Special Paper
방송공학회논문지 제28권 제7호, 2023년 12월 (JBE Vol. 28, No. 7, December 2023)
https://doi.org/10.5909/JBE.2023.28.7.888
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

Exploring the Video Coding for Machines Standard: Current Status and Future Directions

Dongmin Lee[a)], Sangkyun Jeon[a)], Yeonghun Jeong[a)], Joonsoo Kim[b)], and Jeongil Seo[a)‡]

Abstract

Nowadays, as the evolving deep learning and machine vision technology are collaborating, machine vision accuracy is greatly advanced. Accordingly, machines' demand for image/video processing is increasing in various fields, such as autonomous driving, surveillance systems, and smart cities. However, traditional video coding is optimized only for the human visual system, not machine vision. To solve this inefficiency problem in machine vision system caused by traditional video coding methods, MPEG is currently progressing with the standardization of a new video coding technology optimized not for human visual systems, but for machine vision systems. This paper examines the current status of VCM (Video Coding for Machine) standardization in MPEG.

Keyword : Video coding for machines, image and video coding, feature coding, MPEG standards.

Ⅰ. Introduction

Machine vision is the ability of a computer to see. It is a technology where hardware and software systems of machines take on the roles of human visual recognition and decision-making functions. Recently, with the application of deep learning-based technologies in machine vision, it has surpassed the accuracy of previous methods and the range of humans. Based on those advances, machine vision is gradually expanding beyond the existing application scope of industrial automation to autonomous vehicles, smart cities, video surveillance, security, and safety. Furthermore, with the proliferation of demand for video data analysis because of changes in mobile devices, telecommunication technologies, and social media, demand for video processing through machines is also expected to increase.

In an age where more than 80% of internet traffic is made up of video and image data, there is a growing need to improve coding efficiency to help reduce this data load. However, traditional video coding methods have been primarily optimized for delivering optimal video quality to

a) Dept. of Computer Engineering, Dong-A University
b) Electronics and Telecommunications Research Institute, ETRI
‡ Corresponding Author : Jeongil Seo
E-mail: jeongilseo@dau.ac.kr
Tel: +82-2-200-7796
ORCID: https://orcid.org/0000-0001-5131-0939
※ This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00011, Video Coding for Machine, No.2023-0-00076, Software-Centered University (Dong-A University)).
· Manuscript October 7, 2023; Revised November 1, 2023; Accepted November 1, 2023

human viewers. In other words, being optimized for human viewers means that current video encoding methods may not be optimal from a machine's point of view, which means that video coding methods optimized for the machine's position are needed. The emergence of vehicle-to-vehicle connections, IoT devices, large-scale video surveillance networks, smart cities, and quality inspections is leading this new paradigm. These domains impose stringent requirements in terms of latency and scale and require image and video coding solutions aimed at machine vision.

The requirements for this machine vision have influenced new research directions and approaches that are different from previous ones. For instance, recent advancements in deep learning technologies for various classification and regression tasks promote research of proper compression representation and compact features.

Moving Picture Experts Group (MPEG), a group that sets video coding standards, is thinking about a new way of video coding called 'Video Coding for Machines (VCM)'. This paper will explain what's happening with VCM and what might come next[1].

II. Review of VCM Related Standards

In MPEG-7, techical standards analogous to the objectives of VCM were established[2]. They include Compact Descriptors for Visual Search (CDVS) in MPEG-7 Part 13 and its successor, Compact Descriptors for Video Analysis (CDVA) in MPEG-7 Part 15.

1. Compact Descriptors for Visual Search

CDVS is a standard established in 2015, which suggests a normative bitstream of compact visual descriptors standardized for mobile visual search and augmented reality applications[3]. CDVS makes a normative feature extraction pipeline, including interest point detection, local feature selection, local feature description, local feature descriptor aggregation, local feature descriptor compression, and local feature location compression.

The successful standardization of CDVS has notably influenced computer vision algorithms. However, in more intricate analytical tasks such as autonomous driving and video surveillance, these standardized approaches have demonstrated inferior performance compared to end-to-end frameworks.

2. Compact Descriptors for Visual Analysis

CDVA, the successor to CDVS, is a standard established in 2019 that responds to the explosive demand for video analysis in autonomous driving, video surveillance systems, entertainment, and smart cities[4]. CDVA was initiated in Feb. 2015 to have a norm of neural network-based video feature descriptors for machine vision tasks. CDVA includes Nested Invariance Pooling (NIP) and Adaptive Binary Arithmetic Coding (ABAC) as a key technology. In particular, compared to existing methods, the combination of CDVA's NIP descriptors and CDVS's descriptors made remarkable performance.

VCM needs both traditional image/video coding and feature compression methods for human and machine vision. VCM uses HEVC/VVC standards and CDVS/CDVA's ideas to satisfy those visions. Like HEVC/VVC standards, the encoder generates a compressed bitstream from the received video input, and the decoder reconstructs video from the bitstream that the encoder generated. And like CDVS/CDVA, features are extracted from received video input and generate efficient feature descriptors.

III. VCM Standard History

Video Coding for Machines (VCM) aims to enable machines to understand and perform tasks without human intervention. VCM is the technology aimed at compressing

video data to the maximum extent while maintaining the performance required for machines to carry out their tasks effectively. Discussions on standardization of these VCM technology is currently underway in MPEG. MPEG has divided the standardization process of VCM into two tracks: VCM Track 1: Feature coding and VCM Track 2: Image and video coding. Before delving into the activities and future directions of Track 1 and Track 2, let's first understand the scope of VCM standardization as defined by MPEG. The scope of VCM includes its use cases, requirements, performance evaluation metrics, and a broad overview of the standardization progress.

1. Scope of VCM standardization and use cases

As defined by MPEG, the most recent scope of standardization, established in April 2022, is "MPEG-VCM aims to define a bitstream from encoding video, descriptors, or features extracted from video that is efficient in terms of bitrate and performance of a machine task after decoding[5]".

As defined in the standard range, the input and output of the VCM may be in various forms, such as Video, Feature, Descriptors, and more, all of which are necessary for performing machine vision tasks. Examples of VCM structure according to these various input and output combinations are shown in Fig. 1[5].

Also MPEG has identified seven representative use cases for VCM, as illustrated in Fig. 2[1]. The first is the area of surveillance. In modern video surveillance systems, artificial neural networks are employed for tasks like object detection and tracking, and the number of cameras used in the system is exploding, requiring a lot of bandwidth and computational performance to transmit videos obtained from cameras and process them through neural networks.

