IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

Video Coding for Machines with Neural-Network-based Chroma Synthesis

Mateusz Lorkiewicz[1], Sławomir Różek[1], Olgierd Stankiewicz[1], (Member, IEEE), Tomasz Grajek[1], (Senior Member, IEEE), Sławomir Maćkowiak[1] and Marek Domański[1], (Life Senior, IEEE)

[1]Institute of Multimedia Telecommunication, Poznan University of Technology, pl. M. Skłodowskiej-Curie 5, 60-965 Poznan, Poland

Corresponding author: Tomasz Grajek (e-mail: tomasz.grajek@put.poznan.pl).

The work was supported by the Ministry of Science and Higher Education, Poland

ABSTRACT Video coding for machines is an emerging area within video compression technology that has recently attracted considerable research attention. Within the ISO/IEC standardization activities, efforts are underway to develop a new standard optimized for machine vision tasks, rather than just for human-oriented video consumption. In this context, the novel contribution of this work is the application of neural networks to perform chroma synthesis at the decoder side, thus eliminating the need for direct chroma transmission. This concept has been implemented and validated in the Video Coding for Machines Reference Software (VCM-RS), a test model developed by the MPEG Video group, which is briefly described for completeness. Experimental results reported in this paper show that our approach significantly reduces the number of bits required for video representation, achieving bitrate savings of up to 60% for certain video sequences. The presented proposal has been accepted in the current version of the specification of the upcoming VCM standard.

INDEX TERMS chroma representation, chroma synthesis, colorization, video coding for machines

I. INTRODUCTION

According to the reports from data transmission monitoring companies, the vast majority (more than 80%) of information transmitted today is video data [1-2]. Over the last four decades, several generations of human-optimized video coding techniques have been developed in an attempt to cope with the aforementioned increased amount of transmitted video. Among such video coding technologies, the most recent one is VVC [3-4].

In recent years, the volume of video surveillance material has exceeded the level manageable by human observers. In addition, the proliferation of the Internet of Things, semi- and fully autonomous vehicles, AI-based video analytics, the use of intelligent video surveillance, and video-based control in many new fields are just some of the factors contributing to the rapid increase in the amount of video data shared between computers, where direct human consumption of the decoded video is not the primary application [5]. These factors have sparked research interests in video coding, where the decoded data serves as input for machine vision tasks, leading to the proposal of several techniques [6-13].

During the second decade of the 21st century, the abovementioned problems have been recognized, and the term *'video coding for machines* (VCM)' has come into use [6-13].

The need for a standardized video coding framework for machine-centric applications has been recognized by the ISO/IEC Moving Picture Experts Group (MPEG). In 2019, the MPEG Video Coding for Machines (VCM) Ad-Hoc group began works towards a future standard to enable efficient compression for machine vision and hybrid machine-human-vision tasks. By 2021, VCM had outlined initial requirements for applications such as surveillance, intelligent transportation, and smart cities, reflecting the growing role of machine-driven algorithms for visual data processing [5]. Due to the complexity of these domains, the standardization process was split into two tracks: one focused on enhancing traditional video compression methods and the other on compressing visual features (Feature Coding for Machines [14]). For the first track mentioned, in 2022, a revised Call for Proposals (CfP) [15] solicited contributions from global research institutions, leading to the formation of the Video Coding for

Machines (VCM) standardization project, which continues to drive advancements in the field.

II. MPEG VCM CODEC

The MPEG VCM approach is codec-agnostic, separating VCM-specific tools from the core video codec (the "Inner Codec"), such as HEVC or VVC (see Fig. 1). Although special codec designs have also been studied for this purpose [16-17].

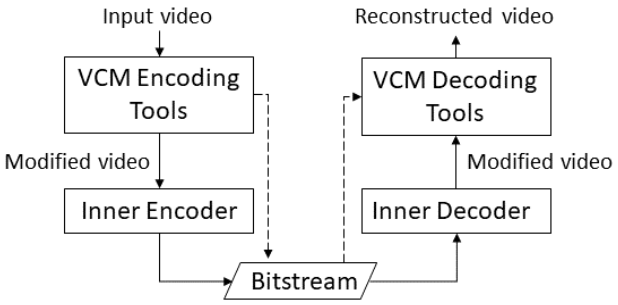


FIGURE 1. The general model of MPEG VCM codec. Dashed lines represent additional data generated by VCM encoding tools and necessary for VCM decoding tools.

The VCM-specific encoding tools transform the video before video is compressed using a standard compression technology. Some of these tools generate additional data that must be transmitted in the compressed bitstream. The VCM-specific encoding tools include spatial and temporal resampling, region-of-interest (RoI) encoding [18-20], and bit-depth truncation. Temporal downsampling reduces the frame rate by discarding and interpolating frames during encoding and decoding. Spatial downsampling reduces the resolution of video frames in a way that balances bitrate reduction with preserving the detail necessary for machine vision, while also reducing computational complexity. RoI encoding prioritizes key regions within a frame for high-quality encoding, while less important areas are grayed or discarded. A RoI usually corresponds to the area of an object and its nearest neighborhood. Bit-depth truncation reduces the luma component's dynamic range (e.g., from 10 to 9 bits) in order to reduce the bitrate without significantly affecting video quality.

After encoding using VCM tools, the resulting video is encoded by an Inner Encoder, typically VVC [3-4], or its extensions in current MPEG experiments. The compression efficiency is evaluated using selected machine vision tasks performed by artificial neural networks. The machine vision tasks usually tested are object detection and object tracking.

