Video Coding for Machines

Dr. Shan Liu

Distinguished Scientist and General Manager

Tencent Media Lab | Tencent Cloud

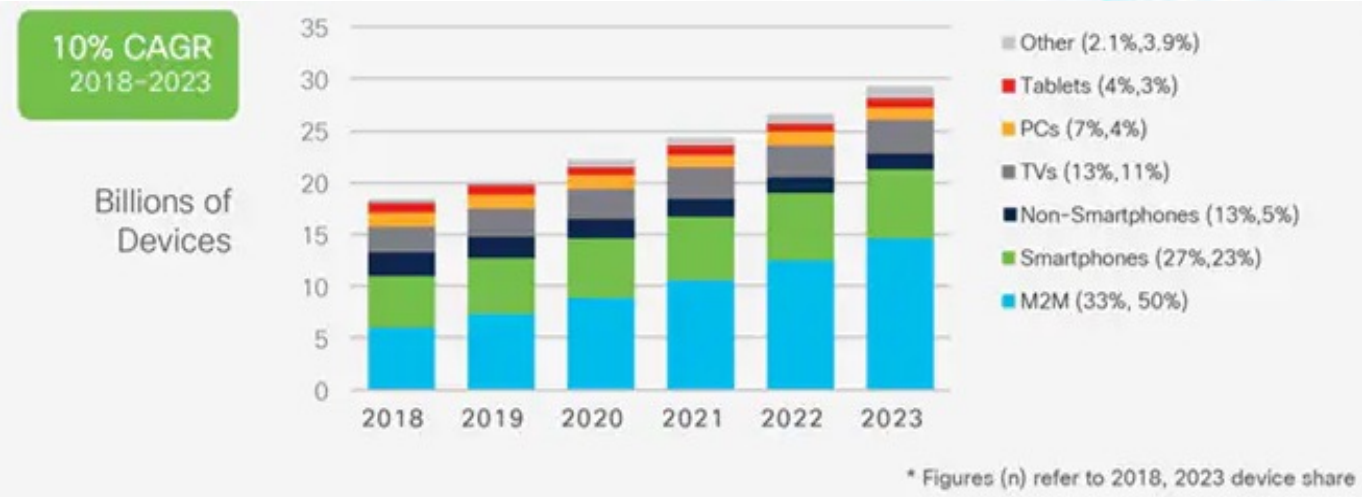
November 2021

Outline

- Why VCM?
- Overview of MPEG VCM
  - Use cases
  - Requirements
  - Framework
  - Evaluation Methodology
  - Anchor generation
  - Proposed technologies
- Future work plan

Why VCM?

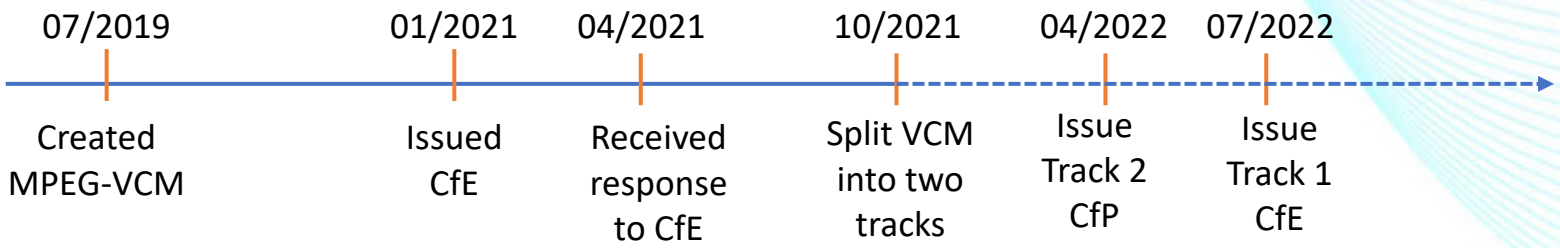
- Video has occupied a very large portion of internet traffic.
- Mand more video are consumed by machines.
  - Automation, analysis and intelligence without or with human intervention ➔ machine vision or hybrid vision
- Machine-to-Machine (M2M) devices and connections are fast growing.
- Machine vision 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

