EVALUATION OF VIDEO CODING FOR MACHINES WITHOUT GROUND TRUTH

Kristian Fischer[1], *Markus Hofbauer*[2], *Christopher Kuhn*[2], *Eckehard Steinbach*[2], *André Kaup*[1]

[1]Multimedia Communications and Signal Processing, FAU Erlangen-Nürnberg
[2]Chair of Media Technology, Technical University of Munich

ABSTRACT

In the emerging field of video coding for machines, video datasets with pristine video quality and high-quality annotations are required for a comprehensive evaluation. However, existing video datasets with detailed annotations are severely limited in size and video quality. Thus, current methods have to either evaluate their codecs on still images or on already compressed data. To mitigate this problem, we propose an evaluation method based on pseudo ground-truth data from the field of semantic segmentation to the evaluation of video coding for machines. Through extensive evaluation, this paper shows that the proposed ground-truth-agnostic evaluation method results in an acceptable absolute measurement error below 0.7 percentage points on the Bjøntegaard Delta Rate compared to using the true ground truth for mid-range bitrates. We evaluate on the three tasks of semantic segmentation, instance segmentation, and object detection. Lastly, we utilize the ground-truth-agnostic method to measure the coding performances of the VVC compared against HEVC on the Cityscapes sequences. This reveals that the coding position has a significant influence on the task performance.

Index Terms— Video Coding for Machines, VVC, Semantic/Instance Segmentation, Pseudo Ground Truth

1. INTRODUCTION

Thanks to a drastic increase in accuracy and applicability, a significant amount of computer vision tasks is solved by neural networks today, such as in autonomous driving or surveillance of public spaces. In such applications, machines are directly interacting with each other without a human being involved. Hence, this type of communication is typically referred to as machine-to-machine (M2M) communication. Cisco underlines the relevance of this emerging topic by stating that half of all global devices and connections will be accounted to M2M communication in 2023 [1].

In practical applications, the input multimedia data are usually compressed and transmitted before being analyzed by the machine. This saves eventual transmission rate or storage space on hard disks. Together with the growing importance of M2M communication, this requires specialized compression schemes in M2M applications. Early work in this area optimized the compression to preserve extracted features [2, 3, 4]. More recent work focuses on coding for M2M communication, where a neural network analyzes the decoded image instead of a human observer [5, 6, 7, 8, 9, 10]. This topic is commonly referred to as video coding for machines (VCM). Additionally, the Moving Picture Experts Group (MPEG) founded an ad hoc group on VCM in 2019, with the objective of standardizing an optimal bitstream for M2M communication [11].

To evaluate codecs for a VCM scenario, an arbitrary algorithm or neural network is applied to the compressed input image and its performance is measured depending on the deteriorations present. Thus, there are three main conditions for a proper dataset: 1.) The dataset needs to contain uncompressed, pristine input data, such that the coding efficiency measurement is not falsified by already existing artifacts. 2.) The dataset requires properly labeled ground-truth (GT) data to measure the performance of the applied algorithm. 3.) The dataset needs to contain video data on top/instead of images, since practical scenarios are usually based on video streams. To the best of our knowledge, there is currently no dataset available that fulfills all three conditions. Thus, previous VCM research either sacrifices the condition of uncompressed input data [5, 6] or evaluates the coding frameworks on single images [7, 8, 9, 10].

To alleviate the problem of requiring hand-labeled annotations, which most datasets do not have, several papers from the research field of future semantic segmentation either turn to simulated data [12] or generate pseudo GT data. Since we focus on real world applications, we do not consider simulation-based approaches. Luc et al. [13] and Nabavi et al. [14] take the predictions of a state-of-the-art semantic segmentation network as pseudo GT to evaluate their future semantic segmentation models. Couprie et al. [15] applied pseudo GT annotations for the task of future instance segmentation for the Cityscapes [16] dataset. Aqqa et al. [17] already measured the VCM performance of four different state-of-the-art object detection networks on different H.264/AVC [18] compression levels taking the predictions on uncompressed data as GT. However, they do not provide any statement on how accurate their measurement is compared to GT-based evaluations.