The novelty of this paper consists in application of an additional new tool extending the set of VCM-related encoding and decoding tools. The new encoding tool is Chroma Removal and Control, whereas the new decoding tool is Chroma Restoration. The idea is to remove chroma data at the input of the Inner Encoder and restore chroma at the output of Inner Decoder, Chroma Restoration exploits some additional data produced by Chroma Removal and Control that are transmitted in the bitstream (see Fig. 2). The bitrate for this additional data is relatively low, therefore the amount of data is reduced at the input to the Inner Encoder is lower, thus increasing compression performance of the system,

III. VIDEO COLORIZATION – EXISTING APPROACHES

Color information plays a crucial role as contextual input in tasks such as object detection, segmentation, and classification, particularly in scenarios requiring differentiation between objects with similar geometric structures.

In the VCM framework, chroma is subject to lossy compression and quantization with the goal of balancing bitrate cost against quality loss. Consequently, although chroma still accounts for a noticeable portion of the overall bitrate, its degradation also leads to a noticeable degradation in the performance of machine vision tasks. This study aims to evaluate whether a more efficient representation can be achieved by completely omitting chroma during transmission and reconstructing it at the decoder side from luminance alone using colorization.

The colorization of a video sequence is a popular topic in computer vision [21]. The development of Artificial Neural Networks (ANN) resulted in the development of multiple methods for automatic colorization of grayscale images and videos. Depending on the use case, an ANN may colorize, e.g., a part of a grayscale image [22-24], may use additional color distribution hints [23, 25-26] or a sparse grid of color samples [23, 27]. A colorization technique may also be additionally guided by indicating a reference image [28], a color palette [29], or even by text prompts [30-31]. Another approach is embodying the color hints in grayscale image [32]. In the case of a video sequence, the colorization algorithm can use temporal hints for the continuity of reproduced colors [33]. The colorization algorithm may colorize the whole image at once, or colorize subsections of the image, according to performed semantic segmentation [34-37].

Convolutional neural networks are used in most methods for the abovementioned use cases. One of the popular approaches is the use of the U-net architecture [23, 25, 38-39] as the main colorization ANN, which is trained as a part of a Generative Adversarial Network (GAN) network [23, 25, 39-41]. In most cases, the ANN estimates all image color components at once. However, in some methods, the color components are estimated separately [37, 41].

With the advancements in colorization methods, their usage for image and video compression was considered [42-44]. Peter [45] proposed to dedicate more bits for luma than for chroma during the encoding, and then restore the color on the decoder site. Some researchers [46-47] proposed to transmit the luma component only, and colorize the sequence in the decoder. In Cui [48], a single color component is first reconstructed and then used to restore the other one.

Most of the respective image and video compression methods use colorization as a postprocessing step after decoding the grayscale content. Additionally, the techniques are dedicated to human consumers and typical natural content.

For video coding for machines, it is essential to understand how image colorization affects the performance of machine vision tasks like object detection, segmentation, and classification.

Zhu et al. [49] examined the impact of inaccurate RGB data on segmentation accuracy in colorized point clouds. Two types of inaccuracies have been considered: incorrect color and similar color information, both of which significantly reduced performance. The study employed the DeepLabV3+ model, assessing results using Overall Accuracy (OA) and Mean Intersection over Union (mIoU).

Singh et al. [50] explored applications of Convolutional Neural Networks (CNN) and how color information influences CNN performance in object recognition. They trained models on original color images, grayscale images, and images with altered color distributions to assess reliance on color cues. The results indicate that, while CNNs can recognize objects without color, accurate color cues obviously enhance classification accuracy, highlighting the role of colorization in object classification tasks.

These considerations underscore the importance of accurate color information in improving the performance of detection, segmentation, and classification algorithms.

Unfortunately, the subject of colorization in video coding for machines remains nearly unexplored. In this paper, the idea of colorization is explored and tested in VCM. Here, colorization is considered as one of the coding steps, which is also the novelty of this work.

IV. MAIN IDEA

The main idea of the paper is to synthesize chroma at the decoder side (from the luma component alone) instead of transmitting it directly. The application of this idea within the MPEG VCM framework [51] is illustrated in Fig. 2. When the proposed new compression tool is used, the chroma is cleared prior to further processing in the encoder (before the Spatial Downsampling tool is activated) thus allowing for bitrate savings in the Inner Encoder. The rationale for that is the observation that exact color impression is mostly not crucial for machine vision, but color differences often play a critical role in machine vision.

In the proposed approach, chroma synthesis (restoration) is implemented with the use of an Artificial Neural Network (ANN). At its input, only the decoded luma samples are available. Obviously, chroma synthesis in such a case provides only an approximation of the original chroma, which may be insufficient for some sequences in some applications. Therefore, we also propose a control mechanism that enables or disables the proposed chroma synthesis tool. The decision is made at the encoder in an automated way, based on correlation metric (see Section V B). Through this paper, we call the MPEG VCM codec with the abovementioned modifications the Modified MPEG VCM Codec.

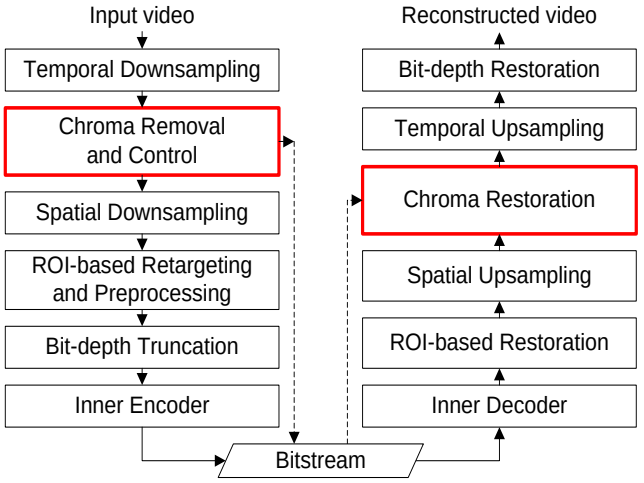


FIGURE 2. The application of the proposed idea within the MPEG VCM framework. The new stages added by this work are marked in red. The dashed lines represent the additional data generated by Chroma Removal and Control that is necessary for Chroma Restoration. For the sake of clarity, the other additional data produced by VCM Coding Tools are omitted.

