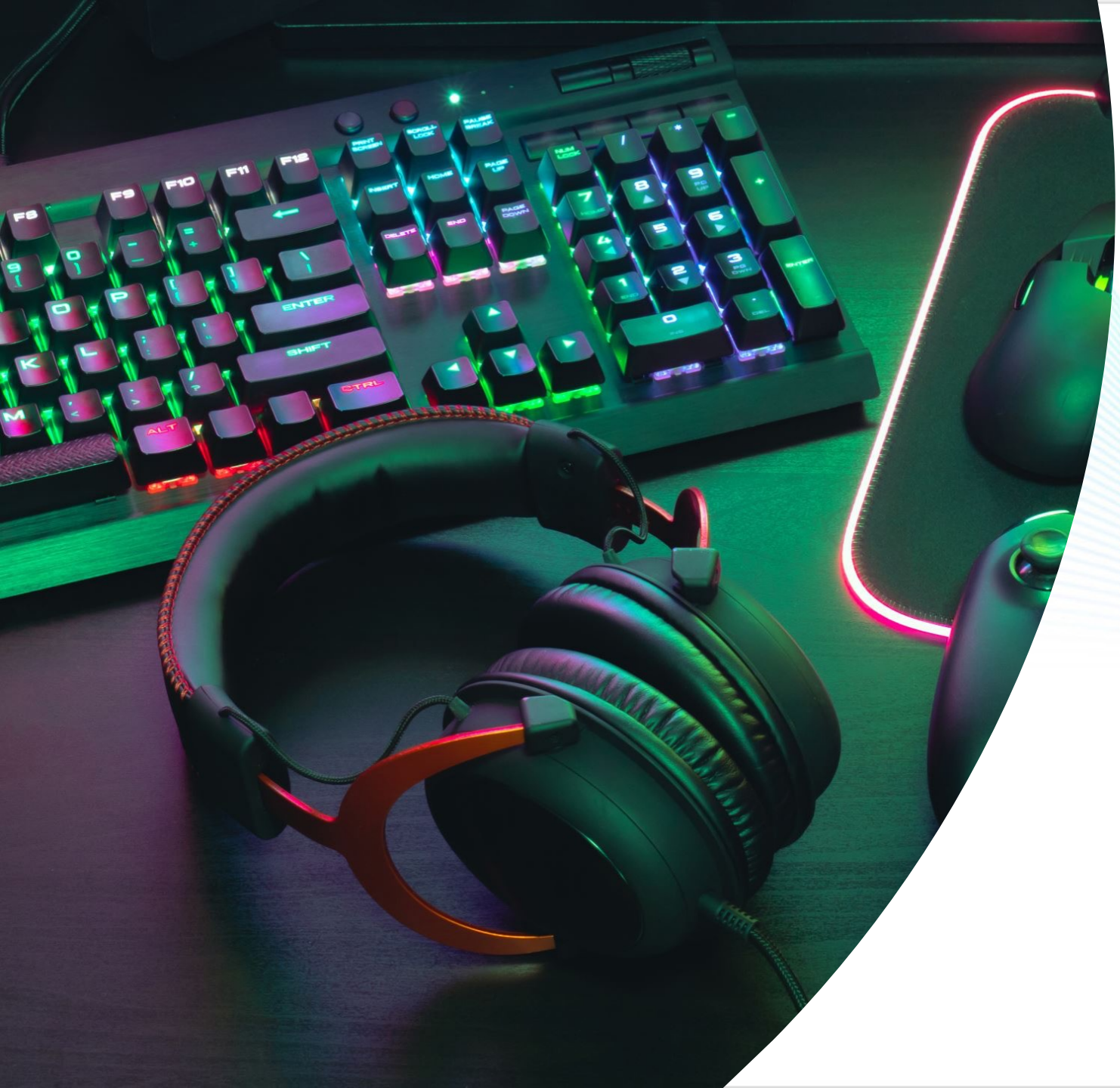


Updates on Video Coding for Machines

VCM AHG Chairs
ISO/IEC JTC 1/SC 29/WG2
2022.07.20

Outline

Motivation of MPEG VCM

Review and current status

Mandates and timeline

Use cases and Requirements

Framework

Evaluation methodologies

Common test conditions

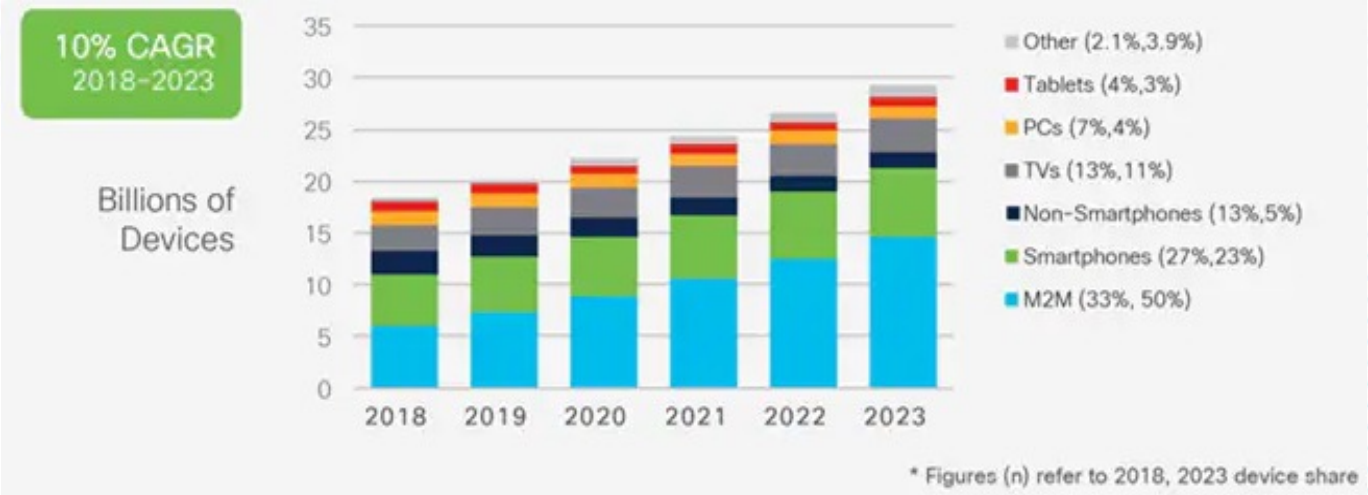
Anchor generation

Proposed technologies

Next plan

Why VCM?

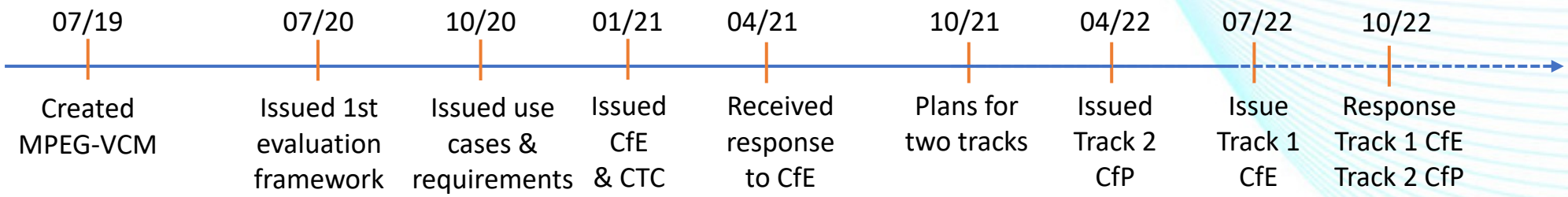
- Video has occupied a very large portion of internet traffic.
- Mand more video are consumed by machines
  - Automation, analysis and intelligence without or with human intervention ➔ machine vision or hybrid vision.
- Machine-to-Machine (M2M) devices and connections are fast growing.
- Machine vision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

- 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.
- More than 100 experts from around 30 institutes worldwide have participated in MPEG VCM AHG activities.