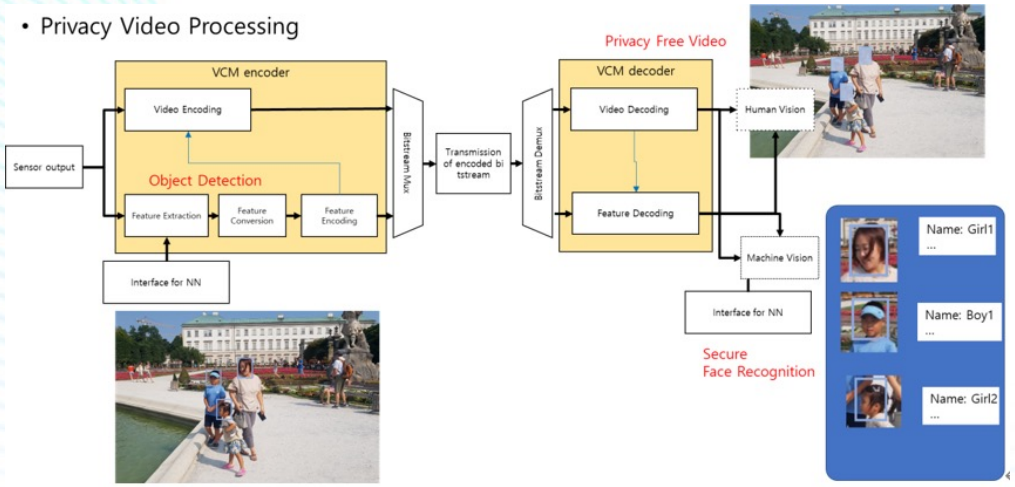
MPEG VCM

- ISO/IEC JTC1/SC29 WG2 committee created the VCM Ad-Hoc Group in July 2019 with the following mandates:
  - Define use cases and requirements for compression for machine vision and hybrid human/machine visions.
  - Collect dataset with ground truth and evaluation metrics.
  - Solicit technology evidence for video compression, feature extraction and feature compression.
  - Develop a framework to evaluate and compare different technology solutions.
  - Develop the standards for video coding for machines.



Use cases

- Typical Use cases
  - Surveillance
  - Intelligent transportation
  - Smart city
  - Intelligent industry
  - Intelligent content
  - Consumer electronics
- Typical machine vision tasks
  - Object detection and tracking
  - Instance segmentation
  - Event detection
  - Action recognition
  - …