V. PRACTICAL IMPLEMENTATION OF THE MODIFIED MPEG VCM CODEC

A. NEURAL NETWORK

As a colorization method, we choose the SIGGRAPH17 method [23], available as an open-source package [52]. In the experiments executed by the authors, two other models (ECCV [22] and DeOldify [53]) were tested, but SIGGRAPH17 gave the best results. The SIGGRAPH17 uses one of the popular approaches to colorization with ANN: U-net-like convolutional network [38], which is trained as a part of the GAN network [40]. The network architecture is presented in Fig. 3. Such a network is a good reference for other approaches. The model is adapted for processing and outputting frames of size 256×256 represented in CIE Lab color space [54]. The network is trained using the ImageNet dataset [55].

The input video samples are represented in the YCbCr color system, and the 4:2:0 chroma subsampling scheme is used. Before colorization, fame contains a luma component only. In order to keep compliance with the software used, the luma samples are converted to RGB, and then RGB samples are converted to CIE Lab color coordinates [56]. Only the L component is inputted to the network (ANN). Lastly, the frame is downscaled to the size of 256×256.

The ANN outputs a and b color components. The ANN input (L) is concatenated with the output to get the color frame. Then, the frame is resized to the original resolution, and color conversion from CIE Lab, through RGB to YCbCr is

performed. Finally, chromas are subsampled to meet the 4:2:0 chroma sampling scheme.

Notably, the Y component of the resulting frame is copied from the original frame. Effectively, only chroma Cb and Cr channels are modified.

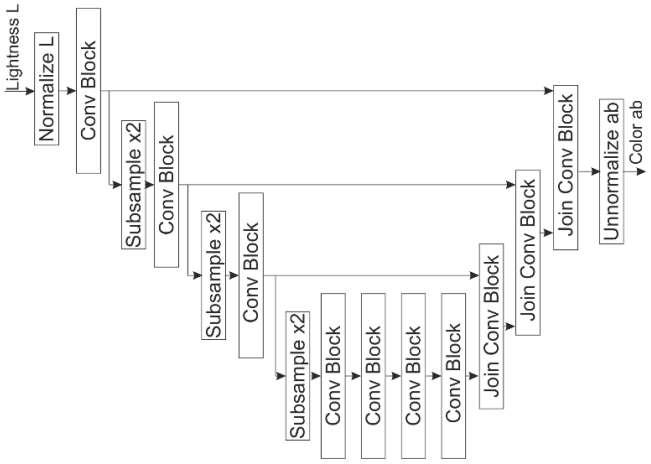


FIGURE 3. The network architecture of the employed ANN model for chroma a and b synthesis for the CIE Lab color space [23].

B. ENCODER OPERATION

The operation of the proposed tool in the MPEG VCM encoder is presented in Fig. 4. As mentioned before, the encoder has to decide whether to enable or disable the proposed colorization based on the potential fidelity of the reconstructed chroma.

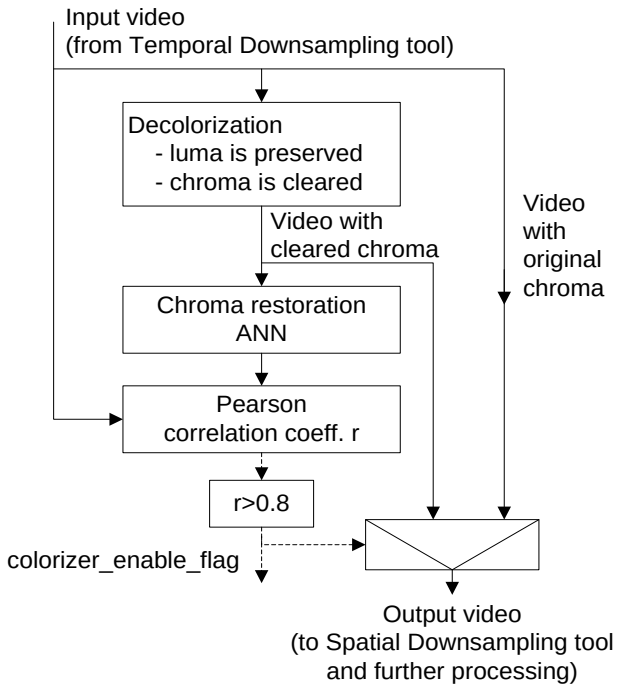


FIGURE 4. Operation of the proposed tool in the encoder.

To estimate that, the chroma is speculatively removed and restored (synthesized). Then, the correlation between the reconstructed chroma and the original chroma is measured using the Pearson Correlation Coefficient (r). High correlation ($r > threshold$) is the condition to turn the proposed tool on (and thus not directly encode the chroma and synthesize it on the decoder side instead). Otherwise, the proposed tool is disabled, and the chroma is encoded as usual. The *threshold* in our method was set to 0.8, based on extensive experiments. In the bitstream, the information about the decision is signaled as a single-bit flag "colorizer_enable_flag" (Fig. 4). This decision is done (and signaled) on the basis of each IDR frame in order to meet random-access requirements in all considered encoding scenarios (see Section VI A).

C. DECODER OPERATION

When the proposed tool is enabled (which is signaled with "colorizer_enable_flag" in the bitstream), it means that the transmitted frame has been decolorized, i.e. contains luma only. In such a case, the chroma components are synthesized using the ANN.

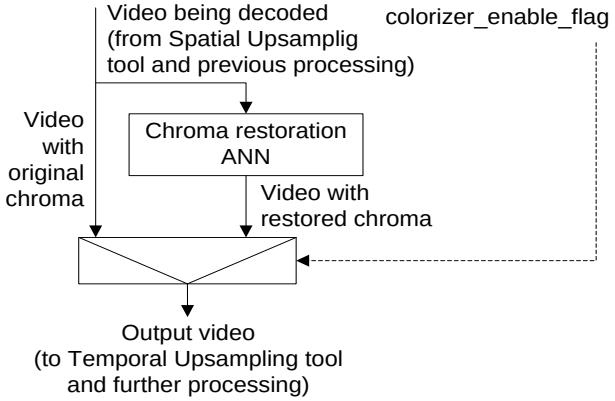


FIGURE 5. Operation of the proposed tool in the decoder.