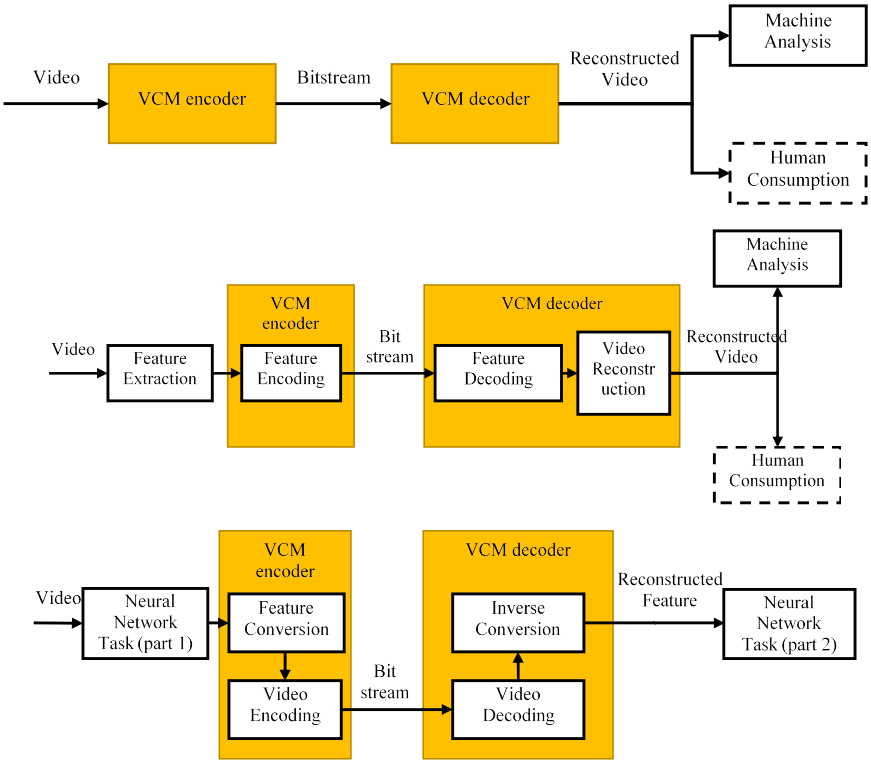

Fig. 1. Example of VCM Structure

Fig. 2. Use cases of VCM

The second is the intelligent transportation system. In intelligent transportation systems, cars must communicate between cars or with other sensors to perform their duties, and many of these communication targets include video. Given that automobiles are the end-users of the transmitted and received videos, the implementation of an appropriate encoding method is essential.

The third is the smart city. In the case of smart cities, a high level of interconnection is made between different node sensors and devices for smart city applications, such as traffic monitoring, density detection and prediction, traffic flow prediction, and resource allocation. Efficient communication between devices and encoding methods that transmit only the necessary information is crucial.

The fourth is the intelligent industrial system. In automated production environments, continuous video transmission and analysis for product defect inspection require efficient encoding methods.

The fifth is the field of intelligent content. Due to the recent development of mobile technology, a huge amount of image/video content is generated. Among them, protecting specific groups from inappropriate content is a significant concern. Efficient encoding methods using machine vision technology, are needed to process live images/videos, short videos, and social media content.

The sixth is consumer electronics. The use of neural networks in consumer products for tasks like situation awareness and providing information to users is increasing. Efficient encoding methods are required to reduce network bandwidth for communication and transmission to external servers.

The seventh is Multi-Tasking with Descriptors. One video stream can serve various machine vision tasks. These tasks may be executed in parallel, sequentially, or hybrid manner. At this time, if a common machine vision task is performed before transmission and the output is transmitted to the input of other machine vision tasks as a descriptor, the bandwidth required for video transmission and the calculation time of the machine task after decoding can be greatly reduced. Therefore, efficient encoding methods are crucial when transmitting descriptors, videos, or both simultaneously.

2. Requirements and Evaluation metrics

Through several standardization meetings, MPEG has deliberated on the requirements for the standardization of VCM, culminating in the establishment of 17 specific requirements as presented in Table 1[5], finalized in April 2022. Among the 17 requirements, 16 are categorized as mandatory requirements (shall), constituting the majority. These requirements can be broadly divided into four categories: general requirements for VCM, feature coding requirements (Track 1), video coding requirements (Track 2) and requirements applicable to both feature and video coding.

Table 1. VCM Requirements

No.	Requirements
1	VCM shall support video coding for machine task consumption purposes.
2	VCM shall support feature coding.(Feature coding)
3	VCM shall support a coding efficiency improvement for at least 30% BD-rate over the VVC standard on machine vision tasks.(Video/Feature coding)
4	VCM shall support a broad spectrum of encoding rates.
5	VCM shall support various degrees of delay configuration.(Video coding)
6	VCM shall be agnostic to network models. (Video/Feature coding)
7	VCM shall be agnostic to machine task types. (Video/Feature coding)
8	VCM shall provide description of the meaning or the recommended way of using the decoded data. (Feature coding)
9	VCM should support the use and inclusion of information, such as descriptors in its bitstream.(Feature/Video coding)
10	A single VCM bitstream shall support any number of instances of machine tasks. (Video/Feature coding)
11	VCM shall support at least the following colour formats; monochrome, RGB, and YUV (YCbCr).(Video coding)
12	VCM shall support at least the following input bit depths: 8-bit and 10-bit.(Video coding)
13	VCM shall allow for feasible implementation within the constraints of the available technology at the expected time of usage.
14	VCM shall support rectangular picture format up to 7680x4320 pixels (8K).
15	VCM shall support fixed and variable rational frame rates for video inputs.
16	VCM shall support any input source from video or image.
17	VCM shall support privacy and security.

Performance evaluation metrics for VCM are defined in the document of Common Test Conditions and Evaluation Methodology for Video Coding for Machines[6]. The performance evaluation metrics used for assessing VCM encoders are determined based on the specific machine vision tasks, typically conducted on identical or similar datasets, ensuring consistency in the evaluation process.

The mAP (mean AP) calculated by averaging AP (Average Precision) for each object class for a specific range of IoU (Intersection over Union) is used to evaluate the performance of object detection task.

TP is a true positive, FP is a false positive, TN is a true negative, FN is a false negative, and T_{IoU} is a Intersection over Union (IoU) threshold[6]. IoU means a matching ratio between the correct answer rectangular area and the prediction rectangular area[1].

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{1}$$

And $TP(T_{IOU})$, $FP(T_{IoU})$, $FN(T_{IoU})$, and $TN(T_{IoU})$ are defined with an T_{IoU} for that category, where true/false represents the output of the neural network, positive/negative represents the label in the ground truth[6]. The Precision refers to the ratio of correct answers to the total predicted results.

$$Precision(T_{IoU}) = \frac{TP(T_{IoU})}{TP(T_{IoU}) + FP(T_{IoU})} \tag{2}$$

VCM performance evaluation uses the average value of AP values (AP@[0.5:0.95]) obtained for every IoU value at 0.05 intervals from 0.5 to 0.95.

