electronics

MDPI

Article

Compression of Multiscale Features of FPN with Channel-Wise Reduction for VCM

Dong-Ha Kim, Yong-Uk Yoon, Gyu-Woong Han, Byung Tae Oh and Jae-Gon Kim *

School of Electronics and Information Engineering, Korea Aerospace University, Goyang 10540, Republic of Korea; donghakim@kau.kr (D.-H.K.); yuyoon@kau.kr (Y.-U.Y.); gwhan@kau.kr (G.-W.H.); byungoh@kau.ac.kr (B.T.O.)
* Correspondence: jgkim@kau.ac.kr

Abstract: With the development of deep learning technology and the abundance of sensors, machine vision applications that utilize vast amounts of image/video data are rapidly increasing in the autonomous vehicle, video surveillance and smart city fields. However, achieving a more compact image/video representation and lower latency solutions is challenging for such machine-based applications. Therefore, it is essential to develop a more efficient video coding standard for machine vision applications. Currently, the Moving Picture Experts Group (MPEG) is developing a new standard called video coding for machines (VCM) with two tracks, each mainly dealing with compression of the input image/video (Track 2) and compression of the features extracted from it (Track 1). In this paper, an enhanced multiscale feature compression (E-MSFC) method is proposed to efficiently compress multiscale features generated by a feature pyramid network (FPN), which is the backbone network of machine vision networks specified in the VCM evaluation framework. The proposed E-MSFC reduces the feature channels to be included in a single feature map and compresses the feature map using versatile video coding (VVC), the latest video standard, rather than the single stream feature compression (SSFC) module in the existing MSFC. In addition, the performance of the E-MSFC is further enhanced by adding a bottom-up structure to the multiscale feature fusion (MSFF) module, which performs the channel-wise reduction in the E-MSFC. Experimental results reveal that the proposed E-MSFC significantly outperforms the VCM image anchor with a BD-rate gain of up to 85.94%, which includes an additional gain of 0.96% achieved by the MSFF with the bottom-up structure.

Keywords: video coding for machines (VCM); feature compression; feature pyramid network (FPN); VVC; multiscale feature compression (MSFC)

check for updates

Citation: Kim, D.-H.; Yoon, Y.-U.; Han, G.-W.; Oh, B.T.; Kim, J.-G. Compression of Multiscale Features of FPN with Channel-Wise Reduction for VCM. *Electronics* **2023**, *12*, 2767. https://doi.org/10.3390/electronics12132767

Academic Editor: George A. Papakostas

Received: 11 May 2023
Revised: 16 June 2023
Accepted: 20 June 2023
Published: 21 June 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Currently, video is the most dominant traffic type on the Internet. Particularly with the convergence of emerging technologies, such as 5G, artificial intelligence (AI) and the Internet of Things (IoT), an increasing number of videos are generated by edge devices and consumed by machines for various vision applications in different fields, including the autonomous vehicle, video surveillance and smart city fields. However, owing to the huge volume of the increasing video data, video compression is still a crucial challenge in machine vision applications. Traditional video coding standards aim to achieve the best quality for human consumption under a certain bitrate constraint by utilizing the characteristics of the human visual system (HVS). However, these standards may be inefficient for machine consumption of vision tasks, such as image classification, object detection and segmentation, owing to their different purposes and evaluation metrics. For example, in the vision task of object detection, important information lies in the position and shape of the objects rather than the background. Therefore, instead of compressing the entire image with high perceptual quality such as traditional standards, focusing on compressing only the crucial information for object detection tasks can lead to much more

efficient compression than conventional methods. Additionally, compressing the entire image to a similar quality level to meet the given bit rate using conventional methods may result in the loss of important information for object detection, such as the shape of objects. Consequently, this has necessitated the development of a more efficient video coding standard for machine consumption. Accordingly, the Moving Picture Expert Group (MPEG) is developing a new standard called video coding for machines (VCM) [1–3].