VI. EXPERIMENTAL VALIDATION

A. CONDITIONS

In order to demonstrate the potential of the presented approach, the method has been implemented in the MPEG VCM reference software [16]. The experiment has been performed according to the MPEG VCM Common Test Conditions (CTC) [51].

The experiment consists of encoding, decoding, and evaluation of SFU-HW and TVD sequences in six scenarios. The comparison is done using the coding results of the current version of the MPEG VCM codec with its default configuration [51].

The MPEG VCM Common Test Conditions (CTC) defines six encoding scenarios, which consist of three configurations: Random Access (RA), Low Delay (LD), and All Intra (AI), in two versions: Inner and End-to-End (E2E) [51]. The difference between Inner and E2E versions of configurations

lays in the scope in which RA/LD/AI temporal requirements are applied. In Inner scenarios, the requirements are applied only on the level of the Inner Codec. In E2E scenarios, the requirements are applied to the whole encoder-decoder system (the Inner Codec and the VCM coding tools, see Fig.1). For LD E2E and AI E2E scenarios, the encoder cannot analyze or use the frames further to this, which is actually encoded. That implies, for example, that the temporal interpolation is to be replaced by extrapolation, which uses only the past frames (in LD E2E), or to be turned off (in AI E2E).

The SFU-HW sequences are divided into four classes according to their resolutions (A: 2560×1600, B: 1920×1080, C: 832×480, D: 416×240).

The mandatory evaluation comprises two machine vision tasks done on the decoded video – Object Detection, done on the SFU-HW dataset, and Object Tracking done on the TVD dataset.

B. EXAMPLES

Example colorized frames compared to the original ones are shown in Fig. 6. As can be seen, the colors are synthesized well, however, in some cases the saturation of colors seems to be lower.

FIGURE 6. Examples of frames: left column - original frame, right column - synthesized chroma [23].

C. RESULTS

The results are presented in Tables I and II as bitrate reduction measured as the Bjøntegaard delta [51] versus the mean average precision (mAP) for object detection and the multiple object tracking accuracy (MOTA) for object tracking. Moreover, the encoding and decoding times are provided for the proposed tool switched on. All the results are provided with respect to the encoder and decoder with the proposed tool switched off. For example, for a given sequence and a given scenario, the value of BD-rate mAP is -14.79%. This means that the average bitrate is reduced by 14.79%, keeping the same mAP for object detection. The relative encoding time of 101% means that switching on the proposed tool increases the VCM encoding time by 1.11%.

From the detailed results (Table II), it can be seen that the method provides interesting results in the cases when the chroma synthesis is turned on. The very small losses in the other cases are caused by the new control bits, which inform the decoder to bypass the colorization tool. Such a loss is negligible as compared to gains achieved by the tool in some other sequences (up to ~60%). Generally, in the context of VCM, the expected error margin due to the noise in metrics like MOTA/mAP is about 1%.

The proposed tool improves the compression efficiency for some sequences. The complexity overhead is mostly below 2.6% for the encoder and below 5% for the decoder.

Table I provides the average results for the sequences tested. The results demonstrate an average bitrate reduction for object tracking of 6.6% to 18.4%, depending on the compression scenario. This percentage reduction is much higher than the percentage increase of the encoding and decoding complexity measured for CPU implementations. Unfortunately, the method does not improve encoding and decoding for object detection. This issue needs some further investigation.

TABLE I
SUMMARIZED RESULTS FOR ALL VCM ENCODING SCENARIOS

Encoding scenario	Object Detection BD-Rate [%] mAP	Object Tracking BD-Rate [%] MOTA	Encoding Time [%]	Decoding Time [%]
Random Access (Inner)	0.04	-14.42	100.64	101.14
Random Access (E2E)	0.03	-12.09	100.57	101.41
Low Delay (Inner)	0.04	-10.79	100.84	101.01
Low Delay (E2E)	0.05	-6.60	100.83	100.95
All Intra (Inner)	0.02	-12.65	100.53	100.64
All Intra (E2E)	0.03	-18.48	100.41	100.82