The performance evaluation metric for the task of image instance segmentation is mAP. However, while in object detection IoU (Intersection over Union) is calculated based on the box area containing the object, in image instance segmentation, binary maps are created according to the shape of the object, and calculations are performed using the overlap of these binary maps.

Multiple Object Tracking Accuracy (MOTA) is used as a performance evaluation index for object tracking. When g_t is called Ground Truth in the t-th frame, MOTA can be calculated as follows[6].

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + mme_t)}{\sum_t (g_t)} \quad (3)$$

FN_t, FP_t, mme_t, and g_t are the number of false negatives, the number of false positives, the number of mismatch error (ID Switching between 2 successive frames), and the number of objects in the ground truth respectively at time t[6]. MOTA is designed to provide a comprehensive assessment of tracking accuracy by considering all these error types.

Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Map (SSIM) values used in traditional video coding are used for video quality evaluations seen by humans, such as hybrid vision. However, VCM primarily focuses on optimizing video compression for machine tasks, not human perception, and does not typically use subjective quality assessment. In VCM, computational complexity, including encoding and decoding times, may be considered to assess practical feasibility and efficiency. VCM defines EncT and DecT. EncT (Encoding time) is the time needed to convert RGB input to bitstream, and DecT (Decoding time) is the time needed to convert bitstream to reconstructed RGB[6].

3. Overview of the standardization progress

In July 2019, during an MPEG meeting held in Gothenburg, Sweden, discussions commenced on the technology for compressing video data while maintaining machine vision performance, leading to the formation of the MPEG VCM AhG (Adhoc Group) in September 2019 to discuss a new video coding method called Video Coding for Machines.

In April 2021, during a meeting, a Call-for-Evidence (CfE) was conducted to verify the feasibility of video coding within the VCM standardization technology. Based on this, in April 2022, a Call-for-Proposal (CfP) for the Image/Video Coding track (Track 2) was issued, and technical proposals were evaluated in October. In July 2022, a CfE for the Feature Coding track (Track 1) was published, and technical evaluations were conducted in October. Then, in April 2023, during the 142nd meeting, a CfP for the Feature Coding track was released.

Ⅳ. VCM Track 1: Feature Coding

The Feature Coding track, as depicted in Fig. 3, involves the technology of receiving the output of a primary machine vision network from images or videos. It then encodes this output, which includes feature maps and data, as part of the encoding process.

In July 2022, MPEG VCM Track 1 issued a Call-for-Evidence (CfE)[7]. For Track 1, essential tasks were defined as object tracking and object segmentation, with object detection as optional. Researchers from Hanbat National

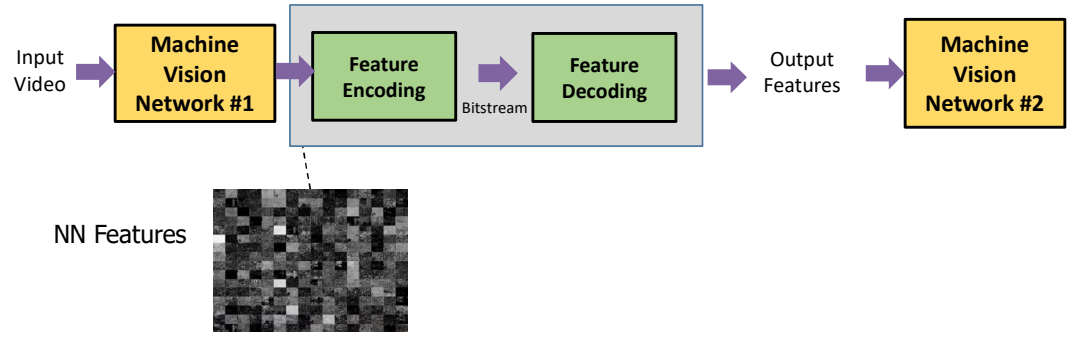

Fig. 3. Pipeline of Feature Coding

University and ETRI proposed a pipeline structure based on Multi-scale feature fusion (MSFF), Single-stream feature codec (SSFC), and Multi-scale feature compression (MSFC)[8]. There is a process of packing and quantizing the output of the features from the SSFC encoder to a format suitable for the input of the VVC encoder, and after performing SSFC decoding through format change, inverse quantization, and unpacking the VVC decoder's output can be restored to multiple layers of features to perform the task.

Korea Aerospace University and ETRI proposed a method of performing VVC encoding by configuring MSFC-transformed features by arranging the output of the features after MSFF in order of importance based on the MSFC framework[9]. The MSFC-transformed features are adaptively selected for compression based on the bitrate used by the VVC encoder. During the decoding process, these features are restored using a method that predicts the removed performance.

Canon proposed a method of compressing features by reducing the dimension of features obtained from MSFC using PCA[10]. Kwangwoon University and ETRI proposed a conversion-based feature compression technology that uses PCA techniques to obtain basis and average in advance for encoded features and used in codecs[11].

Tencent and Wuhan University proposed improving encoding efficiency using the Learning-based Feature Conversion module[12]. For the Feature Encoding/Decoding