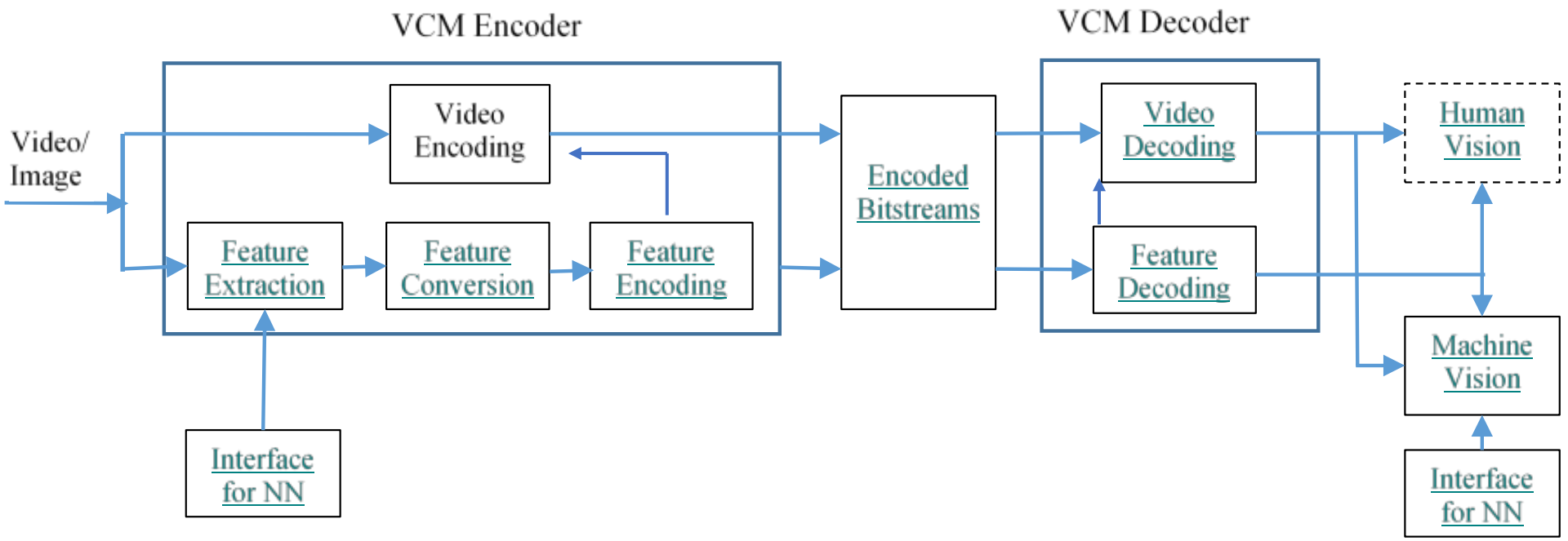
Source: mpeg output document: MDS19841

Source: GTI white paper – Intelligent transportation

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.

An Example of VCM Architecture



Source: mpeg output document: MDS20127

VCM Evaluation Methodology (1)

- 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.
  - Note that HiEve-10 is removed in the last meeting due to lack of details for anchor generations. Update for HiEve-10 is expected in the next meeting.

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

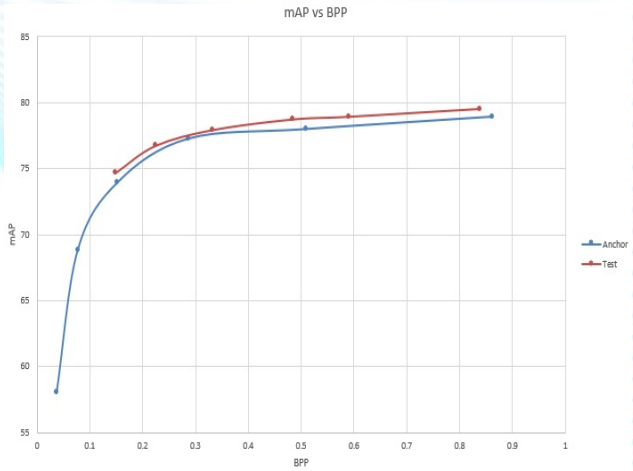
VCM Evaluation Methodology (2)

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

$$BPP = \frac{Total\ bitstream\ size\ in\ bits}{number\ of\ pixels\ in\ \ souce\ images}$$

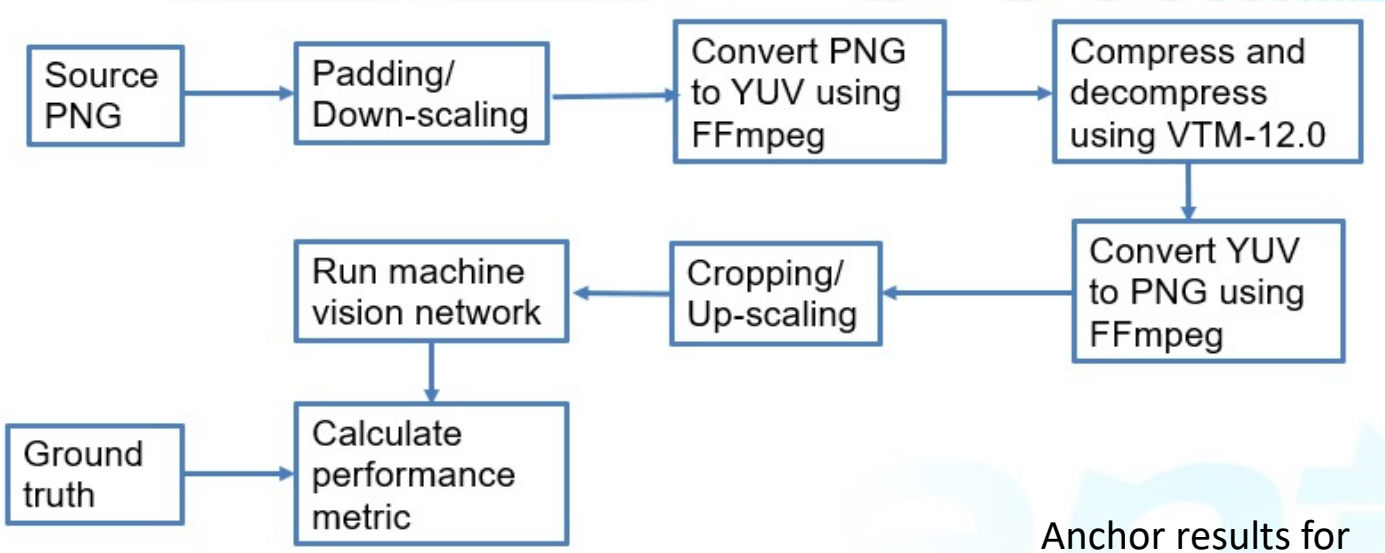
- Bitrate in kbps is used to measure bitstream cost for video dataset.
- BD-rate and BD-mAP/BD-MOTA/BD-fmAP are used to compare a proposed solution to the anchor solution for a single task.
- Note that metric to measure performance for Hybrid vision or multiple machine vision tasks are note yet decided.
- Excel template is used to compute metrics. An example is shown as following.

