

ANALYSIS OF NEURAL IMAGE COMPRESSION NETWORKS FOR MACHINE-TO-MACHINE COMMUNICATION

Kristian Fischer, Christian Forsch, Christian Herglotz, and André Kaup

Multimedia Communications and Signal Processing
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
Cauerstr. 7, 91058 Erlangen, Germany

{Kristian.Fischer, Christian.Forsch, Christian.Herglotz, Andre.Kaup}@fau.de

ABSTRACT

Video and image coding for machines (VCM) is an emerging field that aims to develop compression methods resulting in optimal bitstreams when the decoded frames are analyzed by a neural network. Several approaches already exist improving classic hybrid codecs for this task. However, neural compression networks (NCNs) have made an enormous progress in coding images over the last years. Thus, it is reasonable to consider such NCNs, when the information sink at the decoder side is a neural network as well. Therefore, we build-up an evaluation framework analyzing the performance of four state-of-the-art NCNs, when a Mask R-CNN is segmenting objects from the decoded image. The compression performance is measured by the weighted average precision for the Cityscapes dataset. Based on that analysis, we find that networks with leaky ReLU as non-linearity and training with SSIM as distortion criteria results in the highest coding gains for the VCM task. Furthermore, it is shown that the GAN-based NCN architecture achieves the best coding performance and even out-performs the recently standardized Versatile Video Coding (VVC) for the given scenario.

Index Terms— Neural Compression Networks, Video Coding for Machines, Machine-to-Machine Communication

1. INTRODUCTION

Throughout the recent decades, image and video compression has been dominated by classic hybrid coding methods like Joint Picture Experts Group (JPEG) [1], High Efficiency Video Coding (HEVC) [2], and Versatile Video Coding (VVC) [3]. But with the rise of neural networks, multiple methods were proposed to train neural image compression networks (NCNs) end-to-end by balancing the contrary goals of a small bitstream and best possible image quality [4, 5, 6, 7]. Thereby, all those networks already provide a superior rate-distortion performance than JPEG coding.

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

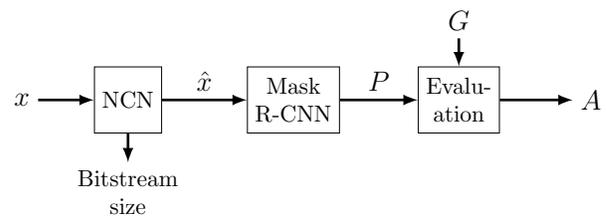


Fig. 1. Investigated framework coding input image x for neural networks. A denotes the accuracy of Mask R-CNN segmenting objects.

Another consequence of the tremendous advances in the field of neural networks is that more and more applications in everyday life, which perform tasks from the field of computer vision, are based on neural networks. Such networks are applied, e.g., for video surveillance, industrial processes, and autonomous driving. In most real-world applications, the multimedia data has first to be transmitted or stored from the capturing device before being analyzed by the neural network. This requires a suitable compression scheme, which is usually optimized for providing the best possible quality for the human visual system. But, as shown in [8], this does not always have to result in a high coding performance, when the decoded frame is analyzed by a neural network instead. Optimizing codecs such that the decoded frame can optimally be analyzed by a neural network is attributed to the field of video coding for machines (VCM), which is targeted by the MPEG ad-hoc group [9] founded in 2019. Besides, several other work was proposed designing or optimizing coding chains with classic hybrid codecs for such machine-to-machine (M2M) communication [10, 11, 12, 13].

Derived from the two before-mentioned developments, this paper deploys NCNs for the VCM task for the first time, investigating which architectures and parametrizations of NCNs are best suited for the VCM task. This provides valuable information to reach the ultimate objective of training such networks end-to-end for M2M communication.

For the investigations, we build up an M2M scenario as shown in Fig. 1, where the decoded frame \hat{x} is analyzed

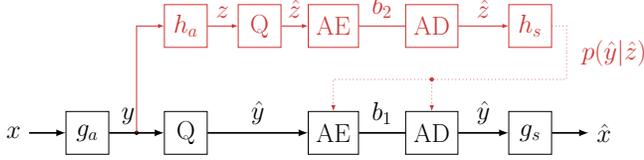


Fig. 2. *b2018* architecture in black; Additional hyperprior coding structure for *bmsbj2018* is depicted in red. AE and AD denote lossless arithmetic en- and decoding, respectively.

by the state-of-the-art instance segmentation network Mask R-CNN [14]. Its accuracy A is measured by comparing the detections P against the ground truth G over the required bitrate to obtain the coding efficiency of the investigated NCNs. Thereby, several architectures and methods of neural compression networks are tested including different non-linearities [4], different distortion metrics during the training process [5, 6], and a Generative Adversarial Network (GAN) [15] structure [7]. Besides, their performance is compared in relation to the commonly used perceptual distortion metrics PSNR, Structural Similarity index (SSIM) [16], and Video Multi-method Assessment Fusion (VMAF) [17]. In the final experiment, we compare the investigated neural compression networks against JPEG and the state-of-the-art video compression methods HEVC and VVC applied in all-intra configuration.