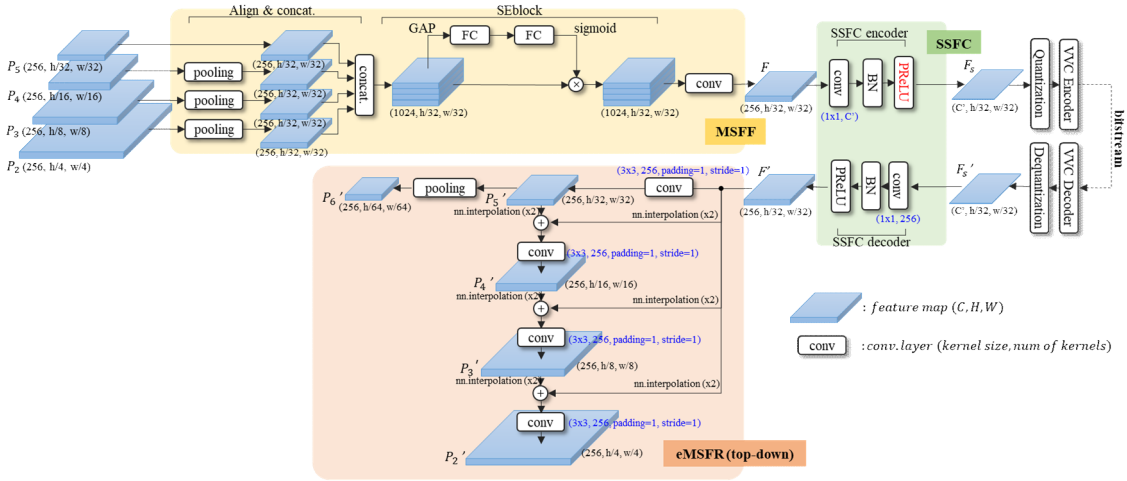

Fig. 4. VCM Feature Codec based on MSFC

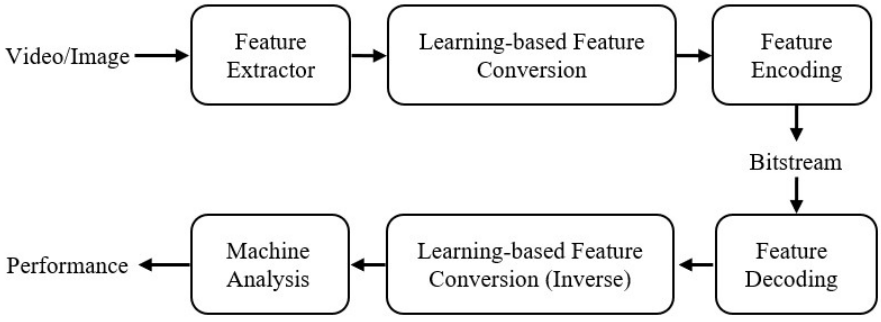

Fig. 5. Learning-based feature compression framework

of images, they employed a deep neural network-based encoding network founded in Cheng2020[13]. For encoding features derived from videos, they utilized the VVC encoder.

The results of the CfE proposal are shown in Table 2 [14]. Performance improvements of 97.58% in object tracking, 98.60% in object segmentation, and 98.34% in object detection were found compared to feature anchors created using a VVC encoder. On the other hand, compared to the VVC anchor, performance improvement was found to be 87.44% in object tracking, 93.04% in object segmentation, and 94.46% in object detection. Based on the confirmation of performance improvement, MPEG published CfP in April 2023.

Regarding CfP anchor-related documents, two anchor experiment results, four crosscheck experiment results, and two split point-related documents were reviewed. ETRI submitted both of the presented anchor experiment results, and they proposed CfP anchors for object detection and object segmentation tasks using OpenImages, both of which were reflected as feature anchors[17,18].

At Qualcomm, they used Amazon Cloud to provide crosscheck results for the SFU object detection feature anchor[19]. This approach was proposed to minimize discrepancies in crosscheck outcomes. However, it was determined that proceeding with the CfP using the existing methodology presented no significant issues. So, it was decided to include it in the CfP document as an option to choose. Canon provided crosscheck experiment results for the anchor submitted by ETRI, and it was confirmed that the differences did not exceed the threshold[20]. China Telecom independently provided crosscheck results for the HiEve object tracking feature anchor and collaborated with the Zhejiang University to provide crosscheck results for the TVD object tracking feature anchor[21,22]. Both results passed without exceeding the threshold for differences.

Sharp presented experimental results using an alternative split point for JED and argued that applying various split points could help prevent overfitting[23]. However, it was summarized that for introducing a new split point, evaluation of the VTP application results and characteristics of the split point need to be defined. Canon's contribution, which provided crosscheck results for the alternative split point presented by Sharp, was not reviewed as no decision was made regarding introducing a new split point[24].

At the meeting in April 2023, 5 technical contributions were submitted for Track 1. Following their previous contribution, m60257, China Telecom presented the performance of a video object tracking task using the layer just before the FPN in the JDE network[25]. The performance presented by China Telecom was achieved by combining the proposed DCT-based encoder/decoder with a feature extractor and a feature reconstructor. Compared to the Feature anchor, it demonstrated a 92% BD-rate performance improvement in object tracking.

Following their previous contribution, m60671, ETRI presented the performance of a video object tracking task by applying the proposed technique to the newly added

Table 2. Results about Object Tracking, Instance Segmentation, Object Detection

	Object Tracking		Instance Segmentation		Object Detection	
	BD-rate over Video	BD-rate over Feature	BD-Rate over Image	BD-Rate over Feature	BD-Rate over Image	BD-Rate over Feature
m60761[8]	-87.44%	-97.58%	-79.21%	-95.56%	-81.11%	-94.15%
m60788[11]	63.69%	-74.43%	-47.46%	-89.48%	-54.51%	-85.06%
m60799[9]	-80.18%	-97.09%	-93.04%	-98.60%	-94.46%	-98.34%
m60802[15]	-	-	-19.35%	-83.38%	-	-
m60803[16]	218.93%	-33.01%	-	-	-	-
m60821[10]	-77.40%	-95.84%	-78.11%	-95.84%	-70.39%	-91.14%
m60925[12]	-64.94%	-92.17%	-69.08%	-92.30%	-	-

HiEve dataset and the new split point[26]. Compared to the Feature anchor, it demonstrated a 93.38% BD-rate performance improvement on the 1080p sequence and a 92.09% improvement on the 720p sequence.

Beihang University updated their previous contributions, m61973 and m61980, presented at the last meeting[27,28]. They submitted a technical contribution replacing the three parts: Compression network, Enhancement network, and VVC with a learnable codec. According to their contribution, there was a BD-rate performance improvement of 91.84% in the instance segmentation task and 94.70% in the object detection task on the OpenImages dataset.

Shanghai Jiao Tong University, Zhejiang University, Tsinghua University, and Lenovo presented a performance improvement method for object detection tasks using the Channelwise dynamic pruning approach[29]. This method compresses by removing non-essential feature channels, and two pruning methods were proposed: pruning-ratio design and pruning-mask design. While the method demonstrated improved BD-rate performance for object detection compared to image anchor through graphs, specific BD-rate values were not provided, and experimental results in the JDE network were not obtained.

Ⅴ. VCM Track 2: Image and Video Coding

The Image and Video Coding track, as illustrated in Fig. 6, refers to a codec that takes video or image input to generate a compressed bitstream, restores it to a format similar to the original video, and then uses the restored image as an input for machine vision networks.

For the Image and Video Coding track, evaluations for the CfE were conducted in April 2021, and evaluations for the CfP proposals took place in October 2022. The evaluations during the CfP compared performance across three tasks: object detection, object segmentation, and object tracking. The datasets for each task are specified in Table 3[6].

Table 3. Key tasks and Target Dataset for VCM Track2 CfP

Target Tasks		Object Detection	Instance Segmentation	Object Tracking
Network		Faster-RCNN	Mask R-CNN	JDE-1088x608
Dataset	Image	FLIR, TVD, OpenImageV6	TVD, Openimagev6	-
	Video	SFU-HW-Objects-v1	-	TVD
Performance Metric		BPP, mAP	mAP	MOTA

The techniques proposed in the Image and Video Coding track can be broadly categorized into a region-based approach (ROI-based approach) grounded on preprocessing and an End-to-End deep learning network approach.