TABLE II
DETAILED RESULTS FOR RANDOM ACCESS, LOW DELAY AND ALL INTRA ENCODING SCENARIOS

	Random Access (Inner)			Low Delay (Inner)			All Intra (Inner)		
Object Detection	BD-Rate [%] mAP	Encoding Time [%]	Decoding Time [%]	BD-Rate [%] mAP	Encoding Time [%]	Decoding Time [%]	BD-Rate [%] mAP	Encoding Time [%]	Decoding Time [%]
Class A	0.03	100.78	99.75	0.03	100.64	100.18	0.01	100.54	100.08
Class B	0.06	99.85	99.79	0.04	100.49	100.52	0.01	100.53	100.72
Class C	0.03	101.21	100.22	0.02	100.55	99.91	0.03	100.51	100.61
Class D	0.04	100.54	100.26	0.05	101.39	100.55	0.02	101.23	99.85
Average	0.04	100.60	100.01	0.04	100.77	100.29	0.02	100.70	100.32
Object Tracking	BD-Rate [%] MOTA	Encoding Time [%]	Decoding Time [%]	BD-Rate [%] MOTA	Encoding Time [%]	Decoding Time [%]	BD-Rate [%] MOTA	Encoding Time [%]	Decoding Time [%]
TVD-1-1	0.05	100.23	99.74	0.06	100.09	99.83	0.03	100.32	100.64
TVD-1-2	0.09	100.19	100.59	0.12	99.80	99.53	0.03	99.94	100.21
TVD-1-3	0.07	100.12	100.19	0.08	100.16	99.87	0.03	100.11	100.09
TVD-2-1	0.04	99.65	100.02	0.07	100.18	99.81	0.08	100.09	99.51
TVD-3-1	**-14.79**	**101.95**	**105.46**	**1.43**	**101.73**	**104.51**	**-23.29**	**100.53**	**102.22**
TVD-3-2	**-25.43**	**101.25**	**104.38**	**-23.38**	**101.71**	**104.29**	**-25.91**	**100.71**	**101.93**
TVD-3-3	**-60.97**	**101.38**	**105.42**	**-53.90**	**102.60**	**104.21**	**-39.52**	**100.73**	**102.13**
Average	-14.42	100.68	102.26	-10.79	100.90	101.72	-12.65	100.35	100.96
	Random Access (E2E)			Low Delay (E2E)			All Intra (E2E)		
Object Detection	BD-Rate [%] mAP	Encoding Time [%]	Decoding Time [%]	BD-Rate [%] mAP	Encoding Time [%]	Decoding Time [%]	BD-Rate [%] mAP	Encoding Time [%]	Decoding Time [%]
Class A	0.03	100.28	99.95	0.06	100.53	100.03	0.01	100.31	99.99
Class B	0.02	100.05	99.77	0.06	100.42	100.72	0.02	100.56	100.65
Class C	0.03	101.11	101.34	0.03	100.59	100.02	0.02	100.34	100.32
Class D	0.03	100.59	100.85	0.06	101.34	99.87	0.02	100.99	100.23
Average	0.03	100.51	100.48	0.05	100.72	100.16	0.03	100.55	100.30
Object Tracking	BD-Rate [%] MOTA	Encoding Time [%]	Decoding Time [%]	BD-Rate [%] MOTA	Encoding Time [%]	Decoding Time [%]	BD-Rate [%] MOTA	Encoding Time [%]	Decoding Time [%]
TVD-1-1	0.04	100.41	99.84	0.08	100.14	99.76	0.02	100.01	99.92
TVD-1-2	0.10	100.03	100.52	0.15	99.88	99.57	0.03	99.90	99.61
TVD-1-3	0.07	100.02	100.03	0.11	100.21	99.89	0.01	100.05	100.21
TVD-2-1	0.03	99.55	100.93	0.09	100.19	99.83	0.03	100.07	99.93
TVD-3-1	**-12.33**	**102.02**	**105.56**	**-2.04**	**101.76**	**104.64**	**-37.76**	**100.44**	**103.29**
TVD-3-2	**-25.48**	**101.15**	**104.18**	**-13.44**	**101.76**	**104.23**	**-34.17**	**100.64**	**102.55**
TVD-3-3	**-47.07**	**101.23**	**105.33**	**-31.18**	**102.65**	**104.25**	**-57.49**	**100.75**	**103.77**
Average	-12.09	100.63	102.34	-6.60	100.94	101.74	-18.48	100.27	101.33

VII. CONCLUSIONS

This paper introduces a novel neural network-based chroma synthesis tool for video coding for machines (VCM), enabling significant bitrate savings while maintaining high performance in machine vision tasks. Integrated seamlessly into the MPEG VCM framework, the method utilizes an automatic control mechanism at the encoder, which dynamically activates chroma synthesis based on correlation metrics. This approach ensures that the proposed tool is applied adaptively, thus preserving machine vision task performance in diverse video scenarios.

The modified codec was carefully tested under several encoding scenarios defined by MPEG in Common Test Condition document [51]. Experimental results show an average bitrate reduction of 12% in Random Access End-to-End configurations for object tracking tasks, confirming the method's ability to optimize network and storage demands without compromising machine vision accuracy. Importantly, the replacement of the decoded chroma by the synthesized chroma was found to have minimal impact on detection and tracking performance, validating the efficiency of the control mechanisms.

This study lays the groundwork for incorporating advanced chroma synthesis techniques into future VCM standards to address the growing need for efficient machine-oriented video coding. Further research is warranted to explore its applicability to broader machine vision tasks, such as segmentation and classification, and to develop adaptive models for diverse data contexts.

Recently, our method has been thoroughly evaluated by experts working for the ISO/IEC MPEG Video Coding for Machines (VCM) group [57-64]. The refined version of our proposal has been accepted in the current version of the specification of the upcoming VCM standard. This advanced version of the standard is currently formally published as a so-called Committee Draft [65], i.e. the standard proposal submitted to the standardization bodies of the ISO/IEC countries.