2. INVESTIGATED IMAGE COMPRESSION NETWORKS

2.1. Basic Neural Compression Network – *b2018*

In hybrid image or video codecs, transform coding is deployed to reduce statistical dependencies by transforming the residual image into a frequency domain. Subsequently, the coefficients are quantized and encoded by an entropy encoder to reduce the bitrate. As transformation, the linear Discrete Cosine Transform (DCT) is commonly selected and non-linear methods such as prediction are added to improve the performance for non-linear signals.

Contrary, for neural compression networks as depicted in Fig. 2 and proposed in [18] and [4], the transform is directly implemented as an analysis neural network $y = g_a(x, \phi)$ generating the latent space y from the input image x with the learned network weights ϕ . Subsequently, the latent space y is quantized into \hat{y} . By encoding \hat{y} losslessly and transmitting it to the decoder side, the decoded output image \hat{x} is obtained by applying an inverse synthesis transformation $\hat{x} = g_s(\hat{y}, \theta)$ parametrized by θ . The weights ϕ and θ are jointly trained by minimizing the loss function

$$\mathcal{L}(\phi, \theta, \psi) = \mathbb{E}_{x \sim p_x} [-\log_2 p_{\hat{y}}(\hat{y}, \psi) + \lambda \cdot d(x, \hat{x})]. \quad (1)$$

Thereby, the first summand holds the entropy H of \hat{y} with the estimated entropy model $p_{\hat{y}}$ and its parametrization ψ . The

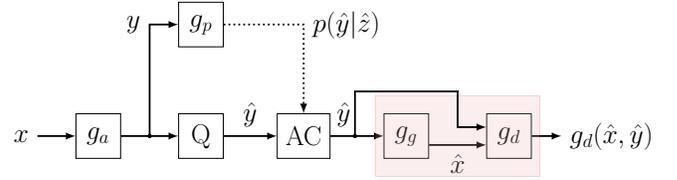


Fig. 3. *HiFiC* architecture; Arithmetic coding (AC) comprises AE and AD for simplicity. The red highlighted area represents GAN structure at the decoder side.

lower the entropy, the less bits b will be required to transmit \hat{y} to the decoder side. The second summand represents the distortion between the original image x and its reconstructed version \hat{x} measured by an arbitrary distortion function d . Typically, Mean Squared Error (MSE) is chosen as d . Similar to hybrid coding methods, λ steers the relaxation between low bitrate and high quality towards either direction.

Both transform networks, g_a and g_s , build an hourglass-shaped auto encoder structure to derive a latent space y with a lower dimensionality than x in order to achieve a more compact representation that can efficiently be transmitted to the decoder. They consist of convolution layers adapting the spatial resolution with a down- or upscaling stride. Subsequently, a Generalized Divisive Normalization (GDN) [19] non-linearity is applied, which is inspired from visual systems occurring in nature and increases statistical independence by normalizing inside a layer. Another non-linearity alternative is the leaky Rectified Linear Unit (LReLU), which is often used in classification and detection networks.

2.2. NCN with Additional Hyperprior – *bmsbj2018* and *mbt2018*

One major drawback of the architecture in *b2018* is that the latent space \hat{y} still holds spatial dependencies. Thus, a hyperprior is added to the *b2018* architecture in [5] and shown in Fig. 2, which includes a second auto encoder consisting of the analysis and synthesis networks h_a and h_s , respectively. This auto encoder obtains the statistical dependencies $p(\hat{y}|\hat{z})$ from a second latent space z . Thus, each element \hat{y}_i can be modeled by a Gaussian distribution with zero mean and the standard deviation being derived from this additional latent space \hat{z} . With this *bmsbj2018* model, the coding performance is significantly increased over *b2018*. Additionally, the *bmsbj2018* model is also proposed to be trained with the multi-scale SSIM metric as distortion function $d(x, \hat{x})$.