Currently, the MPEG VCM group is developing the new standard in two tracks: Track 2 and Track 1, which mainly deal with the compression of the input image/video and the compression of the feature extracted therefrom, respectively [4]. Figure 1a shows a possible processing pipeline for the feature compression considered in the VCM Track 1 [5], whereas Figure 1b show a processing pipeline for the image/video compression of the VCM Track 2. For both pipelines, when versatile video coding (VVC), which is the most recent video coding standard [6], is used as the codec, the machine vision task performances measured at the given bitrates are defined as a feature anchor and an image anchor of VCM, respectively, for the evaluation of the potential technologies [4].

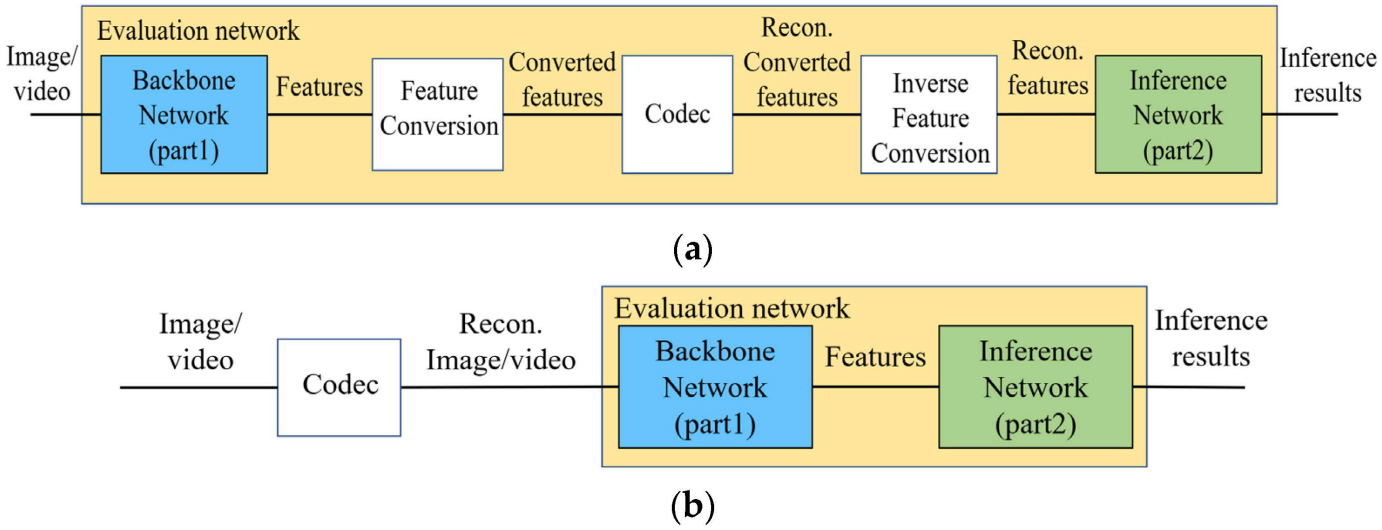


Figure 1. Possible processing pipelines of video coding for machines (VCM) [4]. (**a**) Image/video feature compression of VCM Track 1. (**b**) Image/video compression of VCM Track 2.

The VCM Track 1 explores the compression of multiscale features generated by the feature pyramid network (FPN), which is a backbone network of the machine vision networks selected for object detection tasks and segmentation tasks in the evaluation framework [4]. Generally, the compression of a feature pyramid, which consists of multiscale features with different sizes corresponding to each layer, is inefficient owing to the significant increase in size from the corresponding input image/video.

To effectively reduce the size of multiscale features, first, a framework named multiscale feature compression (MSFC) was proposed [7]. The MSFC framework integrates and compresses the multiscale features into a single feature map. As shown in Figure 2, the MSFC framework consists of three modules: a multiscale feature fusion (MSFF) module, a single stream feature compression (SSFC) module and a multiscale feature reconstruction (MSFR) module. The MSFC framework was introduced to VCM and modified the existing bottom-up MSFR structure to a top-down MSFR structure, which reconstructs low-level features from high-level features [8].