REFERENCES

[1] "Cisco visual networking index: forecast and trends", 2018–2023, online: https://cyrekdigital.com/uploads/content/files/white-paper-c11-741490.pdf (access: 4 December 2024).
[2] E. Lukan, "50 video statistics you can't ignore in 2024," online: https://www.synthesia.io/post/video-statistics, (access: 15 December 2024).
[3] ISO/IEC DIS 23090-3 (2020) / ITU-T Recommendation H.266 (08/2020), "Versatile video coding".
[4] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, J.-R. Ohm, "Overview of the versatile video coding (VVC) standard and its applications," IEEE Trans. on Circuits and Systems for Video Technology, vol. 31, no. 10, pp. 3736–3764, 2021.
[5] ISO/IEC JTC1/SC29/WG2, "Use cases and requirements for video coding for machines," MPEG doc. WG02N0190, April 2022, Online.
[6] L. Duan, J. Liu, W, Yang, T. Huang, W. Gao, "Video coding for machines: a paradigm of collaborative compression and intelligent analytics," IEEE Trans. on Image Processing vol. 29, pp. 8680–8695, 2020.
[7] K. Fischer, F. Brand, C. Herglotz, A. Kaup, "Video coding for machines with feature-based rate-distortion optimization," 22nd Int. Workshop on Multimedia Signal Processing (MMSP), 2020.
[8] L. Galteri, M. Bertini, L. Seidenari, A. Del Bimbo, "Video compression for object detection algorithms," International Conference on Pattern Recognition (ICPR), pp. 3007–3012, 2018.
[9] H. Choi, I. V. Bajic, "High efficiency compression for object detection," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1792–1796, 2018.
[10] J. Chao, E. Steinbach, "Keypoint encoding for improved feature extraction from compressed video at low bitrates," IEEE Trans. Multimedia, vol. 18, no. 1, pp. 25–39, Jan. 2016.
[11] S. Ma, X. Zhang, S. Wang, X. Zhang, C. Jia, and S. Wang, "Joint feature and texture coding: Toward smart video representation via frontend intelligence," IEEE Trans. Circuits Syst. Video Technol., vol. 29, no. 10, pp. 3095–3105, Oct. 2019.
[12] S. Xia, K. Liang, W. Yang, et al. "An emerging coding paradigm VCM: A scalable coding approach beyond feature and signal", IEEE International Conference on Multimedia and Expo (ICME). 2020.
[13] X. Sheng, L. Li, D. Liu and H. Li, "VNVC: A versatile neural video coding framework for efficient human-machine vision," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 7, pp. 4579-4596, July 2024, doi: 10.1109/TPAMI.2024.3356548.
[14] M. E. Hossain Eimon, J. Merlos, A. Perera, H. Kalva, V. Adzic and B. Furht, "Enabling next-generation consumer experience with feature coding for machines," 2025 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 2025, pp. 1-4, doi: 10.1109/ICCE63647.2025.10930026.
[15] ISO/IEC JTC 1/SC 29/WG 2, "Call for proposals for video coding for machines", MPEG doc. N191, April 2022.
[16] ISO/IEC JTC 1/SC 29/WG 4, "Algorithm description of tools in VCM reference software", MPEG WG4 Doc. N664, April 2025.
[17] H. Yu, “Container: VCM-CE2 description: neural network based inner coding,” ISO/IEC JTC1/SC29/WG4 Doc. MPEG M64618, Geneva, July 2023.
[18] O. Stankiewicz et al., "Region-of-interest-based video coding for machines," 2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Niagara Falls, ON, Canada, 2024, pp. 1-6, doi: 10.1109/ICMEW63481.2024.10645441.
[19] Y. Lee, S. Kim, K. Yoon, H. Lim, S. Kwak and H. -G. Choo, "Machine-attention-based video coding for machines," 2023 IEEE International Conference on Image Processing (ICIP), Kuala Lumpur, Malaysia, 2023, pp. 2700-2704, doi: 10.1109/ICIP49359.2023.10222037.
[20] Yang W, Huang H, Hu Y, et al. "Video coding for machines: Compact visual representation compression for intelligent collaborative analytics", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
[21] I. Žeger, S. Grgic, J. Vuković and G. Šišul, "Grayscale image colorization methods: overview and evaluation," IEEE Access, vol. 9, pp. 113326-113346, 2021, doi: 10.1109/ACCESS.2021.3104515.
[22] R. Zhang, Ph. Isola, A. Efros, "Colorful image colorization," Proceedings of ECCV, 2017.
[23] R. Zhang, Jun-Yan Zhu, Ph. Isola, X. Geng, A. . Lin, T. Yu, A. Efros, "Real-time user-guided image colorization with learned deep priors," ACM Transactions on Graphics (TOG), vol. 9, no 4. 2017.
[24] X. Cong, Y. Wu, Q. Chen, C. Lei, "Automatic controllable colorization via imagination," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2609-2619, 2024.
[25] H. Lee, D. Kim, D. Lee, J. Kim, J. Lee, "Bridging the domain gap towards generalization in automatic colorization," European Conference on Computer Vision, 2022.
[26] G. Kim, et al., "BigColor: colorization using a generative color prior for natural images," European Conference on Computer Vision, 2022.
[27] Y. Xiao, P. Zhou, Y. Zheng, C. -S. Leung, "Interactive deep colorization using simultaneous global and local inputs," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1887-1891, 2019.
[28] F. Fang, T. Wang, T. Zeng, G. Zhang, "A Superpixel-based variational model for image colorization," IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 10, pp. 2931-2943, 2020.