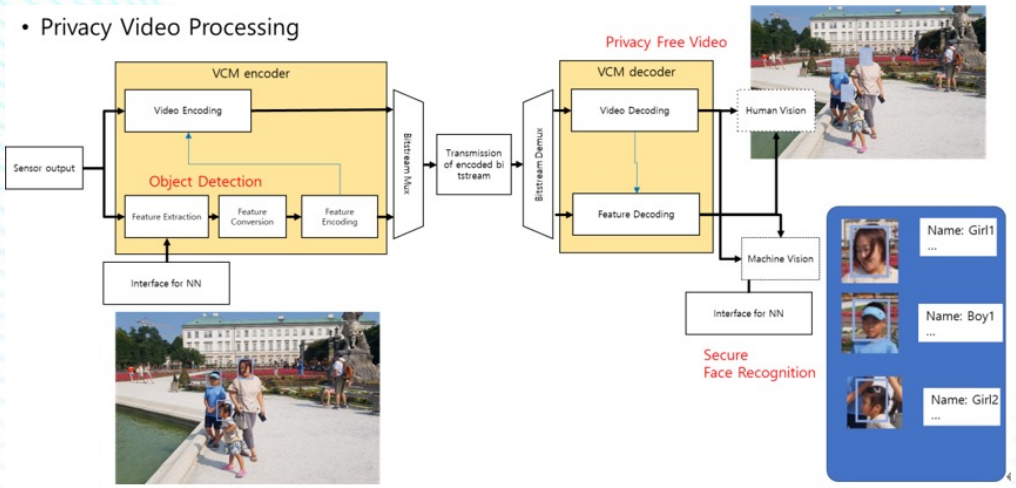
Use Cases

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.