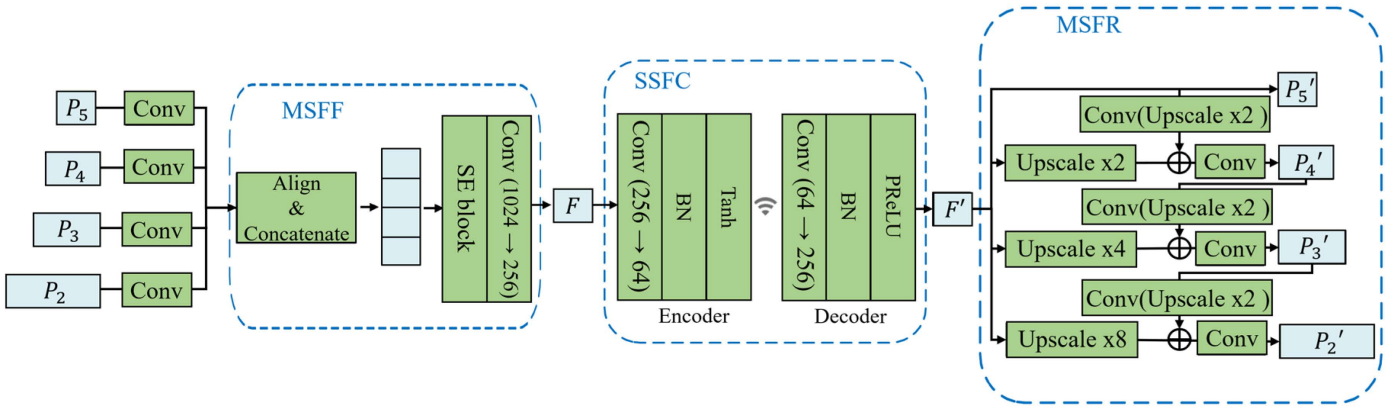


Figure 2. Structure of the existing multiscale feature compression (MSFC) frameworks [7,8].

However, the performance of feature compression using the existing MSFC model was not superior to the performance of the image anchor defined by the VCM group. This implies that the conventional approach of compressing images using the HVS is still more efficient. Therefore, this paper proposes an MSFC-based feature compression method that outperforms the performance of image anchors in terms of compression performance. In this paper, we propose an enhanced MSFC (E-MSFC) [9] based on the existing MSFC methods [7,8] to efficiently compress the multiscale features of FPN for object detection tasks. The proposed E-MSFC further reduces the feature channels to be packed into a single feature map and compresses the single feature map using VVC. For the further channel-wise reduction of the feature maps, the architectures of the MSFF and MSFR modules in the existing MSFC are extended in the E-MSFC and the SSFC module is replaced by VVC for the compression of the single feature map. The proposed E-MSFC method achieves much higher performance compared to the existing methods and significantly outperforms the VCM image anchor. It is expected that the proposed multiscale feature compression method using VVC can be a potential candidate solution for the feature compression of VCM.

The rest of this paper is organized as follows. In Section 2, the existing MSFC framework is briefly described and Section 3 presents the proposed E-MSFC by comparing it to the existing methods. In addition, an extension of the MSFF with a bottom-up structure in the E-MSFC is described in Section 3. The experimental conditions and results are presented in Section 4. Lastly, the conclusion is presented in Section 5.

2. Multiscale Feature Compression

The existing MSFC framework [7] compresses the feature pyramid of P-layer features labeled with $P_x(x = 2, 3, 4, 5)$, which is extracted from an FPN of the Faster R-CNN [10,11]. As shown in Table 1, the extracted features from FPN constitute a feature pyramid, which consists of five multiscale feature maps with different sizes. The compression of all the features may be undesirable as the overall size of the feature maps is tens of times larger than that of the input image. However, owing to the characteristics of FPN, there is a lot of redundancy in the feature pyramid extracted from FPN. The redundancy can be removed using an efficient compression method and a compact representation can be obtained. An example of such a method that can compress multiscale features in a compact representation is the MSFC. As shown in Figure 2, the MSFC framework consists of MSFF, SSFC and MSFR modules, which are briefly described in the following subsections.

Table 1. Details of the feature pyramid of multiscale features [7].