The authors gratefully acknowledge that this work has been supported by the Deutsche Forschungsgemeinschaft (DFG) under contract number KA 926/10-1.

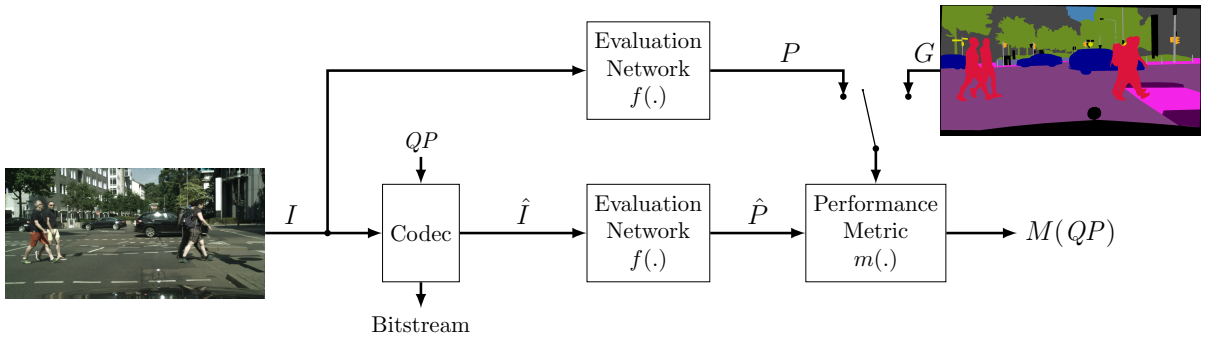

Fig. 1: VCM evaluation framework. The switch either selects the traditional evaluation framework with true GT data G, or the GT-agnostic evaluation with the pseudo GT data P.

Inspired by those approaches, this paper proposes to apply a GT-agnostic evaluation framework to evaluate VCM scenarios. We use the network predictions on uncompressed, pristine input data as pseudo GT. Thereby, this paper contributes an extensive analysis of the GT-agnostic evaluation framework and quantifies the measurement error compared with the traditional GT-based method. Second, with the help of the GT-agnostic method, we compare the two latest video codecs High Efficiency Video Coding (H.265/HEVC) [19] and its successor Versatile Video Coding (H.266/VVC) [20] for inter coding in *randomaccess* configuration on the non-labeled Cityscapes sequences.

2. ANALYZING VCM WITH GT-AGNOSTIC EVALUATION

2.1. Traditional VCM Evaluation with True GT

In a typical VCM scenario for image coding as depicted in Figure 1, the input image I is first encoded into a compact bitstream representation, which is steered by the user-defined quantization parameter (QP) towards either a low bitrate or a low distortion. Afterwards, the encoded bitstream is decoded resulting in the deteriorated output image $\hat{I}$. In classic compression frameworks focusing on human perception, $\hat{I}$ is compared against I to calculate distortion metrics representing the human visual system. In VCM however, $\hat{I}$ is fed into an arbitrary evaluation network $f(.)$, which fulfills a certain task resulting in the predictions $\hat{P}$. Subsequently, those predictions are compared against the ground-truth data G by a performance metric m that measures the performance of the evaluation network on the distorted input. The resulting performance score M can be formulated as

$$M(QP) = m(\hat{P}(QP), G) = m(f(\hat{I}(QP)), G). \quad (1)$$

We investigate the scenarios of semantic and instance detection/segmentation, using the mean intersection over union (mIoU) and the average precision (AP) [21] as performance metrics, respectively. Both return values between 0 and 1, with 0 being not accurate and 1 representing perfectly matched GT instances by the neural network. Ultimately, we plot the resulting performance $M(QP)$ over the required bitrate. This indicates how well the neural network performs on compressed input images.