1. Region-based Approach Methods

In VCM, ROI (Region-of-Interest) refers to the area used for extracting key features in machine vision. As illustrated in Fig. 7, the region-based approach (ROI-based approach) involves using a network or part of a network intended for

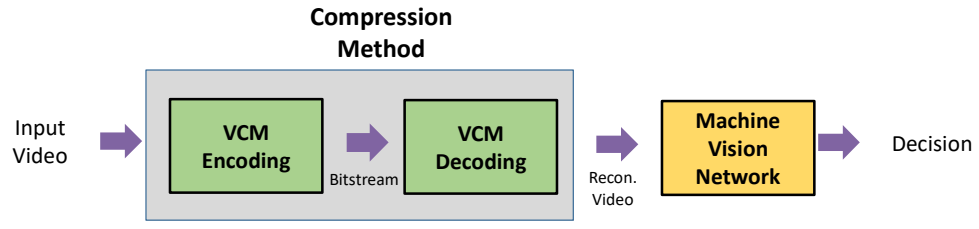

Fig. 6. Coding Pipeline of Image and Video Coding Track

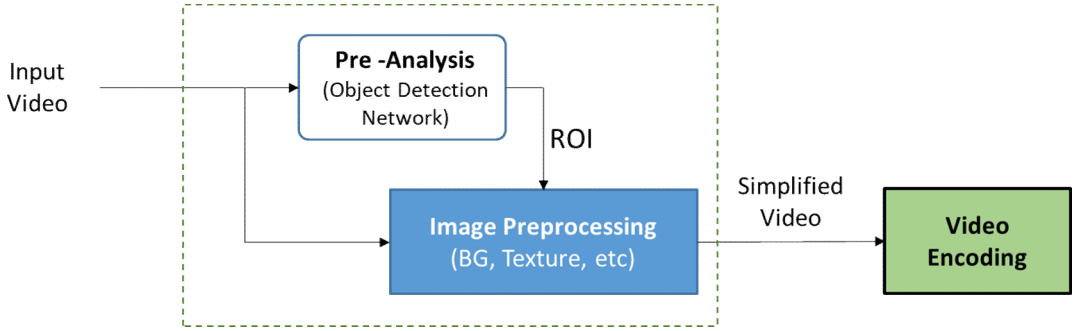

Fig. 7. region-based approach framework

machine vision on the input video. It detects the areas or frames in the video where the main objects are present and then encodes only those specific areas or frames or augments those areas or frames with more data.

The basic region-based coding method identifies the ROI and transmits only the corresponding portion. Poznan University proposed a method that detects the ROI, simplifies the background, and compresses video using conventional video coding[30-32]. ETRI and Myongji University proposed a technique that employs machine vision to obtain descriptors, descriptors for the obtained results, and object information images, which are then applied to another machine vision task, object tracking[33]. Alibaba and the City University of Hong Kong proposed a method that leverages the Yolo v7 object detection network for extracting object boundaries. This approach then utilizes the identified boundaries to temporally and spatially reduce information, effectively lowering the bitrate[34]. Florida Atlantic University and OP Solutions proposed a coding technique that uses machine vision to locate areas containing objects, reconstructs the video using only this area for transmission, and then restores it to its original video[35].

Not only sending the ROI area but methods assigning different bitrates to the ROI were also proposed. Ericsson suggested a technique that uses machine vision to identify regions containing objects[36]. This information about the region is then input into the VVC Encoder to adjust each region's quality parameter (QP), thus improving the encoding efficiency. ETRI and Konkuk University proposed a method for allocating different performance metrics based on object recognition importance, as illustrated in Fig. 8. In the CfE technical proposal, a coding method was suggested to distinguish between the foreground and background, then divide and transmit them as two streams[37]. In response to the CfP technical query, a method was proposed where the background and objects are composited into a new frame and then encoded with different qualities [38].

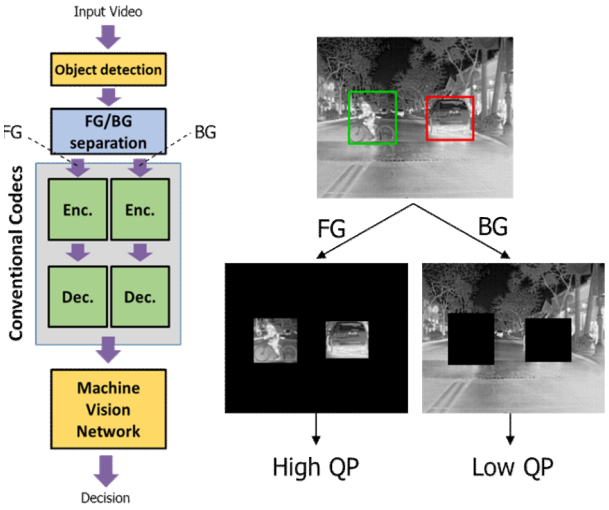

Fig. 8. Examples of Methods Transmitting Different Qualities Based on Importance

Methods for sampling video footage have also been proposed. CAS-ICT and China Telecom introduced a video

encoding technique that temporally utilizes frame samples [39]. This method regenerated intermediate frames using an interpolation network in the decoder to produce the original frames. Tencent and Wuhan University proposed a solution that employs spatiotemporal sampling techniques to reduce data size and uses deep neural networks for post-filtering during decoding[40,41].

2. Deep Learning Network-Based Compression Method

With the recent advancements in deep learning technology, there's active research into using deep learning networks for compressing various multimedia, including videos. Compression techniques based on deep learning input images or videos into deep neural networks and encode them by extracting limited-form hidden vectors.

For typical image compression, deep neural networks are trained to enhance the quality of the reconstructed image while representing the hidden vector with fewer bits for improved compression efficiency. VCM employs a deep learning network and incorporates the error function of the machine vision network during encoding. Incorporating the error function ensures that the network is trained more favorably for machine vision tasks. For example, when targeting an object detection network, the cost function L for the compression network can be represented by combining the bitrate R, the image error, and the detection network's cost.

$$L_{overall} = R + \lambda_{mse} L_{mse} + \lambda_{detect} L_{detect} \quad (4)$$

At Zhejiang University, a deep neural network-based encoding technique was proposed, composed of Key frame Encoder (MIKEnc) and Key frame Decoder (MIKDec)[42]. In this method, the encoder, MIKEnc, consists of a preprocessing network named PreP that performs image resizing and normalization, an NN-FE that extracts features from the image, and an NN-FC module that further compresses these features. On the other hand, the decoder,