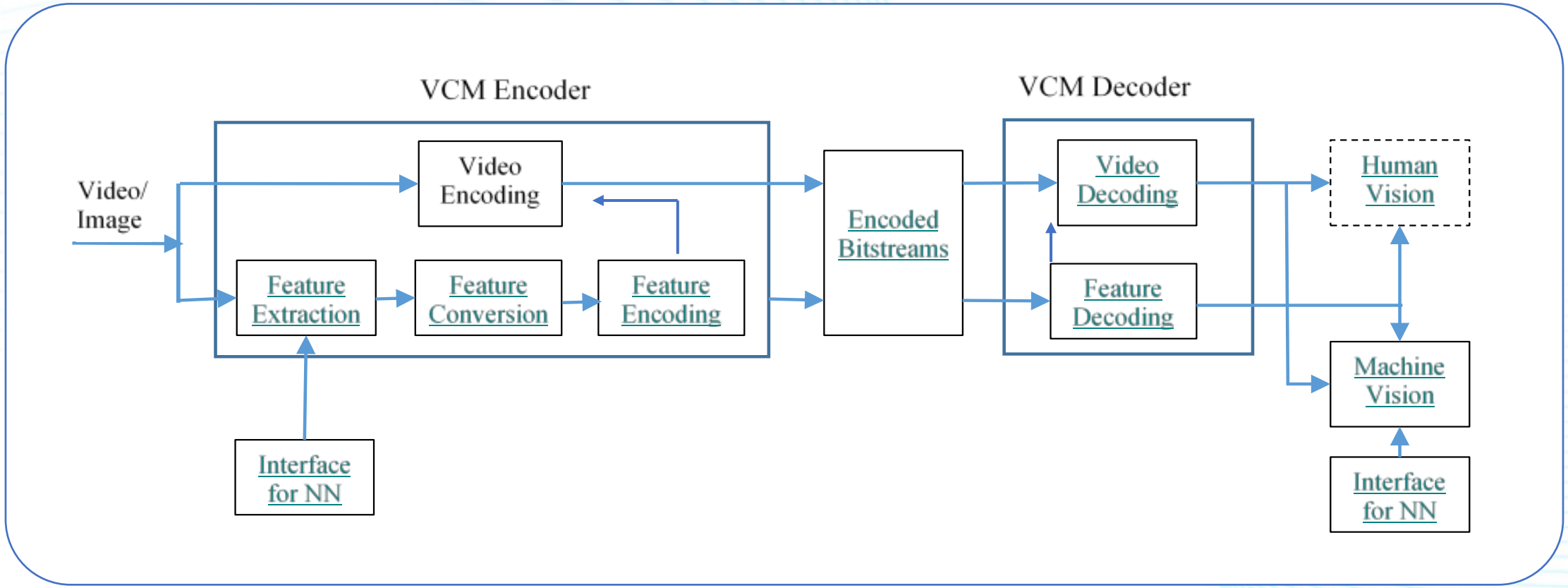
Feature extraction

- 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)

- Three machine vision tasks are selected to cover the main tasks identified in the use cases:
  - Object detection, instance segmentation and Object tracking.
- Four 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~~

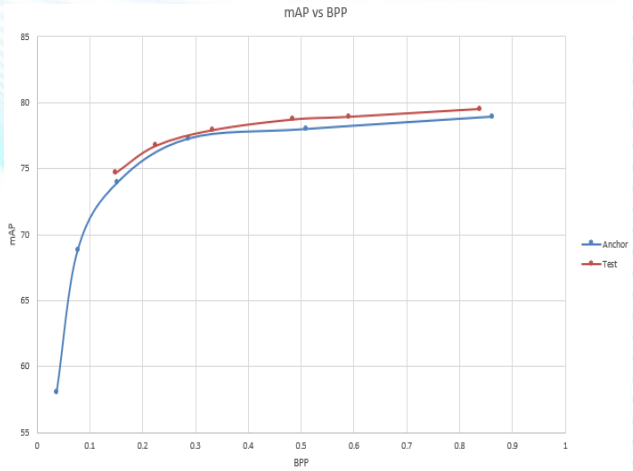
VCM Evaluation Methodology (2)

- 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 CTC

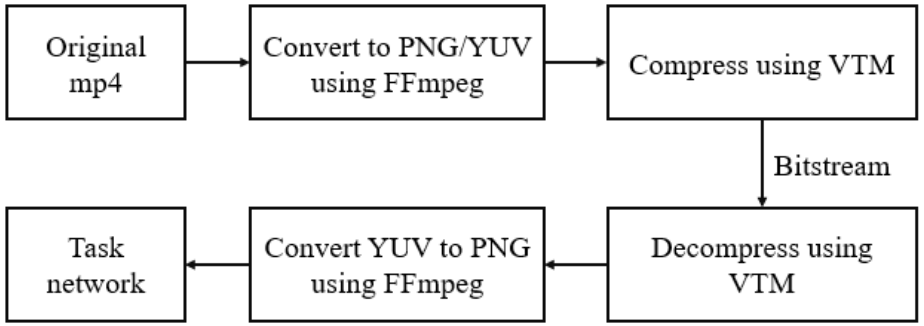
- VVC (VTM12.0) is used to produce compression anchor.
- Test conditions:

Machine Task	Evaluation Dataset	Source	Configuration	QP
Object detection	OpenImageV6	https://storage.googleapis.com/openimages/web/index.html	AI	{22, 27, 32, 37, 42, 47}
	FLIR (IR)	https://www.flir.com/oem/adas/adas-dataset-form/	AI	{22, 27, 32, 37, 42, 47}
	TVD (image)	https://multimedia.tencent.com/resources/tvd	AI	{22, 27, 32, 37, 42, 47}
Instance segmentation	OpenImageV6	https://storage.googleapis.com/openimages/web/index.html	AI	{22, 27, 32, 37, 42, 47}
	TVD (image)	https://multimedia.tencent.com/resources/tvd	AI	{22, 27, 32, 37, 42, 47}
Object tracking	TVD (video)	https://multimedia.tencent.com/resources/tvd	RA	Refer to Appendix A
Video Object Detection	SFU-HW-Objects-v1	Video: ftp://hevc@mpeg.tnt.uni-hannover.de/testsequences/ Label: https://dx.doi.org/10.25314/7d8efc0a-3943-4738-b7a5-72badb04d765	RA	Refer to Appendix A

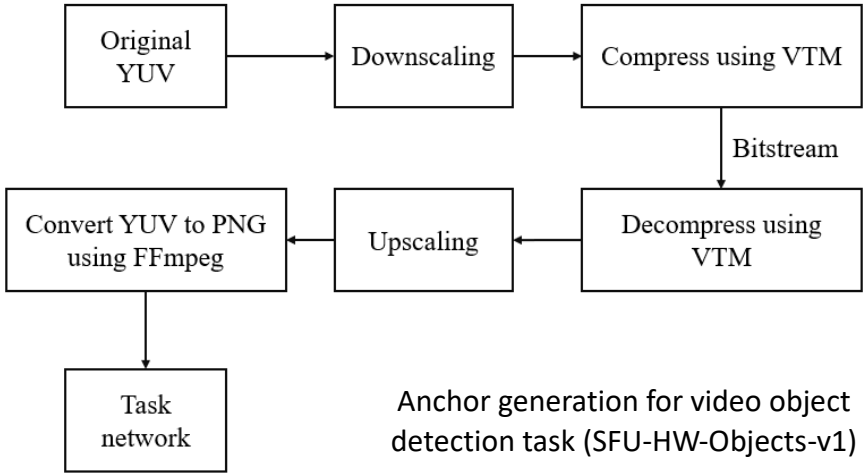
- Recommended training sources:

Database	Location	Access credentials
OpenImageV6-train*	https://storage.googleapis.com/openimages/web/index.html	Public
TVD-train+	https://multimedia.tencent.com/resources/tvd	Public

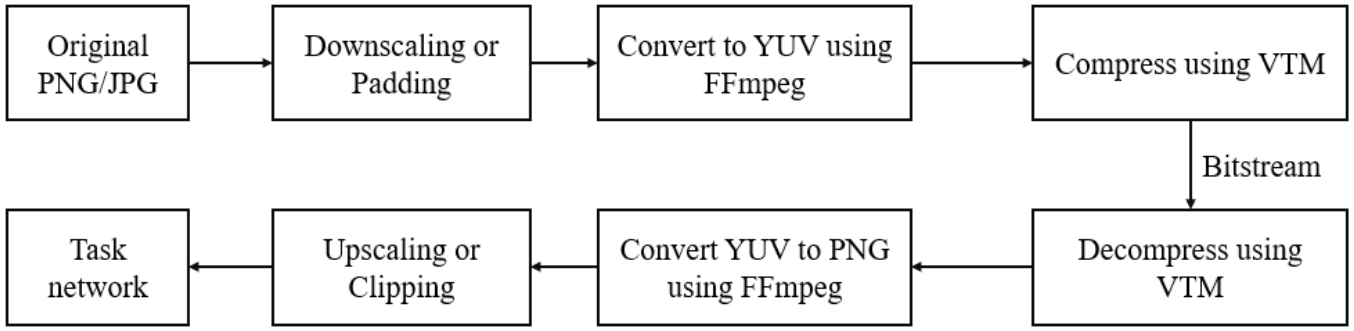
VCM Anchor Generation



Anchor generation for video object tracking task (TVD (Video))



Anchor generation for video object detection task (SFU-HW-Objects-v1)