	Size (Channel, Height, Width)	**Bit-Depth**
Input image	3, h, w	24
P_2	256, h/4, w/4	32
P_3	256, h/8, w/8	
P_4	256, h/16, w/16	
P_5	256, h/32, w/32	
P_6	256, h/64, w/64	

2.1. MSFF Module

The MSFF module generates a single-frame feature map in a compact form with the size reduced as much as possible before the compression of the extracted FPN features. In this process, some features are resized in order to be aligned in a feature map, and redundancy between feature channels is reduced by CNNs. As shown in Figure 2, the MSFF [7] takes multiscale feature maps, P_x, and generates a single feature map, which is denoted as F. Each scale feature is resized and aligned to the size of the top-level feature map, P_5, using a convolutional layer, after which the features are concatenated. The concatenated feature map, which is illustrated as blue boxes in the MSFF, is reweighted according to the estimated importance of feature maps through a squeeze-and-excitation (SE) block [12]. Lastly, the channel-wise reduction is performed on the reweighted features to remove the redundancy between them using a convolutional layer. This ensures the fusion of the multiscale features of the pyramid into a single feature map, F, which consists of 256 feature channels, and each feature map has the same size as P_5.

2.2. SSFC Module

As shown in Figure 2, the SSFC module [7] consists of an encoder and a decoder. The encoder compresses the fused single feature map, F, through the channel-wise reduction on the feature maps from 256 to 64 channels using a convolution layer, subsequent batch normalization (BN) and the Tanh activation function. Lastly, the encoded feature elements with 32-bit depth are uniformly quantized to n-bit depth ($n = 2, 4, 8$) for further compression. In the decoder, the feature map with 256 channels is restored using a single convolution layer and the decoded single feature map, $F\prime$, is obtained through a subsequent batch normalization and the PReLU activation.

2.3. MSFR Module

As shown in Figure 2, a feature pyramid of multiscale features is reconstructed from the decoded single feature map, F', in the MSFR module [8]. The feature maps, P'_x, are reconstructed using a top-down architecture as follows. The feature map, F', is used as P'_5 as shown in (1). In this architecture, the feature map is upscaled using the nearest interpolation and the upscaled feature map is reconstructed using the convolutional layer. For the reconstruction of P'_4, after upscaling F' to the resolution of P'_4, as given by (2), it is added to the upscaled and convolved feature map of P'_5, after which a convolution layer is used. The feature maps P'_3 and P'_2 are reconstructed in a similar way as given by (3) and (4), respectively. In addition, the feature map, P'_6, is generated from P'_5 using a max pooling layer as in (5). The reconstruction of the feature maps, P'_x, mentioned above is summarized as follows:

$$P'_5 = F' \tag{1}$$

$$P'_4 = Conv\left(Conv\left(Nearest\left(P'_5, \times 2\right)\right) + Nearest\left(F', \times 2\right)\right) \tag{2}$$

$$P'_3 = Conv\left(Conv\left(Nearest\left(P'_4, \times 2\right)\right) + Nearest\left(F', \times 4\right)\right) \tag{3}$$

$$P_2' = Conv\left(Conv\left(Nearest\left(P_3', \times 2\right)\right) + Nearest\left(F', \times 8\right)\right) \quad (4)$$

$$P_6' = MaxPool\left(P_5'\right). \quad (5)$$

3. Proposed Enhanced MSFC

The SSFC module in the existing MSFC compresses the single feature map, F, which is the output of the MSFF module, using channel-wise reduction and uniform quantization. In contrast, in the exploration phase of VCM, a single feature map is compressed in various ways, such as using a neural network [13–15] or a conventional video codec [16–20]. In addition, compression using a conventional video codec or a neural network exhibits improved performance compared to simply quantizing the features [7,8]. Therefore, to enhance the machine vision task performance, this paper proposed an enhanced MSFC (E-MSFC) [9] model that combines the existing MSFC model and the feature compression method explored in the VCM.

As mentioned above, a single feature map fusing multiscale features can be compressed by utilizing neural networks or VVC. In the case of neural-network-based compression, a codec that outperforms VVC can be devised by training it to specialize in feature compression. However, this may require the training of different compression networks for given different target bitrates. For example, according to VCM's common test conditions (CTCs) [21], the performance of the machine vision task is compared to the anchor at six predefined bitrate points; thus, the different feature compression networks should be trained for the six bitrate points. In addition, devising an enhanced MSFC framework is time-consuming because the feature compression network may require re-training with evolving MSFF and/or MSFR structures.