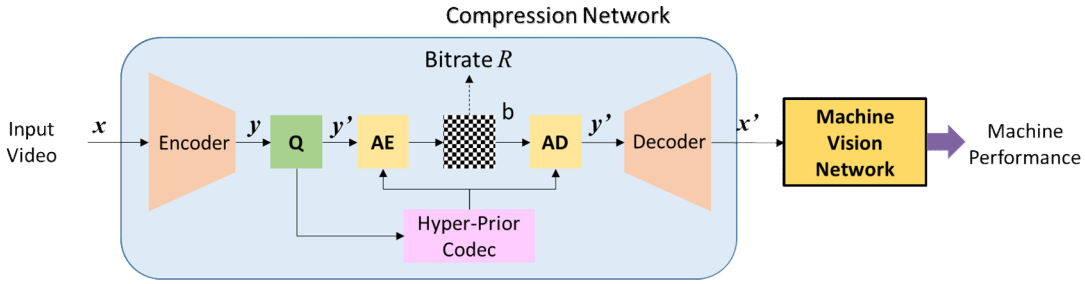

Fig. 9. Deep learning network-based compression framework

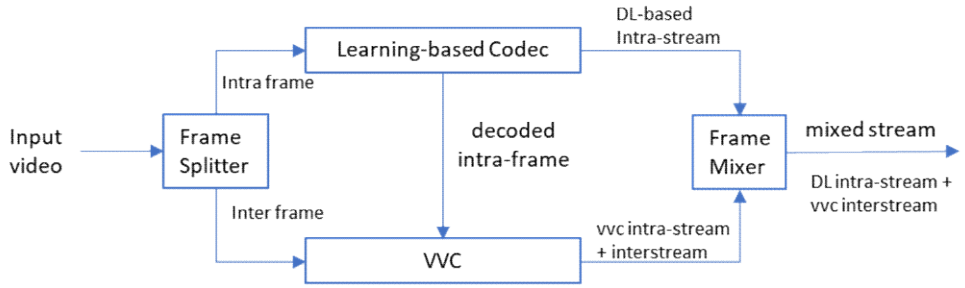

Fig. 10. Nokia's Hybrid Codec Structure

MIKDec, comprises an NN-FR that decodes the features, an NN-IR that reconstructs the original image from the feature vector, and a post-processing network that carries out image resizing and normalization. Tencent and Wuhan University proposed an End-to-end learning-based Solution [43] utilizing Variable-rate intra-coding and Scale space flow Coding[44]. For the Intra Coding Network, they enhanced the existing Cheng2020 with attention model[13] by adding a Scalingnet to support variable bitrates. Nokia introduced a hybrid encoding technique that employs deep learning-based compression for Intra-frame Coding and uses VVC coding for Inter-frame Coding[45,46].

Nokia has structured a design that forms an autoregressive context only within a limited subgroup to reduce complexity compared to the conventional LIC. The method presented by Nokia showed improved performance on three benchmark datasets compared to the conventional LIC and the VVC codec[47]. However, MPEG has concerns regarding the use of LIC technology due to difficulties in utilizing video data and its heavy dependence on the training dataset.

The evaluation results for Track 2 CfP are provided in the CfP test report and CfP response report documents [48, 49]. The top performance for each task, as listed in the CfP test report, is shown in Table 4.

Table 4. CfP Best Performances

Task	Dataset	BD-rate change
Object tracking	TVD (videos)	-57%
Instance segmentation	OpenImages	-51%
	TVD	-57%
Object detection	OpenImages	-47%
	FLIR	-53%
	TVD	-65%
	SFU (videos)	-36%

When comparing performance relative to VVC for each task, it was found that there was a performance improvement of 57% in the object tracking task, 45% in the object segmentation task, and 39% in the object detection task. The results in Table 4 represent the top performance for each task; therefore, the integrated performance is expected to be somewhat lower than the above mentioned results. Fig. 11 is an example of video restoration using the proposed technology.

Fig. 11. Restoration results using the CfP proposed technology

3. Updates on the Standardization of Track 2

VCM, as shown in Fig. 12, is currently developing the VCM track 2 reference software. Within this software, MPEG conducted a Core Experiment (CE) to evaluate the performance of its major algorithms. The term "Inner Codec" (Encoder, Decoder) mentioned in the reference software refers to existing video coding technologies like AVC, HEVC, VVC, or any other video encoding technology with similar functionalities.

Based on October 2022, five CEs (Core Experiments) were established. By April 2023, it was decided to continue with CE 1, 2, and 3, while CE 4 and 5 were to be merged into other CEs.

CE 1, titled "RoI-Based Coding Methods", is a comparative experiment that extracts Regions of Interest (RoI) and reconstructs each frame to process the video through this method. Due to issues with the Evaluation Metric and the Candidate Selection method, the results of CE 1 could not be concluded, prompting the decision to continue this CE until the next session. Five additional proposals were included in the Hybrid Codec CE in the April 2023 meeting. However, as VTM CE and Hybrid Codec CE each use their respective anchors, a performance comparison is not feasible.

CE 2 revolves around "Neural Network-Based Intra Frame Coding". It is a comparative experiment using deep neural networks on the Intra Frame coding algorithm. For CE 2, while there was a general positive outlook on the feasibility of the Hybrid Coding method, there was a sentiment that validation concerning training was necessary for the Learned Image Codec (LIC) method. With LIC, there's a strong dependency on the training dataset and the network, raising concerns about its generalization capabilities. While there were no disagreements on the potential for performance improvement, consensus was not reached on including the Hybrid Codec method in the TuC (Technologies under Consideration). Hence, it was decided to proceed with CE 2 in the next meeting.

CE 3, named "Frame Level Spatial Resampling", is about the comparative experiment of resizing each frame by resampling for encoding and then upsampling it during decoding. After the 141st meeting, four tests related to Spatial Resampling were conducted, with three test results being submitted and crosschecks completed for two test results. Based on the outcomes, Spatial Resampling demon-

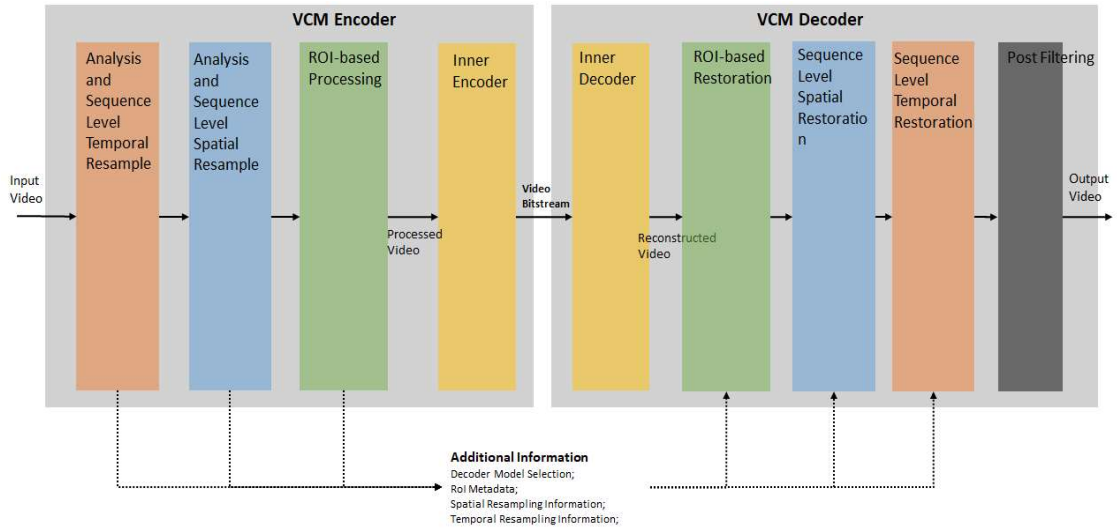

Fig. 12. Block diagram of VCM track 2 reference software

strated some performance improvements, leading to the decision to continue this CE until the next meeting.