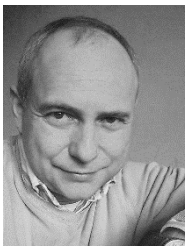
[29] H. Chang, O. Fried, Y. Liu, S. DiVerdi, A. Finkelstein, "Palette-based photo recoloring," ACM Trans. Graph., vol. 34, no. 139, 2015.
[30] Z. Chang, S. Weng, P. Zhang, Y. Li, S. Li, B. Shi, "L-CoIns: Language-based colorization with instance awareness," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19221-19230, 2023.
[31] W. Yang, F. Cai, Y. Shu, Z. Zhang, Q. Liu and Y. Ding, "Colorize at will: Harnessing diffusion prior for image colorization," IEEE Access, vol. 12, pp. 107287-107296, 2024, doi: 10.1109/ACCESS.2024.3435485.
[32] T. Ye, Y. Du, J. Deng and S. He, "Invertible grayscale via dual features ensemble," in IEEE Access, vol. 8, pp. 89670-89679, 2020, doi: 10.1109/ACCESS.2020.2994148.
[33] R. Ward, D. Bigioi, S. Basak, J. G. Breslin and P. Corcoran, "Latent colorization: latent diffusion-based speaker video colorization," in IEEE Access, vol. 12, pp. 81105-81121, 2024.
[34] G. Kong, H. Tian, X. Duan and H. Long, "Adversarial edge-aware image colorization with semantic segmentation," in IEEE Access, vol. 9, pp. 28194-28203, 2021, doi: 10.1109/ACCESS.2021.3056144.
[35] T. -T. Nguyen-Quynh, S. -H. Kim and N. -T. Do, "Image colorization using the global scene-context style and pixel-wise semantic segmentation," in IEEE Access, vol. 8, pp. 214098-214114, 2020, doi: 10.1109/ACCESS.2020.3040737.
[36] W. Ye, H. Chen, Z. Zhang, Y. Liu, S. Weng and C. -C. Chang, "Hybrid scheme of image's regional colorization using mask R-CNN and Poisson editing," in IEEE Access, vol. 7, pp. 115901-115913, 2019, doi: 10.1109/ACCESS.2019.2936258.
[37] M. Gain and R. Debnath, "A novel unbiased deep learning approach (DL-Net) in feature space for converting gray to color image," in IEEE Access, vol. 11, pp. 78918-78933, 2023, doi: 10.1109/ACCESS.2023.3263948M.
[38] O. Ronneberger, P. Fischer, T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in: N. Navab, J. Hornegger, W. Wells, A. Frangi, (eds) Medical image computing and computer-assisted intervention – MICCAI 2015, Lecture Notes in Computer Science, vol 9351, 2015.
[39] H. Shafiq and B. Lee, "Image colorization using color-features and Adversarial Learning," IEEE Access, vol. 11, pp. 132811-132821, 2023, doi: 10.1109/ACCESS.2023.3335225.
[40] I. Goodfellow, et al. "Generative adversarial nets," Communications of the ACM, vol. 63, no. 11, pp. 139 – 144, 2014.
[41] K. Du, C. Liu, L. Cao, Y. Guo, F. Zhang and T. Wang, "Double-channel guided generative adversarial network for image colorization," in IEEE Access, vol. 9, pp. 21604-21617, 2021,
[42] M. H. Baig, L. Torresani, "Multiple hypothesis colorization and its application to image compression," Computer Vision and Image Understanding, vol. 164, pp. 111-123, 2017.
[43] Y. Xiao et al., "Interactive deep colorization and its application for image compression," IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 3, pp. 1557-1572, 2022.
[44] H. Wang, X. Liu, "Overview of image colorization and its applications," 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), pp. 1561-1565, 2021, DOI: 10.1109/IAEAC50856.2021.9390626.
[45] P. Peter, L. Kaufhold, J. Weickert, "Turning diffusion-based image colorization into efficient color compression," IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 860-869, 2017.
[46] A. Singh, A. Chanani, H. Karnick, "Video colorization using CNNs and keyframes extraction: an application in saving bandwidth," in: N. Nain, S. Vipparthi, B. Raman (eds), Computer Vision and Image Processing. CVIP 2019, vol 1148. Springer, Singapore.
[47] Z. Pan, F. Yuan, J. Lei, S. Kwong, "Video compression coding via colorization: A Generative Adversarial Network (GAN)-Based approach," 2019, DOI: 10.48550/arXiv.1912.10653.
[48] K. Cui, A. Boev, E. Alshina and E. Steinbach, "Color image restoration exploiting inter-channel correlation with a 3-Stage CNN," IEEE Journal of Selected Topics in Signal Processing, vol. 15, no. 2, pp. 174-189, 2021.
[49] Q. Zhu, J. Cao, Y. Cai, L. Fan, "Evaluating the impact of point cloud colorization on semantic segmentation accuracy," IEEE 8th International Conference on Vision, Image and Signal Processing (accepted), 2024, DOI: 10.48550/arXiv.2410.06725.
[50] A. Singh, A. Bay, A. Mirabile, "Assessing the importance of colours for CNNs in object recognition," 2020, DOI: 10.48550/arXiv.2012.06917.
[51] "Common Test Conditions for Video Coding for Machines," Doc. ISO/IEC JTC1/SC29/WG4 N0543, Sapporo, July 2024.
[52] https://github.com/richzhang/colorization.
[53] https://github.com/jantic/DeOldify.
[54] CIE Colorimetry 15 (Third ed.). CIE. 2004. ISBN 3-901-906-33-9
[55] O. Russakovsky, et al., "ImageNet large scale visual recognition challenge. International Journal of Computer Vision, vol. 115, pp. 211–252, 2015.
[56] L. W. MacDonald, M, Ronnier Luo (eds.), "Colour Imaging," Wiley 1999.
[57] O. Stankiewicz, S. Różek, M. Lorkiewicz, T. Grajek, S. Maćkowiak, M. Wawrzyniak, J. Stankowski, Marek Domański "[VCM] Neural-network-based chroma reconstruction", ISO/IEC JTC1/SC29/WG4 m70249, October 2024, Kemer, Turkey.
[58] Q. Cheng, H. Wang, L. Fang, E. Lu, Y. Zhang, "[VCM] Crosscheck of m70249", ISO/IEC JTC1/SC29/WG4 m70550, October 2024, Kemer, Turkey.
[59] M. Domański, M. Lorkiewicz, H. Żabiński, T. Grajek, O. Stankiewicz, S. Różek, S. Maćkowiak, J. Stankowski, "[VCM] Improved neural-network-based chroma reconstruction", ISO/IEC JTC1/SC29/WG4 m70724, February 2025, Geneve, Switzerland.
[60] Qingsu Cheng, Huifen Wang, Li Fang, Enmin Lu, Yuan Zhang, "[VCM] Crosscheck of m70724", ISO/IEC JTC1/SC29/WG4 m71578, February 2025, Geneve, Switzerland.
[61] M. Xu, "[VCM] Crosscheck of m70724", ISO/IEC JTC1/SC29/WG4 m71696, February 2025, Geneve, Switzerland.
[62] A. Kim, "[VCM] Crosscheck of m70724", ISO/IEC JTC1/SC29/WG4 m71755, February 2025, Online meeting.
[63] O. Stankiewicz, J. Stankowski, M. Lorkiewicz, S. Różek, H. Żabiński, T. Grajek, S. Maćkowiak, M. Domański, "[VCM] Neural-Network-based chroma reconstruction with improved numerical stability", ISO/IEC JTC1/SC29/WG4 m72375, April 2025, Online meeting.
[64] E. Lu, H. Wang, L. Fang, Q. Cheng, Y. Zhang, "[VCM] Crosscheck of m72375", ISO/IEC JTC1/SC29/WG4 m72438, April 2025, Online meeting.
[65] ISO/IEC/JTC1/SC29/WG4 N662, Doc. W25197, "Text of ISO/IEC CD 23888-2 Video coding for machines," April 2025.