The successor of *bmsbj2018*, *mbt2018* [6], employs non-zero-mean Gaussian distributions to code \hat{x} . Additionally, an autoregressive context model is added to *bmsbj2018* to further improve the entropy coding step.

2.3. GAN-based NCN – *HiFiC*

The last considered model *High Fidelity Compression (HiFiC)* [7] is a GAN-based expansion of *bmsbj2018*. Its basic struc-

ture is depicted in Fig. 3. Similar to *bmshj2018*, the encoding network g_a generates a quantized latent space \hat{y} , which is entropy coded with the help of an additional hyperprior \hat{z} derived from g_p . The decoder is designed as a GAN being conditioned on \hat{y} . There, the generator network g_g is supposed to fool the discriminator network g_d by creating output images \hat{x} derived from \hat{y} that g_d falsely classifies as real world data, which results in superior subjective quality than *bmshj2018* or *mbt2018*. Besides, *HiFiC* is additionally trained with a distortion metric measured in a feature space of a neural network [20], which was also shown to be beneficial for VCM coding with VVC and Mask R-CNN in [13] by a similar metric.

3. ANALYTICAL METHODS

3.1. Dataset

In order to evaluate a framework as shown in Fig. 1, the 500 uncompressed images with a size of 1024×2048 pixels from Cityscapes [21] validation set are encoded with the different neural compression networks. These images are captured from a car’s windshield observing different road scenes. For each image, a pixel-wise annotation of eight different classes of road users is provided. With that, the Average Precision (AP) is calculated as proposed for Cityscapes [22] over the whole dataset and for each class in order to measure the Mask R-CNN accuracy. Ultimately, the AP values are weighted (wAP) according to the number of instances for each class as proposed in [8] and [13].

3.2. Employed Implementations

To compress the Cityscapes images at full resolution, we utilize the pre-trained NCN models provided by the original authors in [23] without further re-training. For *bmshj2018* and *mbt2018*, eight models exist covering different areas of the rate-distortion relaxation, whereas for *b2018* and *HiFiC* only four and three models are supplied, respectively.

As state-of-the-art hybrid intra video coding reference, the HEVC test model (HM 16.20) [24] and VVC test model (VTM 10.0) [3] are selected. Before applying these two codecs, the Cityscapes images provided as PNGs are first converted into YUV format with 4:2:0 downscaling and vice-versa before applying Mask R-CNN. In order to fit to the bitrate ranges provided with the NCN models, Quantization Parameter (QP) values of 12 to 42 in steps of 5 are chosen. Lastly, JPEG compression is investigated using the OpenCV library [25] with quality levels from 10 to 90 in steps of 10.

To detect the road users from the compressed images, the Detectron2 [26] framework is deployed. It provides a Mask R-CNN model with a ResNet-50 [27] backbone that has already been trained on the Cityscapes training images.

Table 1. BDR in % with respect to the listed quality metric using *b2018* with GDN as anchor for four quality levels.

	PSNR	VMAF	SSIM	wAP
<i>b2018-LReLU</i>	1.9	-2.4	2.0	-13.2

Table 2. BDR in % with respect to the listed quality metric using the corresponding codec trained with MSE as anchor for eight quality levels.

	PSNR	VMAF	SSIM	wAP
<i>bmshj2018-SSIM</i>	32.2	18.5	-35.2	-6.3
<i>mbt2018-SSIM</i>	46.7	36.6	-31.1	-1.0

3.3. Quality Metrics

In order to measure the performance of the different codecs for the human visual system, the quality metrics PSNR, SSIM, and VMAF are obtained. The wAP of Mask R-CNN being applied to the compressed images is taken to measure the performance for the M2M scenario. In order to quantify the resulting rate-distortion curves, the Bjøntegaard delta rate (BDR) [28] is calculated, which measures the bitrate savings for an identical quality. In addition to common BDR using PSNR, SSIM, and VMAF as quality metric, PSNR is also substituted with wAP to measure the VCM coding performance as it is recommended by MPEG VCM group [29].

4. EVALUATION RESULTS

4.1. Choice of Non-Linearity

The first experiment conducts a comparison between a *b2018* model build with GDN non-linearities that are optimized for compressing natural content for the human visual system and a model build with LReLUs. The BDR values for the different quality metrics are provided in Table 1. Choosing LReLU over GDN as non-linearity requires more bits to achieve the same PSNR. Contrary, the coding performance when coding for Mask R-CNN is significantly improved by selecting the LReLU model, which saves 13.2 % bitrate for the same wAP. Similar ReLU activations are also used in the ResNet backbone of Mask R-CNN, which is one possible explanation, why the LReLU-based *b2018* model outperforms the GDN-based model for M2M communication.