			Reference: VCM Anchor (VTM-12.0)		Test: tested		BD-rate	BD-mAP
Scale	Dataset	QPISlice	BPP	mAP	BPP	mAP	mAP	
100%	OpenImageV6	22	0.863	78.929	0.839	79.504	-18.00%	0.58
		27	0.509	77.989	0.590	78.913		
		32	0.287	77.263	0.484	78.709		
		37	0.153	73.963	0.334	77.916		
		42	0.078	68.842	0.225	76.723		
		47	0.037	58.021	0.149	74.674		

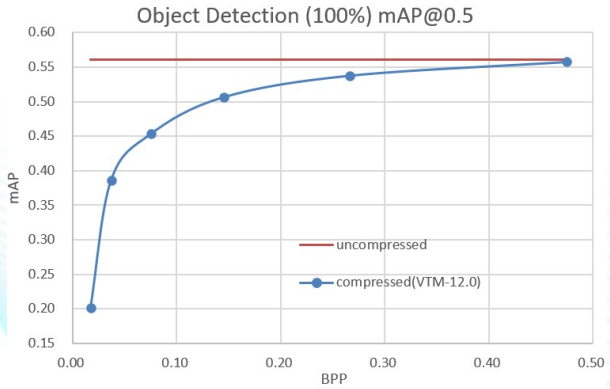


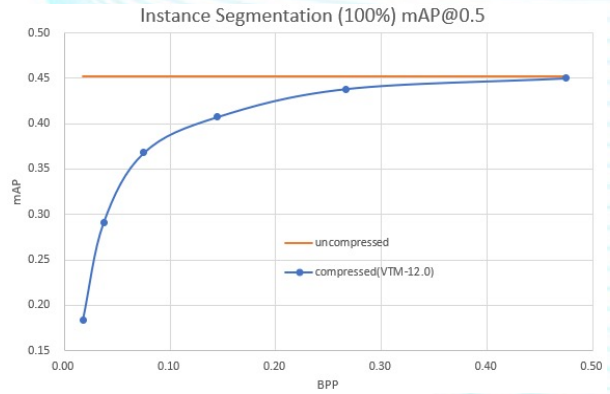
VCM Anchor Generation

- The state-of-art video codec, VVC, is adopted as the anchor solution for video compression.
  - Either 100% scale or pareto-front results for 100%, 75%, 50% and 25% are used.
- Pipeline for anchor generations is as following:



