Test, and Linear Correlation
- Enhancing Image Coding with Compressed Feature Residuals
- …

3. Channel Reordering
- Reorder according to means
- Coding block aligned resizing

4. …

EE2: Learning based Codec

EE2 was launced with the target of studying the performance of learning-based compression for machine vision tasks including end-to-end training-based compression method.

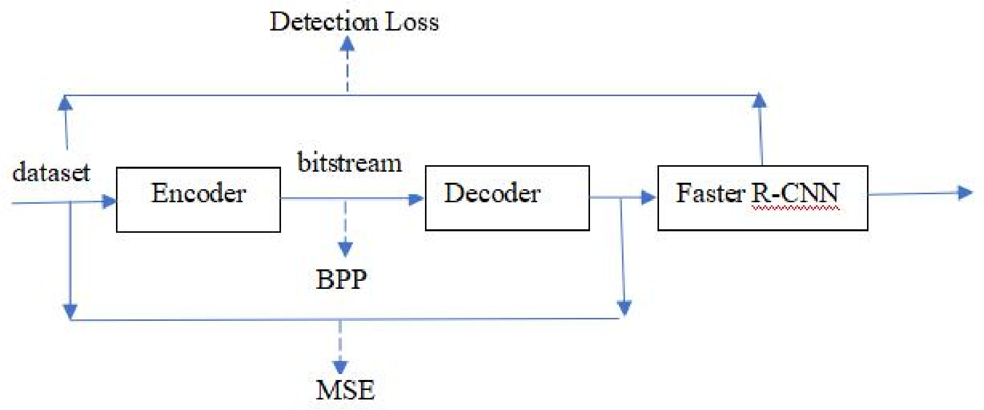
Four experiments

1. Compressor 1: Cheng2020
2. Compressor 2: bmshj2018_hyperprior
3. Compressor 3: mbt2018-mean with inverted bottleneck
4. Compressor 4: MS-SSIM optimized Cheng2020-anchor

Three tasks

Object Detection, Instance Segmentation, Object Tracking

Jointly Optimized Using Machine Task Accuracy, Fidelity Measures, and bpp



Dataset	EE2 subtest (Task)	Compressor 2	Compressor 3	Compressor 4
OpenImages	Object detection	-17.92%	-68.61%	-30.12%
	Instance segmentation	-17.46%	-14.99%	-31.44%

TVD	Object detection	7.66%	-35.64%	-6.81%
	Instance segmentation	11.16%[3]	20.75%	-98.54%[3]
	Object tracking	1881.82%[4]	2393.07%[4]	204.01%[4]
SFU-HW	Object detection	2938.99[4]	2016.85%[4]	Not available

EE3: Multi-tasks and Hybrid Vision

EE3 was launced with the target of studying the performance of technologies that support multi-tasks with hybrid machine/human vision.

Three experiments

1. Region-based coding with machine attention
2. OoI(Object of Interest) based coding with machine description
3. End-to-end learning-based compression trained on object detection and inferenced on other tasks

Three tasks

Object Detection, Instance Segmentation, Object Tracking

Compressing Image or Videos for Down--stream Tasks with the Guidance of Up-stream Features

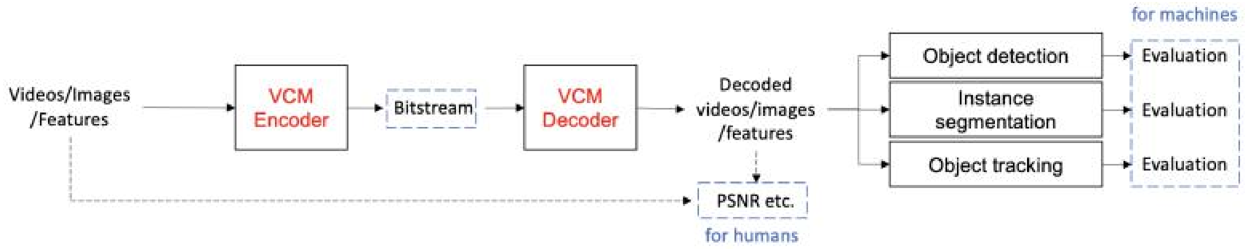


IMAGE			Test Task 1 Object Detection				Test Task 2 Object Segmentation			
Scale	Dataset	QP	bpp	mAP	Normalised mAP	weight	BPP	mAP	Normalised mAP	weight
100%	TVD	22	0.209	54.585	97.329	0.6	0.209	44.291	98.065	0.4
		27	0.118	52.023	92.762		0.118	43.094	95.413	
		32	0.064	47.667	84.994		0.064	37.916	83.950	
		37	0.035	41.579	74.140		0.035	33.470	74.106	
		42	0.019	30.302	54.030		0.019	23.620	52.297	
		47	0.010	17.853	31.834		0.010	15.368	34.026	

Relevant study in SG16

Q5/16	Q12/16	Q21/16	FG-VM	FG-AI4AD
Artificial intelligence-enabled multimedia applications	Intelligent visual systems and services	Multimedia framework, applications and services	Vehicular Multimedia	AI for autonomous and assisted driving

VCM
Video coding for machines

- VCM focus on the video/image/feature coding technology
- Supporting V2X, video surveillance, unmanned aerial vehicle, smart manufacturing applications related to machine vision including Q5, Q12, Q21, FG-VM, FG-AI4AD of ITU-T SG16.

DCM: Data coding for machine-intelligence

- **Work scope**
  - Applications oriented machine intelligence and human-machine intelligence
  - Representation and data coding for video, audio and other data information
  - Propose national standard suggestion
  - Encourage Chinese experts to participate in international standardization and improve international influences

2020.01 — Establish DCM standard group

2020.09 — Establish Requirement group

2020.11 — DCM White Paper

2021.04、7 — Establish 6 AHG
- Anchor generation
- Exploration of new technology

2021.11 — now

Thank You!

Q & A