MATEUSZ LORKIEWICZ received the M.S. degree in the Faculty of Electronics and Telecommunications, Poznan University of Technology (PUT), in 2018. Currently a Ph.D. student and assistant at the Institute of Multimedia Telecommunication in the Faculty of Computing and Telecommunications, PUT. He published several papers (proceedings of international conferences) on audio supervision and video encoding control. He is a co-author of ISO standardization papers in the field of depth estimation and free-viewpoint television. His professional interests include audio and video signal processing, video compression, video compression control algorithms, and artificial neural networks.

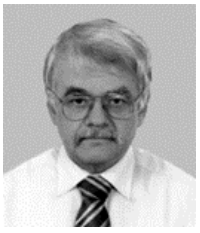
SŁAWOMIR RÓŻEK received the M.S. degree in the Faculty of Electronics and Telecommunications, Poznan University of Technology (PUT), in 2018, and another M.S. degree in the field of Automatics and Robotics, PUT, in 2020. Currently a Ph.D. candidate and assistant at the Institute of Multimedia Telecommunication in the Faculty of Computing and Telecommunications, PUT. He is actively involved in ISO standardization activities, where he contributes to the development of Video Coding for Machines (VCM) standard. He is a co-inventor in several patents and pending patent applications. His professional interests include video compression algorithms and the design of analog and digital electronics.

OLGIERD STANKIEWICZ (Member, IEEE) received his habilitation from the Faculty of Computing and Telecommunications, Poznan University of Technology (PUT), in 2020, and previously his PhD from the Faculty of Electronics and Telecommunications, PUT, in 2014. Currently, he is an assistant professor and chief of the Laboratory of Multimedia at PUT, where he has been working since receiving his Master of Engineering degree in 2006. In 2005 he won second place in the IEEE Computer Society International Design Competition (CSIDC), held in Washington D.C. He is actively involved in ISO standardization activities where he contributes to the development of the video coding standards in JCT-3V, MPEG-I, FTV, and VCM. In the years 2011-2014, he was a coordinator of the development of MPEG reference software for 3D-video coding standards based on AVC. Since 2023 he is a coordinator of the development of MPEG reference software for VCM video coding technology. He has published over 100 papers (journals, proceedings of international conferences, also MPEG/JPEG databases) on free view television, depth estimation, view synthesis, and hardware implementation in FPGA. His professional interests include signal processing, video compression algorithms, computer graphics, and hardware solutions. He is co-inventor in several patents and pending patent applications in European and US patent offices.

TOMASZ GRAJEK (M'14—SM'18) received his M.Sc. and Ph.D. degrees from Poznan University of Technology in 2004 and 2010 respectively. He has been leading several projects for industrial research and development. At present he is an Assistant Professor at the Institute of Multimedia Telecommunication, PUT. He is an author and co-author of over eighty papers (journals, proceedings of international conferences, and also MPEG/JPEG databases) on digital video compression, entropy coding and modeling of advanced video encoders. He holds ten patents (EPO and US) as well as several patent applications. He is a Senior Member of Institute of Electrical and Electronics Engineers (IEEE) and Member of Polish Society for Theoretical and Applied Electrical Engineering (PTETiS). His professional interests include signal processing, video compression algorithms, modelling of video encoders.



SŁAWOMIR MAĆKOWIAK received his Ph.D. degree in electrical engineering from Poznan University of Technology, Poland, in 2002. He is an Assistant Professor at the Institute of Multimedia Telecommunication, Poznan University of Technology. His professional activity combines teaching and applied research in the field of multimedia signal processing. He has been involved in numerous industrial research and development projects, particularly in cooperation with telecommunication partners. He is a co-author of many scientific publications and patents in the area of video coding, multimedia systems, and machine vision. He is a member of the Polish Committee for Standardization, Technical Committee for Multimedia, and serves as an expert in ISO/IEC JTC1/SC29 standardization working groups. His research interests include machine vision, data analysis, and artificial neural networks.



MAREK DOMAŃSKI (Life Senior Member, IEEE) received M.S., Ph.D., and Habilitation degrees from Poznan University of Technology, Poland in 1978, 1983, and 1990, respectively. Since 1993, he is a Professor at Poznan University of Technology, where he headed the Chair (Department) of Multimedia Telecommunications and Microelectronics. Since 2005, he has been the head of the Polish delegation to MPEG. He co-authored highly ranked technology proposals submitted in response to MPEG calls for scalable video compression (2004), 3D video coding (2011), and immersive video coding (2019). He also led the team that developed one of the very first AVC decoders for TV set-top boxes (2004) and several AVC, HEVC, and AAC-HE codec implementations and improvements. He is the author of 3 books and over 300 papers in journals and proceedings of international conferences. The papers are mainly on image, video, and audio compression, virtual navigation, free-viewpoint television, image processing, multimedia systems, 3D video and color image technology, digital filters, and multidimensional signal processing. He is co-inventor in several patents and pending patent applications in European and US patent offices. He has been general chairman/co-chairman and the host of several international conferences: Picture Coding Symposium, PCS 2012; IEEE Int. Conf. on Advanced and Signal-based Surveillance, AVSS 2013, European Signal Processing Conference, EUSIPCO 2007; 73rd and 112th Meetings of MPEG; Int. Workshop on Signals, Systems, and Image Processing, IWSSIP 1997 and 2004; Int. Conf. Signals and Electronic Systems, ICSES 2004, and others. He served as a member of various steering, program, and editorial committees of international journals and international conferences.