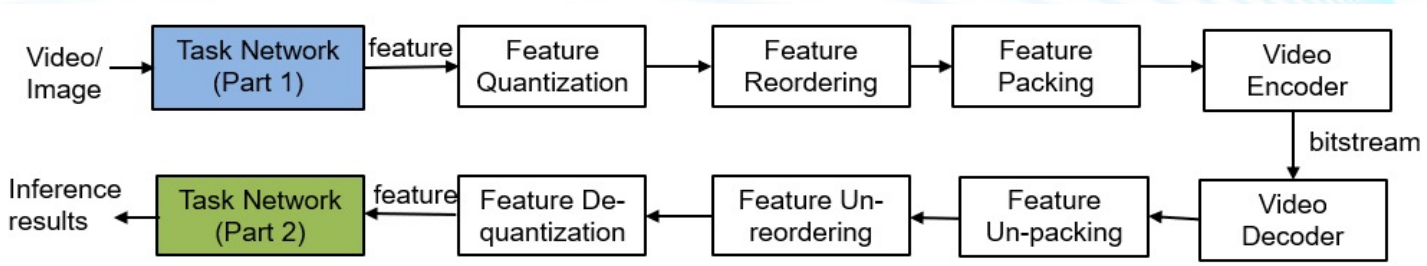
Anchor results for TVD dataset



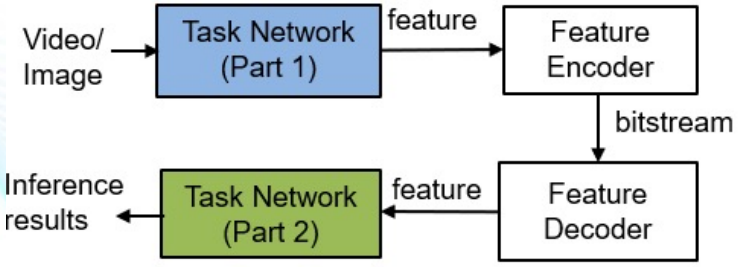


Proposed Technologies

- The proposed technologies can be classified into two categories
  - Category 1 (Track 1): Feature (map) coding
    - (1a) Encode Features as images/video
    - (1b) Encode features directly

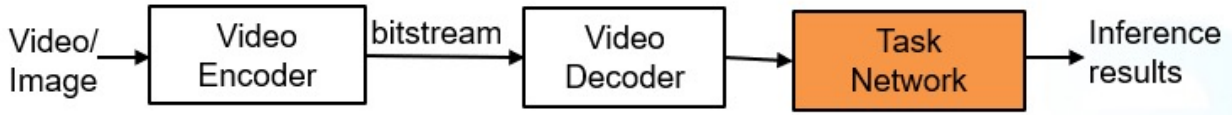


Category 1a



Category 1b

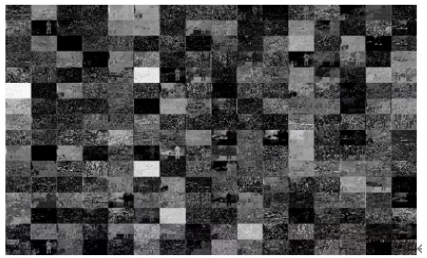
  - Category 2 (Track 2): Image/video coding



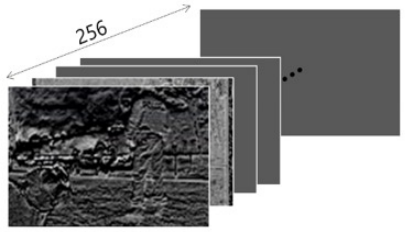
Category 2

Category 1: Feature Map Coding (1)

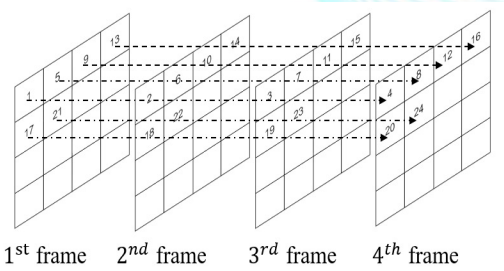
- Diverse contributions are proposed, make it harder to compare solutions
  - Different task networks
    - Mask R-CNN with Resnet-50, YOLOv3,Faster R-CNN with ResNeXt101-FPN, Mobile Netv2 + YOLOv5, etc
  - Different partition of task networks
    - The output of the stem layer, the output of whole backbone network, etc.
  - Different quantization and normalization scheme
    - Uniform or non-uniform quantization with different bit-width
    - Normalization using mean/standard division or using the max/min range
  - Different feature reordering/packing
    - Spatial packing, temporal packing, or multi-frame packing



spatial packing



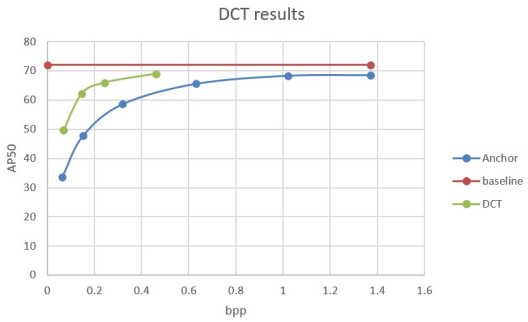
temporal packing



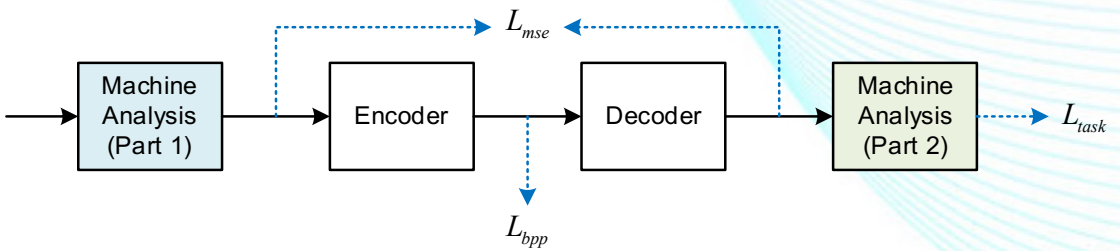
multi-frame packing

Category 1: Feature Map Coding (2)