2.2. Proposed GT-Agnostic Evaluation with Pseudo GT

Without the presence of properly annotated GT data G, we propose that pseudo GT labels are sufficient to evaluate the coding efficiency in VCM scenarios. As depicted in Figure 1, we consider the predictions P of the evaluation network on the pristine input data I as pseudo GT for our GT-agnostic evaluation framework:

$$G_{pseudo} = P = f(I). \quad (2)$$

With that, the overall performance measurement changes to

$$M_{pseudo}(QP) = m(f(\hat{I}(QP)), G_{pseudo}). \quad (3)$$

The predictions P can be considered as pseudo GT data for the predictions $\hat{P}$ derived from the compressed inputs. It is shown in the VCM-related literature as well as later in Sec. 4.1 that the predictions on compressed inputs have a lower performance than the pseudo-GT. This is similar to future instance segmentation, where the network segmenting the current frame has a higher performance than the networks predicting the segmentation masks based on the past.

In general, $M(QP)$ and $M_{pseudo}(QP)$ differ significantly. For the metric result $M(QP)$ measured on true GT, the maximum possible value is below 1, since real-world networks are typically not perfectly matching the GT. For the pseudo-GT based performance $M_{pseudo}(QP)$, this maximum is 1 when the predictions $\hat{P}$ on the compressed inputs are equal to the predictions on pristine data P.

However, the leading metric for comparing the coding quality of two codecs regarding VCM is the Bjøntegaard Delta Rate (BDR) [22], such as used in [5, 7, 9]. For VCM, it defines how much rate is saved while keeping the performance of the network the same for a certain quality range. Therefore, the absolute values of $M(QP)$ and $M_{pseudo}(QP)$ are not important, but rather their relative behavior over the required bitrate.

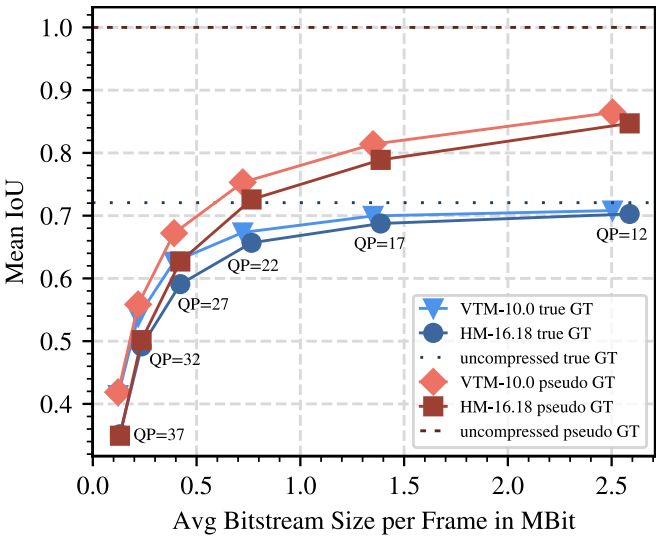

Fig. 2: Mean IoU measured with true GT (blue) and pseudo GT (red) for DeepLabV3+ over the bitrate when coding with VTM-10.0 and $QP \in \{12, 17, 22, 27, 32, 37\}$ for the 500 Cityscapes validation images. The dotted lines represent the corresponding mean IoU on uncompressed input images.

3. ANALYTICAL METHODS

To quantify the measurement error of the GT-agnostic VCM evaluation over the traditional evaluation method, we select the Cityscapes validation dataset comprising 500 pristine and annotated images with a spatial resolution of 1024×2048 pixels. We compress these images with the two standard-compliant codecs HEVC test model (HM-16.18) [23] and VVC test model (VTM-10.0) [24] for two different QP ranges. The first set of QP is $QP \in \{22, 27, 32, 37\}$ as defined in the JVET common testing conditions (CTC) [25]. Additionally, we test higher bitrate and quality ranges for $QP \in \{12, 17, 22, 27\}$ to obtain values that are closer to the neural network performance on pristine data similar to the investigations in [8]. To represent the human visual system, we also consider PSNR and VMAF [26].

As an evaluation network for the task of semantic image segmentation, we employ the state-of-the-art model DeepLabV3+ [27] with a Mobilenet [28] backbone and the pre-trained models from [29]. For instance segmentation, we use Mask R-CNN [30] with the weights from [31]. For object detection with Faster R-CNN, we trained a model similar to [7] on Cityscapes. Both R-CNN networks have a ResNet-50 backbone with a feature-pyramid structure [32].