4.2. Influence of Training Distortion Metric

Another important influence on the performance of neural compression networks is the selected distortion metric throughout the training process. Here, the performances of the two compression networks *bmshj2018* and *mbt2018* are compared depending on whether they were trained with MSE or SSIM. The BDR results are listed in Table 2. Naturally, training the models on SSIM performs worse when measuring the output quality with PSNR as well as for VMAF, but immensely increases the coding performance with respect to

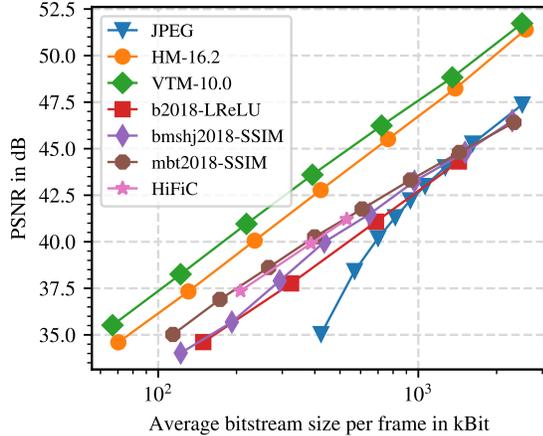


Fig. 4. PSNR over required bits for investigated codecs and the parametrization resulting in the best performance for VCM task and for the 500 Cityscapes validation images.

SSIM. Having the goal to achieve a high detection accuracy measured by wAP, the model should be trained with SSIM as distortion metric as well, saving up to 6.3 % and 1.0 % of bitrate, respectively. This can have multiple explanations. First, the model trained with SSIM focuses, as well as the evaluation network Mask R-CNN, on the structural information [16] of the content. Second, the *bmsbj2018* authors stated that their model trained on SSIM focuses on regions of low contrast by omitting information in high contrast areas. This accommodates the Mask R-CNN, which is struggling to segment objects that do not differ much from the background because they are for example located in the shadow of a building, which can occur throughout the Cityscapes dataset, and which gets amplified when adding quantization to the image.

4.3. Comparison of Neural Compression Networks against State-of-the-Art Compression Methods

In the final analysis, all chosen models from Section 2 with their best found parametrization are compared against the classic codecs JPEG, HEVC, and VVC. Figures 4 and 5 provide the rate-PSNR and rate-wAP curves, respectively. Table 3 lists the BDR of all codecs with the *b2018-LReLU* model as anchor. Among the NCNs, *mbt2018* with the corresponding training distortion metric performs best for PSNR, VMAF, and SSIM. However, the classic video codecs HEVC and VVC still achieve higher BDR savings.

Regarding the investigated VCM use case, *HiFiC* outperforms all other codecs, even performing better than the upcoming video coding standard VVC. The reason for this seems to be caused in the GAN-based structure of *HiFiC* and the neural-network-based distortion metric. During training, the network is pushed towards producing compressed images \hat{x} resulting in a high activation of the discriminator network g_d . That can be compared to the investigated VCM inference case providing images \hat{x} that result in the best possible detection and segmentation accuracy of Mask R-CNN.

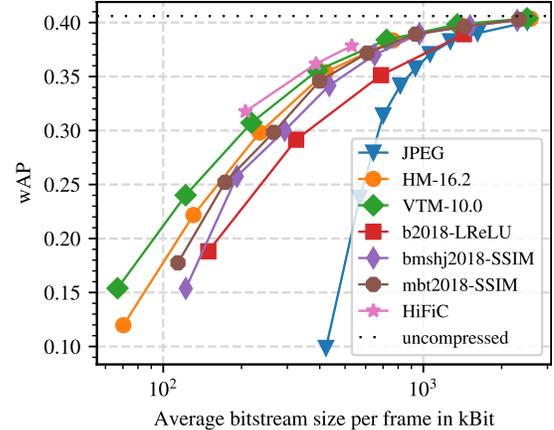


Fig. 5. wAP over required bits for investigated codecs and the parametrization resulting in the best performance for VCM task and for the 500 Cityscapes validation images. The black dotted line represents the accuracy when applying Mask R-CNN to the uncompressed Cityscapes images.

Table 3. BDR in % with respect to the listed quality metric using *b2018* with LReLU and trained on MSE as anchor. Highest bitrate savings for each quality metric is set in bold.