Anchor generation for object detection and instance segmentation tasks

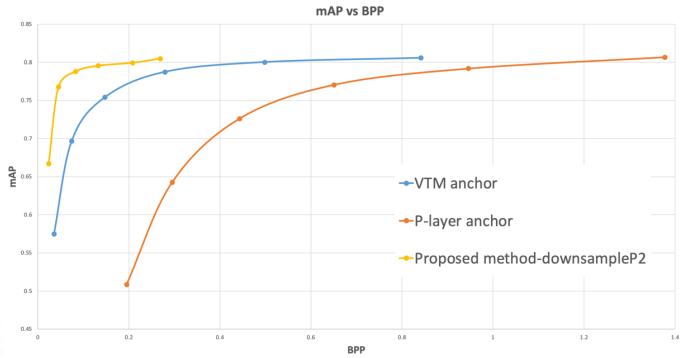
Object detection: OpenImageV6
Object detection: FLIR
Object detection: TVD (Image)
Instance segmentation: OpenImageV6
Instance segmentation: TVD (Image)

VCM Reporting Templates

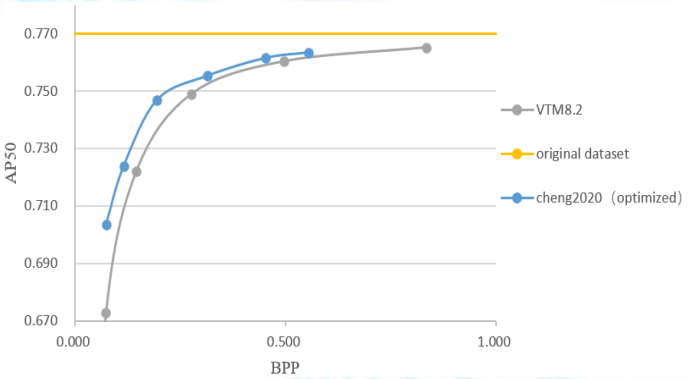
Object Detection	End-to-End BD-Rate [%]					End-to-End BD-mAP		
	mAP	Pareto mAP	Luma	Chroma Cb	Chroma Cr		EncT	DecT
100% Scale	100.00%	--	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
75% Scale (optional)	100.00%	--	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
50% Scale (optional)	100.00%	--	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
25% Scale (optional)	100.00%	--	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
OpenImageV6 Average	100.00%	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
FLIR Average	100.00%	#VALUE!	100.00%	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
TVD Average	100.00%	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
SFU-HW-Objects-v1 Average	100.00%	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
Overall Average	100.00%	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#NUM!	#NUM!

Instance Segmentation	End-to-End BD-Rate [%]					End-to-End BD-mAP		
	mAP	Pareto mAP	Luma	Chroma Cb	Chroma Cr		EncT	DecT
100% Scale	100.00%	--	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
75% Scale (optional)	100.00%	--	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
50% Scale (optional)	100.00%	--	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
25% Scale (optional)	100.00%	--	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
OpenImageV6 Average	100.00%	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
TVD Average	100.00%	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
Overall Average	100.00%	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#NUM!	#NUM!

Object Tracking	End-to-End BD-Rate [%]					End-to-End BD-MOTA		
	MOTA	Pareto MOTA	Luma	Chroma Cb	Chroma Cr		EncT	DecT
100% Scale	100.00%	--	100.00%	100.00%	100.00%	#VALUE!	#DIV/0!	#DIV/0!
Scale1 (optional)	#VALUE!	--	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
Scale2 (optional)	#VALUE!	--	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
Scale3 (optional)	#VALUE!	--	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#DIV/0!	#DIV/0!
TVD Average	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	0.00%	0.00%
Overall Average	#VALUE!	#DIV/0!	#VALUE!	#VALUE!	#VALUE!	#VALUE!	#NUM!	#NUM!