We measure the performance of the semantic segmentation using the mIoU, which is the standard metric for evaluating on the Cityscapes dataset [16]. For instance segmentation, we measure the weighted AP (wAP), which weights the AP for the eight Cityscapes classes depending on the frequency of their occurrence as we have proposed earlier in [7].

Table 1: BDR values of VTM over HM in % with the corresponding quality or performance metric as reference for the 500 labeled Cityscapes validation images. $QP \in \{12, 17, 22, 27\}$.

	True GT	Pseudo GT	Diff.
PSNR	−16.62	-	-
VMAF	−30.51	-	-
mIoU DeepLabV3+	−30.06	−25.25	−4.81
oAcc DeepLabV3+	−40.04	−34.37	−5.67
frwAcc DeepLabV3+	−39.69	−34.07	−5.62
wAP Mask R-CNN	−7.48	−12.34	4.86
wAP Faster R-CNN	−9.00	−12.38	3.38

Table 2: BDR values of VTM over HM in % with the corresponding quality or performance metric as reference for the 500 labeled Cityscapes validation images. $QP \in \{22, 27, 32, 37\}$.

	True GT	Pseudo GT	Diff.
PSNR	−22.74	-	-
VMAF	−26.01	-	-
mIoU DeepLabV3+	−28.31	−27.62	−0.69
oAcc DeepLabV3+	−43.62	−42.96	−0.66
frwAcc DeepLabV3+	−43.49	−43.52	0.03
wAP Mask R-CNN	−13.30	−13.06	−0.24
wAP Faster R-CNN	−10.66	−11.22	0.56

4. EXPERIMENTAL RESULTS

4.1. Analysis of GT-agnostic VCM Framework

In Figure 2, we draw the rate-performance curves of HM and VTM coding the Cityscapes validation dataset depending on whether the mIoU for semantic segmentation has been measured with or without true GT. The blue colors represent the traditional performance measurement with true GT data, whereas the red colors represent the GT-agnostic measurement. There, the differences between both evaluation approaches regarding absolute mIoU values can be clearly seen. As mentioned in Sec. 2.2, the performance for the uncompressed GT-agnostic case is 1. For the traditional method, the performance for a low QP value of 12 is almost equal to the equivalent uncompressed case. For pseudo-GT, the performance drops drastically by around 15 percentage points when using input data compressed with $QP = 12$ instead of pristine input images. A possible explanation for this behavior is that the applied DeepLabV3+ network is not perfect, as indicated by the mIoU value of around 70 % evaluated on the true GT data. These model imperfections make the IoU measurement more sensitive to small noise-like differences between the uncompressed and the slightly compressed data, which ultimately results in this large drop. Considering the four highest QP values, the influence of the noise declines and the basic behavior between HM and VTM is very similar for both measurement methods.

Table 3: BDR values of VTM over HM in % and *randomaccess* configuration for the corresponding quality or performance metric using proposed pseudo GT for the Cityscapes validation sequences and $QP \in \{22, 27, 32, 37\}$.

	BDR
PSNR	−33.19
VMAF	−38.34
mIoU DeepLabV3+	−33.62
wAP Mask R-CNN	−21.04
wAP Faster R-CNN	−20.86

However, as mentioned in Sec. 2.2, the evaluation of VCM is focused on the relative differences between the different codecs regarding BDR instead of the absolute performance values. Thus, we calculate the BDR values of VTM over HM and list them in Table 1 for high bitrates and in Table 2 for the CTC-recommended values. This confirms the visual impression from Figure 2. For low QP values, the difference between measuring the BDR with or without the GT data is 4.81 percentage points for the semantic segmentation with DeepLabV3+ regarding mIoU. For the JVET-recommended bitrate ranges, the total difference between calculating the BDR with true GT or pseudo GT reduces to only 0.69 percentage points.

We additionally investigated other standard semantic segmentation metrics such as the overall accuracy (oAcc) and frequency weighted accuracy (frwAcc), resulting in similar differences between both evaluation methods. Besides, those values are also confirmed for the task of instance segmentation with Mask R-CNN and object detection with Faster R-CNN, with absolute errors of 0.24 percentage points and 0.56 percentage points, respectively. This shows that the GT-agnostic measurement is a well-suited tool for evaluating VCM scenarios without having hand-labeled GT data for the JVET-CTC recommended QP range.