	PSNR	VMAF	SSIM	wAP
JPEG	38.0	-6.5	41.6	96.6
HM-16.20	-56.5	-61.3	-49.7	-33.5
VTM-10.0	-66.8	-71.3	-60.5	-43.2
<i>bmsbj2018-MSE</i>	-32.0	-44.5	-23.1	-15.5
<i>bmsbj2018-SSIM</i>	-11.9	-33.5	-53.5	-21.3
<i>mbt2018-MSE</i>	-49.2	-55.3	-41.6	-26.9
<i>mbt2018-SSIM</i>	-28.6	-39.8	-62.4	-27.9
<i>HiFiC</i>	-27.7	-49.4	-51.4	-52.8

5. CONCLUSIONS

This paper analyzed several neural compression networks according to their performance, when Mask R-CNN is applied to analyze the compressed images. The experiments first revealed that the *b2018* model with LReLU as activation function achieved a superior wAP-rate performance than the GDN-based model. Besides, training models with SSIM was shown to result in bitrate savings compared to standard training with MSE as distortion metric, when coding for Mask R-CNN. Moreover, the GAN-based network *HiFiC* outperformed all other NCNs and the state-of-the-art codecs for the given scenario. Additional experiments in future might find whether this is mostly caused by the GAN structure or the feature-based distortion metric applied in training. Derived from these promising results, future work will now aim for superior NCN coding performance for machines. This could be achieved by improving the training process with enhanced error metrics representing the behavior of image analysis networks and end-to-end training with Mask R-CNN.

6. REFERENCES

- [1] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. 18–34, Feb. 1992.
- [2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. 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, Sept. 2012.
- [3] J. Chen, Y. Ye, and S. H. 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 SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, July 2020.
- [4] J. Ballé, "Efficient nonlinear transforms for lossy image compression," in *Proc. Picture Coding Symposium (PCS)*, 2018, pp. 248–252.
- [5] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Proc. Int. Conference on Learning Representations (ICLR)*, Apr. 2018.
- [6] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in Neural Information Processing Systems*, vol. 31, pp. 10771–10780, Dec. 2018.
- [7] F. Mentzer, G. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," in *Proc. Advances in Neural Information Processing Systems*, Dec. 2020.
- [8] K. Fischer, C. Herglotz, and A. 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.
- [9] Y. Zhang and P. 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.
- [10] A. D. Bagdanov, M. Bertini, A. D. Bimbo, and L. Seidenari, "Adaptive video compression for video surveillance applications," in *IEEE International Symposium on Multimedia*. Dec. 2011, IEEE.
- [11] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *Proc. International Conference on Quality of Multimedia Experience (QoMEX)*. June 2016, IEEE.
- [12] L. Galteri, M. Bertini, L. Seidenari, and A. Del Bimbo, "Video compression for object detection algorithms," in *Proc. International Conference on Pattern Recognition (ICPR)*, Aug. 2018, pp. 3007–3012.
- [13] K. Fischer, F. Brand, C. Herglotz, and A. 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.
- [14] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2980–2988.
- [15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Dec. 2014, vol. 27, pp. 2672–2680.
- [16] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [17] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," Tech. Rep., Netflix, <https://medium.com/netflix-techblog/>, June 2016.
- [18] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Proc. International Conference on Learning Representations (ICLR)*, Apr. 2017.
- [19] J. Ballé, V. Laparra, and E. P. Simoncelli, "Density modeling of images using a generalized normalization transformation," in *Proc. International Conference on Learning Representations (ICLR)*, Jan. 2016.
- [20] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 586–595.
- [21] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 3213–3223.
- [22] M. Cordts and M. Omran, "The cityscapes dataset," <https://github.com/mcordts/cityscapesScripts>, 2017.
- [23] J. Ballé, S. J. Hwang, N. Johnston, and D. Minnen, "Tensorflow-compression," <https://github.com/tensorflow/compression>.
- [24] Joint Collaborative Team on Video Coding, "High efficiency video coding (HEVC)," <https://jvet.hhi.fraunhofer.de/>.
- [25] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.
- [26] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," <https://github.com/facebookresearch/detectron2>, 2019.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.
- [28] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," *ITU-T VCEG and ISO/IEC MPEG document VCEG-MM33*, Apr. 2001.
- [29] S. Liu, W. Gao, X. Xu, S.-P. Wang, C.-C. Lin, and T.-H. Li, "MPEG-M55583: [VCM] common test conditions, evaluation methodology and reporting template for VCM," Tech. Rep., Moving Picture Experts Group (MPEG) of ISO/IEC JTC1/SC29/WG2, Oct. 2020.