- Category 1a: Packed features are coded using video codec, such as
  - VVC or HEVC codec
  - VVC + Deep CABAC with PCA transformation for feature data reduction
  - Resulted bitstreams are much larger than those from VCM anchor solution
- Category 1b: Encode features directly
  - Vector quantization + entropy coding
  - DCT/DWT + quantization + entropy coding
  - End-to-end trained feature compression, which achieves close performance as coding images using VTM codec



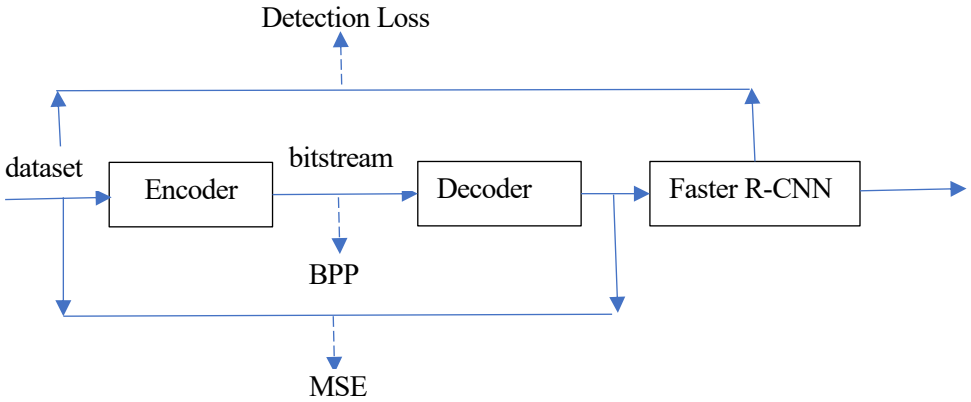
Source: mpeg document: M58000



Source: mpeg document: M58033

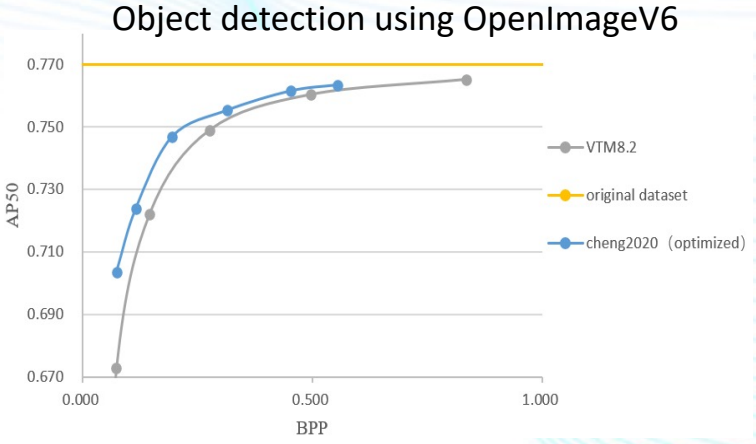
Category 2: Image/Video Coding (1)

- End-to-end trained Learning based image compression
  - Image compression network: Cheng2020, bmshj2018_hyperprior, or modified mbt2018-mean network
  - Joint training with VCM object detection network in which its parameters are fixed
- MS-SSIM optimized Cheng2020 network (source: M58050)
  - Can achieve BD-rate gain 23.56% for instance segmentation using OpenImageV6 dataset