4.2. Performance HEVC vs. VVC for Inter Coding

After demonstrating that the GT-agnostic evaluation method performs quite on par with the traditional one, we show the coding gains of VTM over HM for coding Cityscapes videos. For our evaluations, we compressed the frames 16 to 24 of a Cityscapes validation sequence with the originally labeled 20^{th} frame of one Cityscapes sequence being placed in the middle. The chosen codec configuration was *randomaccess*. In Table 3 we present the BDR values for the different distortion and performance metrics. The coding gains for VTM over HM are similar when coding for DeepLabV3+ compared to using PSNR and VMAF. For the tasks of instance segmentation and detection however, the coding gains of VTM are significantly lower compared to the other metrics. These values show the same tendency as in the intra coding case listed in Table 2 and extensively discussed in [7].

When coding videos for neural networks in practical applications, it is important that the detection performance is constant regardless of the position were the frame has been coded. However, as drawn in Figure 3, the detection quality of Mask R-CNN strongly fluctuates depending on when the frame is coded in the *randomaccess* order, with the highest detection performance for the I-frame. This indicates that *randomaccess* coding might be infeasible for certain practical VCM applications.

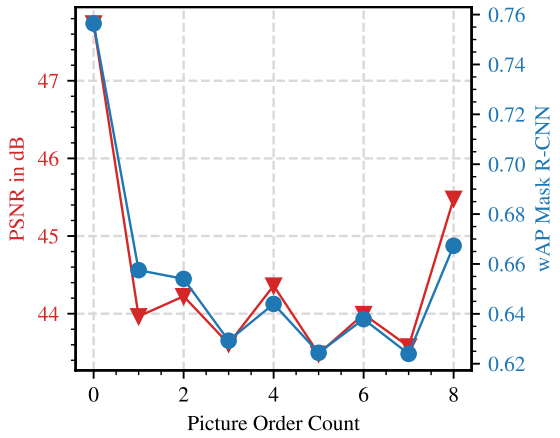

Fig. 3: PSNR (red) and pseudo-GT-based wAP (blue) for Mask R-CNN over the picture order count averaged over all sequences compressed with VTM-10.0 and $QP = 22$.

5. CONCLUSION

In this paper, we propose to evaluate VCM scenarios by employing pseudo-GT data derived from the predictions of computer vision models trained on uncompressed multimedia data. A broad analysis for the QP values defined in the JVET CTCs and three different use cases on the Cityscapes dataset revealed that measuring the coding performance between HM and VTM with this GT-agnostic method only results in negligible errors of up to 0.69 % in the worst case compared with the measurement using true GT data. Thus, our work suggests that future VCM evaluations can also be done by considering uncompressed video datasets without hand-labeled annotations such as the HEVC test sequences or self-captured datasets in practical applications, where annotating is commonly a cumbersome and costly process. The main limitation of this proposed evaluation scheme is that suitable machine-learning tasks and models have to be defined beforehand.

Finally, applying the GT-agnostic method to evaluate *randomaccess* coding on the Cityscapes sequences showed that the coding gains for instance segmentation and detection are not as high as for human-visual-based metrics. It has to be further researched whether the fluctuating detection performance depending on the coding order is negatively affecting practical applications. This could be of special interest for tasks requiring video data as input such as tracking.

6. REFERENCES

[1] Cisco, "Cisco annual internet report (20182023)," Tech. Rep., Cisco, Feb. 2020.

[2] Jianshu Chao and Eckehard Steinbach, "Preserving sift features in jpeg-encoded images," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 301–304.

[3] Jianshu Chao and Eckehard Steinbach, "Sift feature-preserving bit allocation for h. 264/avc video compression," in *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, pp. 709–712.

[4] Jianshu Chao, Hu Chen, and Eckehard Steinbach, "On the design of a novel jpeg quantization table for improved feature detection performance," in *2013 IEEE International Conference on Image Processing*. IEEE, 2013, pp. 1675–1679.