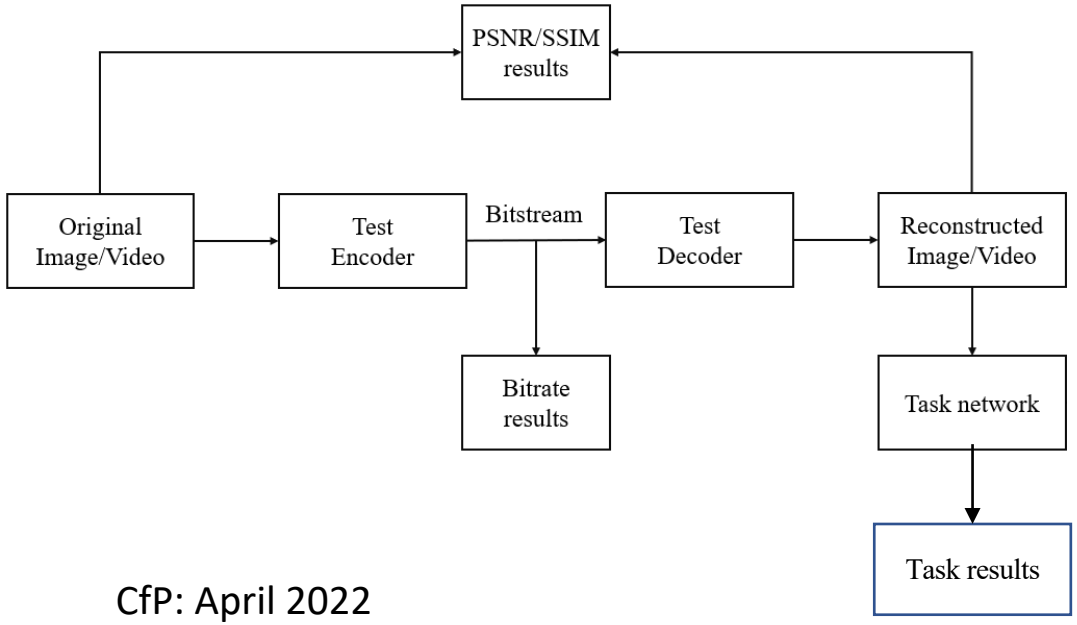
Source: mpeg document: M60240 for segmentation (60+% saving)



Source: mpeg document: M56445 for detection (20+% saving)

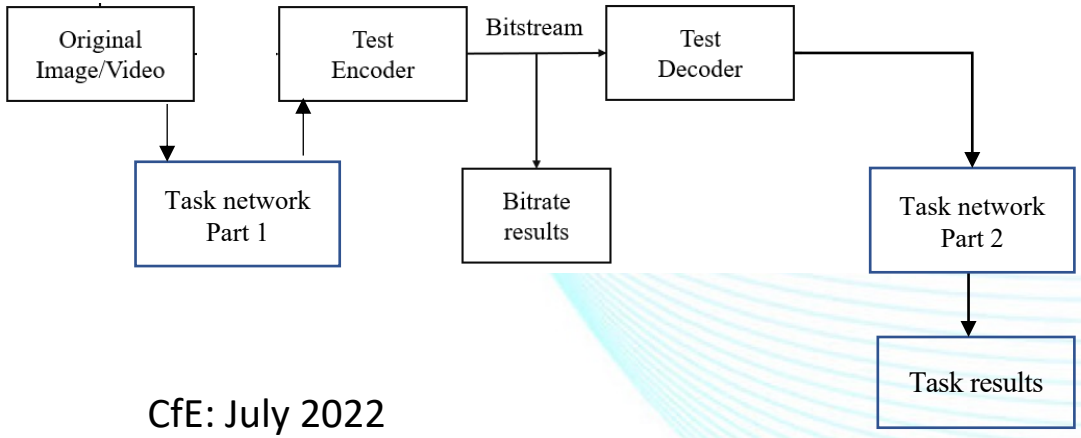
Proposed Technologies

Image/Video Coding (Track 2)