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

End-to-end training framework

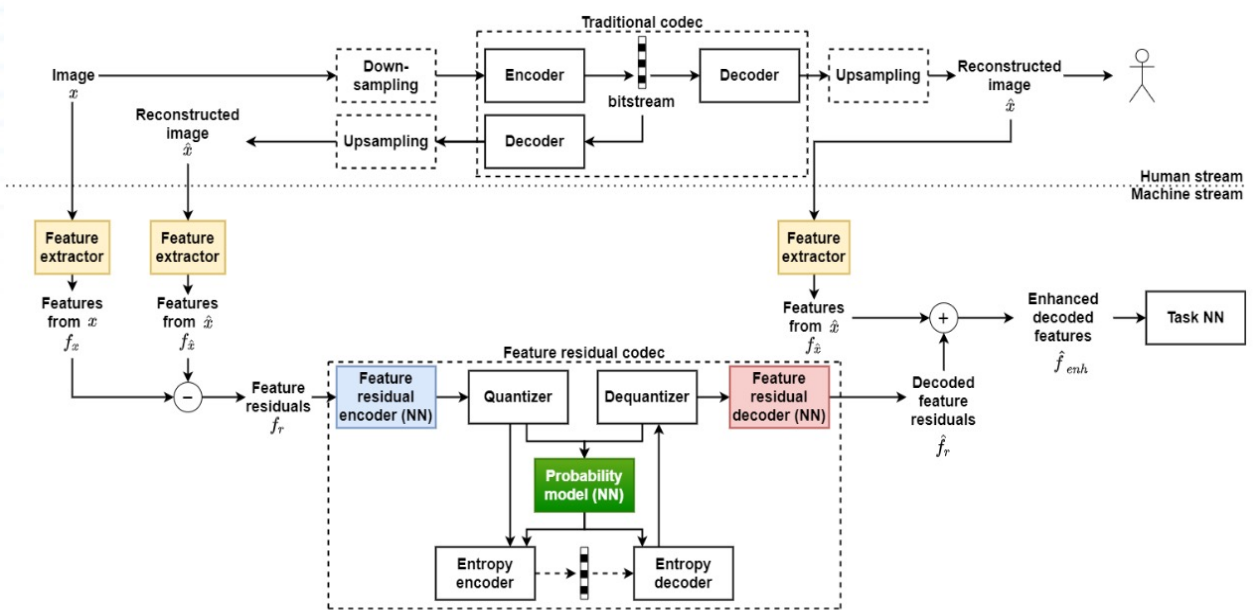


Performance of E2E trained Cheng2020 network: BD-rate -22.80%

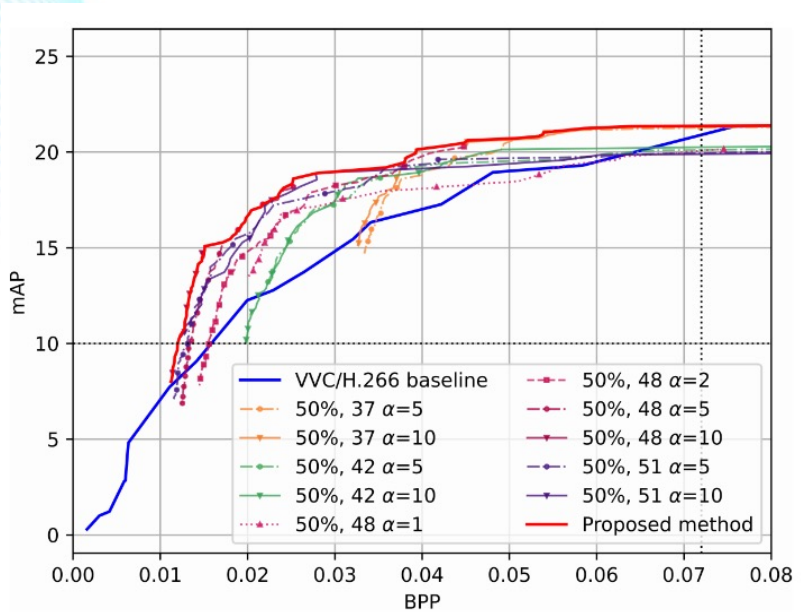
Source: mpeg document: M56445

Category 2: Image/Video Coding (2)

- 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%



System Architecture



Source: mpeg document: M58072

Category 2: Image/Video Coding (3)

- Region adaptive coding (source: M56572)
  - For a given image, two images are generated using the bounding boxes generated from VCM object detection network.
    - Foreground image: contains only foreground object
    - Background image: the rest of the image without foreground object
  - Both images are selectively scaled and coded using VVC codec
    - Foreground image is coded with relatively high QP.
  - This solution achieve 30.76% BD rate gain for object detection using FLIR dataset, compared to the anchor.

Two-Track Work Plan

- In October 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

Q & A

shanl@tencent.com