[5] Hyomin Choi and Ivan V. Bajic, "High efficiency compression for object detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 1792–1796.

[6] Leonardo Galteri, Marco Bertini, Lorenzo Seidenari, and Alberto Del Bimbo, "Video compression for object detection algorithms," in *Proc. International Conference on Pattern Recognition (ICPR)*, Aug. 2018, pp. 3007–3012.

[7] Kristian Fischer, Christian Herglotz, and André Kaup, "On intra video coding and in-loop filtering for neural object detection networks," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Oct. 2020, pp. 1147–1151.

[8] Kristian Fischer, Fabian Brand, Christian Herglotz, and André Kaup, "Video coding for machines with feature-based rate-distortion optimization," in *Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Sept. 2020, pp. 1–6.

[9] Nam Le, Honglei Zhang, Francesco Cricri, Ramin Ghaznavi-Youvalari, and Esa Rahtu, "Image coding for machines: an end-to-end learned approach," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, pp. 1590–1594.

[10] Kristian Fischer, Felix Fleckenstein, Christian Herglotz, and André Kaup, "Saliency-driven versatile video coding for neural object detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 2021, pp. 1505–1509.

[11] Yuan Zhang and Patrick Dong, "MPEG-M49944: Report of the AhG on VCM," Tech. Rep., Moving Picture Experts Group (MPEG) of ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Oct. 2019.

[12] Christopher B Kuhn, Markus Hofbauer, Ziqin Xu, Goran Petrovic, and Eckehard Steinbach, "Pixel-wise failure prediction for semantic video segmentation," in *Proc. IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 614–618.

[13] Pauline Luc, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun, "Predicting deeper into the future of semantic segmentation," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 648–657.

[14] Seyed S. Nabavi, Mrigank Rochan, and Yang Wang, "Future semantic segmentation with convolutional LSTM," in *Proc. British Machine Vision Conference (BMVC)*, Sept. 2018.

[15] Camille Couprie, Pauline Luc, and Jakob Verbeek, "Joint future semantic and instance segmentation prediction," in *Proc. European Conference on Computer Vision (ECCV)*, Sept. 2018, pp. 154–168.

[16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

[17] Miloud Aqqa, Pranav Mantini, and Shishir Shah, "Understanding how video quality affects object detection algorithms," in *Proc. Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, Jan. 2019, pp. 96–104.

[18] Thomas. Wiegand, Gary J. Sullivan, Gisle Bjontegaard, and Ajay Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 13, no. 7, pp. 560576, July 2003.

[19] Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[20] Jinale Chen, Yan Ye, and Seung Hwan Kim, "JVET-S2002: Algorithm description for versatile video coding and test model 10 (VTM 10)," Tech. Rep., Joint Video Exploration Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, July 2020.

[21] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik, "Simultaneous detection and segmentation," in *Proc. European Conference on Computer Vision (ECCV)*, Sept. 2014, pp. 297–312.

[22] Gisle Bjontegaard, "Calculation of average PSNR differences between RD-curves," *ITU-T VCEG and ISO/IEC MPEG document VCEG-MM33*, Apr. 2001.

[23] Joint Collaborative Team on Video Coding, "High efficiency video coding (HEVC)," .

[24] Joint Collaborative Team on Video Coding, "Versatile video coding (VVC)," .

[25] Frank Bossen, Jill Boyce, Karsten Suehring, Xiang Li, and Vadim Seregin, "JVET-T2010: JVET common test conditions and software reference configurations for SDR video," Tech. Rep., Joint Video Experts Team (JVET) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Oct. 2020.

[26] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara, "Toward a practical perceptual video quality metric," Tech. Rep., Netflix, https://medium.com/netflix-techblog/, June 2016.

[27] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. European Conference on Computer Vision (ECCV)*, Sept. 2018, pp. 833–851.

[28] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 4510–4520.

[29] Gongfan Fang, "DeepLabv3Plus-Pytorch," https://github.com/VainF/DeepLabV3Plus-Pytorch.

[30] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick, "Mask R-CNN," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2980–2988.

[31] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[32] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 936–944.