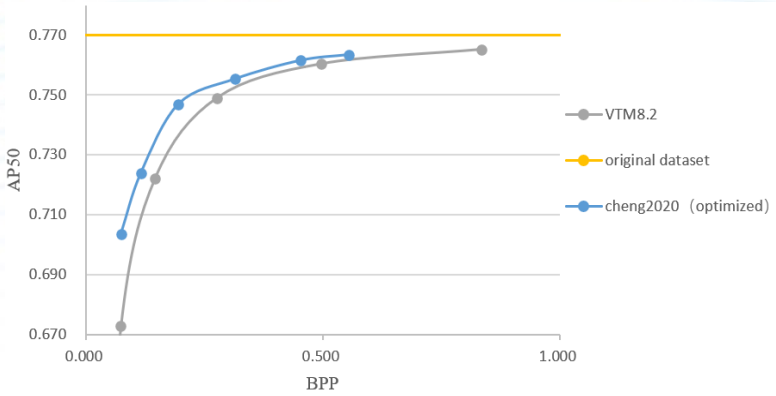
CfP: April 2022
CfP Response: October 2022

Feature (map) Coding (Track 1)



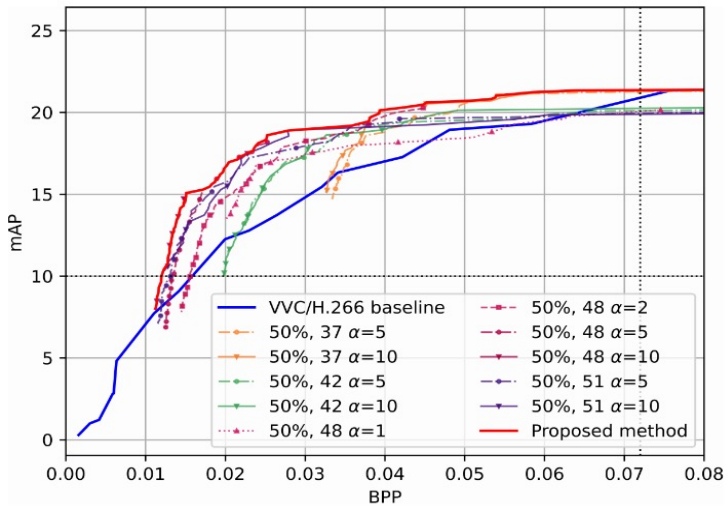
CfE: July 2022
CfE Response: October 2022

Examples of Proposed Technologies



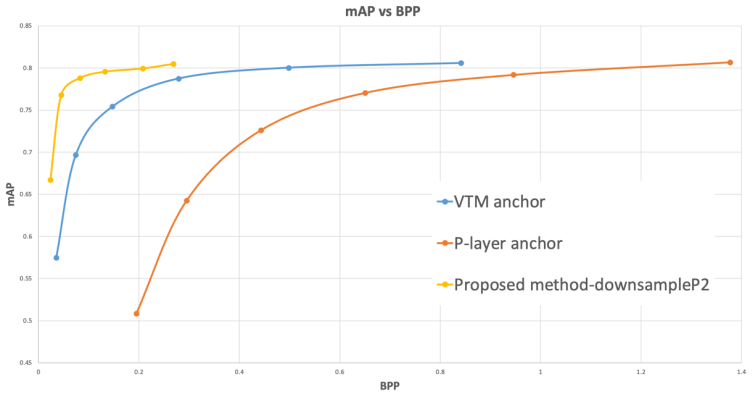
Source: mpeg document: M56445, for detection (20+% saving)

2021.04, image/video coding (e2e nn)



Source: mpeg document: M58072, for detection (40+% saving)

2021.10, image/video coding (hybrid)



Source: mpeg document M60240, for segmentation (60+% saving)

2022.07, feature coding (e2e nn)

Thank You