CE 4, titled "Temporal Resampling", is a comparative experiment on an algorithm that uses criteria, such as the presence or absence of regions of interest or targets or other filtering technologies, to delete certain frames for encoding. Then, during decoding, removed frames are replenished using methods like interpolation networks. After the 141st meeting, three tests for CE 4 were conducted, and valid experimental results were submitted for two tests. Given that the experiments showed significant performance enhancements and adopted similar methodologies, it was decided to adopt the CE results in Ref.SW and candidate WD. As the results were adopted into Reference SW, discussions on CE 4 concluded, deciding not to proceed further.

CE 5 is about "Post Filtering", a comparative experiment on tools to enhance the resolution and quality of each decoded video frame. Although two tests were planned for CE 5, only one was carried out. Furthermore, it was decided to merge CE 5 with CE 2 for additional experiments.

VI. Conclusion

The MPEG VCM is currently standardizing in two distinct areas: Feature Coding (Track 1) and Image and Video Coding (Track 2). Both tracks demonstrated higher performance than a VVC anchor. Therefore, it is anticipated that the standardization of VCM will proceed smoothly according to the planned schedule and is expected to replace conventional video coding technologies in various fields requiring machine vision. Particularly, feature coding, not limited to machine vision applications but also applicable in general video coding, is predicted to be feasible for use in Neural Network Video Coding (NNVC), which is currently under discussion for standardization by the Joint Video Exploration Team (JVET).

References

[1] Hyon-Gon Choo, Won-Sik Cheong, Jeong-il Seo, “Standardization Trends in Video Coding for Machines,” Broadcasting and Media Magazine, 28(1), 38-52.

[2] L. Duan, J. Liu, W. Yang, T. Huang and W. Gao, "Video Coding for Machines: A Paradigm of Collaborative Compression and Intelligent Analytics," in IEEE Transactions on Image Processing, vol. 29, pp. 8680-8695, 2020.
doi: https://doi.org/10.1109/tip.2020.3016485

[3] “Information technology on multimedia content description interface part 13: Compact descriptors for visual search,” ISO/IEC 15938-13, Sep. 2015.

[4] “Information technology on multimedia content description interface part 15: Compact descriptors for video analysis,” ISO/IEC 15938-15, Jul. 2019.

[5] Use cases and requirements for Video Coding for Machines, ISO/IEC JTC1/SC29/WG2 N190, 2022.04

[6] Common Test Conditions and Evaluation Methodology for Video Coding for Machines, ISO/IEC JTC1/SC29/WG2 N192, 2022.04.

[7] Call for Evidence on Video Coding for Machines, ISO/IEC JTC 1/SC29/WG2 N215, 2022.07.

[8] Heeji Han et. al.,, [VCM] Response from Hanbat National University and ETRI to CfE on Video Coding for Machines, ISO/IEC JTC1/SC29/WG2/m60761, 2022.10.

[9] Yong-Uk Yoon et. al.,, [VCM] Response to VCM CfE: Multi-scale feature compression with QP-adaptive feature channel truncation, ISO/IEC JTC1/SC29/WG2/m60799, 2022.10.

[10] C. Rosewarne, R. Nguyen, [VCM Track 1] Response to CfE on Video Coding for Machine from Canon, ISO/IEC JTC1/SC29/WG2/m60821, 2022.10.

[11] Minhun Lee, et. al.,, [VCM Track 1] Response to CfE: A transformation-based feature map compression method, ISO/IEC JTC1/SC29/WG2/m60788, 2022.10.

[12] Yong Zhang et. al.,, [VCM] Response to VCM Call for Evidence from Tencent and Wuhan University - a Learning-based Feature Compression Framework, ISO/IEC JTC1/SC29/WG2/m60925, 2022.10.

[13] Cheng, Z., et. al., Learned image compression with discretized gaussian mixture likelihoods and attention modules. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7939-7948), 2020.
doi: https://doi.org/10.1109/ cvpr42600.2020.00796

[14] CfE response report for Video Coding for Machines, ISO/IEC JTC 1/SC29/WG2 N247, 2022.10

[15] Hanming Wang, Zijun Wu, Tao Han, Yuan Zhang, [VCM][Response to CfE] An End-to-End Image Feature Compressing Method with Feature Fusion Module, ISO/IEC JTC1/SC29/WG2/m60802, 2022.10.

[16] Hanming Wang, Zijun Wu, Tao Han, Yuan Zhang, [VCM][Response to CfE] An End-to-End Video Feature Compressing Method with Feature Fusion Module, ISO/IEC JTC1/SC29/WG2/m60803, 2022.10.

[17] Jooyoung Lee, Se Yoon Jeong, Younhee Kim, (ETRI), [FCVCM] OpenImages object detection feature anchor update, m62564

[18] Jooyoung Lee, Se Yoon Jeong, Younhee Kim, (ETRI), [FCVCM] OpenImages instance segmentation feature anchor update, m62565
[19] Liangping Ma, [FCVCM] Cloud-based crosscheck for the SFU object detection feature anchor, m62671
[20] Chris Rosewarne, Rose Nguyen (Canon), [FCVCM] Crosscheck of OpenImages feature anchors for object detection and instance segmentation (m62564 and m62565), m62995
[21] Zijun Wu, Yuan Zhang, Yin Yu (China Telecom), [FCVCM] Crosscheck for HiEve object tracking feature anchor in m61876, m63026
[22] Hanming Wang, Yuan Zhang (China Telecom), Li Zheng (Zhejiang University), [FCVCM] Crosscheck for the TVD object tracking feature anchor (m62504), m63024
[23] Tianying Ji, Sachin Deshpande (Sharp), [FCVCM] Alternative split point for JDE feature map compression, m63021
[24] C. Rosewarne, R. Nguyen (Canon), [FCVCM] Crosscheck of Alternative split point for JDE feature map compression (m63021), m63442
[25] Yunyu Chen (China Telecom), Lichuan Wang(China Telecom), Yuan Zhang (China Telecom), [FCVCM] Feature Compression For Object Tracking Based On The Combination Of DCT And Neural Network , m63028
[26] Heeji Han, Minseok Choi, Haechul Choi (HNU), Soon-heung Jung, Jin Young Lee, Joungil Yun, Sangwoon Kwak, Hyon-Gon Choo, Won-Sik Cheong (ETRI), [FCVCM] Experimental results of enhanced MSFC based on split point for HiEve in object tracking task, m63045
[27] Shengxi Li, Chaoran Chen, Zifu Zhang, Tie Liu, Xin Deng, Mai Xu (Beihang University), M. Rafie, Zhuoyi Lv (vivo), [VCM Track 1] Hybrid Loss Training for Feature Compression on Object Detection and Instance Segmentation, m63172
[28] Shengxi Li, Chaoran Chen, Zifu Zhang, Tie Liu, Xin Deng, Mai Xu (Beihang University), M. Rafie, Zhuoyi Lv (vivo), [VCM Track 1] Hybrid Loss Training Based on Reversed SIMO in Feature Compression for Object Detection and Instance Segmentation, m63174
[29] Ge Zhang, Kaiyuan Dong, Weiyao Lin (Shanghai Jiao Tong University), Hualong Yu (Ningbo Innovation Center, Zhejiang University), Xiao Ma, Cheng Zhong, Zhongchao Shi (LENOVO (BEIJING) LIMITED)., Adaptive feature compression with dynamic pruning gate, m63317
[30] Marek Domanski et. al., [VCM] Poznan University of Technology Proposal A in response to CfP on Video Coding for Machines, ISO/IEC JTC1/SC29/WG2/m60727, 2022. 10.
[31] Marek Domanski et. al., [VCM] Poznan University of Technology Proposal B in response to CfP on Video Coding for Machines, ISO/IEC JTC1/SC29/WG2/m60728, 2022. 10.
[32] Marek Domanski et. al., [VCM] Poznan University of Technology Proposal C in response to CfP on Video Coding for Machines, ISO/IEC JTC1/SC29/WG2/m60729, 2022. 10.
[33] Sang-Kyun Kim et. al., [VCM] CfP response: Region-of-interest based video coding for machine, ISO/IEC JTC1/SC29/WG2/m60758, 2022. 10.
[34] S. Wang et. al., [VCM] Video Coding for Machines CfP Response from Alibaba and City University of Hong Kong, ISO/IEC JTC1/SC29/WG2/m60737, 2022. 10.
[35] Hari Kalva et. al., [VCM] Response to VCM CfP from the Florida Atlantic University and OP Solutions, ISO/IEC JTC1/SC29/WG2/ m60743, 2022. 10.
[36] Christopher Hollmann et. al., [VCM] Response to Call for Proposals from Ericsson, ISO/IEC JTC1/SC29/WG2/m60757, 2022. 10.
[37] Yegi Lee et. al., [VCM] Response to CfE: Object detection results with the FLIR dataset, ISO/IEC JTC1/SC29/WG11/m56572, 2021.04
[38] Yegi Lee et. al., [VCM Track2] Response to VCM CfP: Video Coding with machine-attention, ISO/IEC JTC1/SC29/WG2/m60738, 2022. 10.
[39] Jianran Liu et. al., [VCM] Video Coding for Machines CfP Response from Institute of Computing Technology, Chinese Academy of Sciences (CAS-ICT) and China Telecom, ISO/IEC JTC1/SC29/WG2/ m60773, 2022. 10.
[40] ZIzheng Liu et. al., [VCM] Response to VCM Call for Proposals - an EVC based solution, ISO/IEC JTC1/SC29/WG2/ m60779, 2022. 10.
[41] ZIzheng Liu et. al., [VCM] Response to VCM Call for Proposals from Tencent and Wuhan University - an ECM based solution, ISO/IEC JTC1/SC29/WG2/ m60780, 2022. 10..
[42] Ke Jia et. al., [VCM] Response to the CfP on Video Coding for Machine from Zhejiang University, ISO/IEC JTC1/SC29/WG2/ m60741, 2022. 10.
[43] Wen Gao et. al., [VCM]Response to VCM Call for Proposals from Tencent - an End-to-end Learning based Solution, ISO/IEC JTC1/ SC29/WG2/m60777, 2022. 10.
[44] E. Agustsson, et. al., Scale-space flow for end-to-end optimized video compression, IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020).
doi: https://doi.org/10.1109/cvpr42600. 2020.00853
[45] Honglei Zhang et. al., [VCM] Response to the CfP of the VCM by Nokia (A), ISO/IEC JTC1/SC29/WG2/m60753, 2022. 10.
[46] Honglei Zhang et. al., [VCM] Response to the CfP of the VCM by Nokia (B), ISO/IEC JTC1/SC29/WG2/m60754, 2022. 10.
[47] Honglei Zhang, Francesco Cricri, Hamed Rezazadegan Tavakoli, Emre Aksu, Miska M. Hannuksela, Leveraging progressive model and overfitting for efficient learned image compression.
doi: https://doi.org/10.48550/arXiv.2210.04112
[48] C. Rosewarne, [VCM Track 2] CfP test report, ISO/IEC JTC1/SC29/WG2/m61010, 2022. 10..
[49] CfP response report for Video Coding for Machines, ISO/IEC JTC1/SC29/WG2/N248, 2022. 10.

Introduction Authors

Dongmin Lee

- Mar. 2020 ~ : Dong-A University, Department of Computer Engineering & AI, Undergraduate Student
- ORCID : https://orcid.org/0009-0007-2009-8753
- Research interests : Video Processing, Deep Leanring, Computer Vision

Sangkyun Jeon

- Mar. 2022 ~ : Dong-A University, Department of Computer Engineering & AI, Undergraduate Student
- ORCID : https://orcid.org/0009-0004-4287-6745
- Research interests : Video Processing, Deep Leanring, Computer Vision

Yeonghun Jeong

- Mar. 2020 ~ : Dong-A University, Department of Computer Engineering & AI, Undergraduate Student
- ORCID : https://orcid.org/0009-0003-0162-3664
- Research interests : Video Processing, Deep Leanring, Computer Vision

Joonsoo Kim

- Feb. 2017 : Seoul National University, Department of Electrical & Computer Engineering, Ph.D.
- Feb. 2017 ~ : Electronics & Telecommunications Research Institute(ETRI), Immersive Media Research Lab. Senior Researcher
- ORCID : https://orcid.org/0000-0002-6470-0773
- Research interests : Light field, VR/AR, Computer Vision, Video Coding for Machine

Jeongil Seo

- Aug. 2005 : Kyungpook National University, Electronics Engineering, Ph.D.
- Mar. 1998 ~ Oct. 2000 : LG Semicon, Junior Researcher
- Nov. 2000 ~ Feb. 2023 : Electronics & Telecommunications Research Institute(ETRI), Immersive Media Research Lab. Director
- Mar. 2023 ~ : Dong-A University, Department of Computer Engineering & AI, Associate Professor
- ORCID : https://orcid.org/0000-0001-5131-0939
- Research interests : Multimedia, Audio & Video Coding, Deep Learning, Computer Vision