To compress using VVC, the feature map of each channel is packed in a single-frame feature map, as shown in Figure 3. The packed feature map is then compressed using VVC. When VVC is used for feature compression, the task performances for the six bitrate points can be easily evaluated according to the CTCs of VCM. The performance of the feature compression using VVC may be lower than those of neural-network-based approaches [13–19]. However, although VVC is optimized for the compression of content to be consumed by humans based on HVS, the feature map can be effectively compressed using VVC. For example, Figure 4 indicates that the intraprediction of VVC works effectively in the compression of the feature map. In addition, it is relatively easy to develop an E-MSFC framework with MSFF and/or MSFR module modifications.

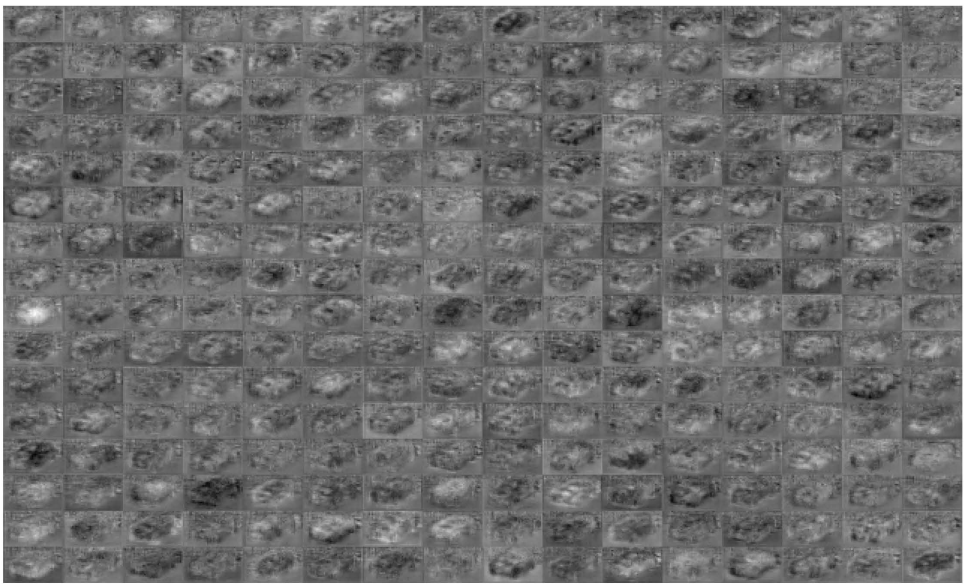


Figure 3. Single-frame feature map fusing multiscale features, F.

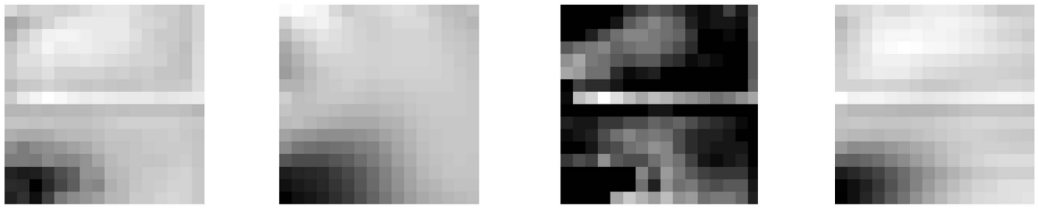


Figure 4. Intraprediction of VVC in the compression of a feature map (from the left: original block, intrapredicted block, residual block and reconstructed block).

Therefore, as shown in Figure 5, the proposed E-MSFC employs VVC as the core codec to compress the single feature map instead of the existing SSFC module. Thereafter, to develop an improved MSFC model in terms of the VCM performance, the structure of the MSFF module and MSFR modules of the existing MSFC were modified and extended in the proposed E-MSFC model. In the overall pipeline of feature compression, the proposed methods add an improved MSFC, which enhances performance with the slightly expanding structure of the existing MSFC, between the backbone and inference networks of the vision network. Therefore, considering the entire pipeline, excluding the feature map compression using the existing VVC, there is a minor increase in complexity with the proposed method.

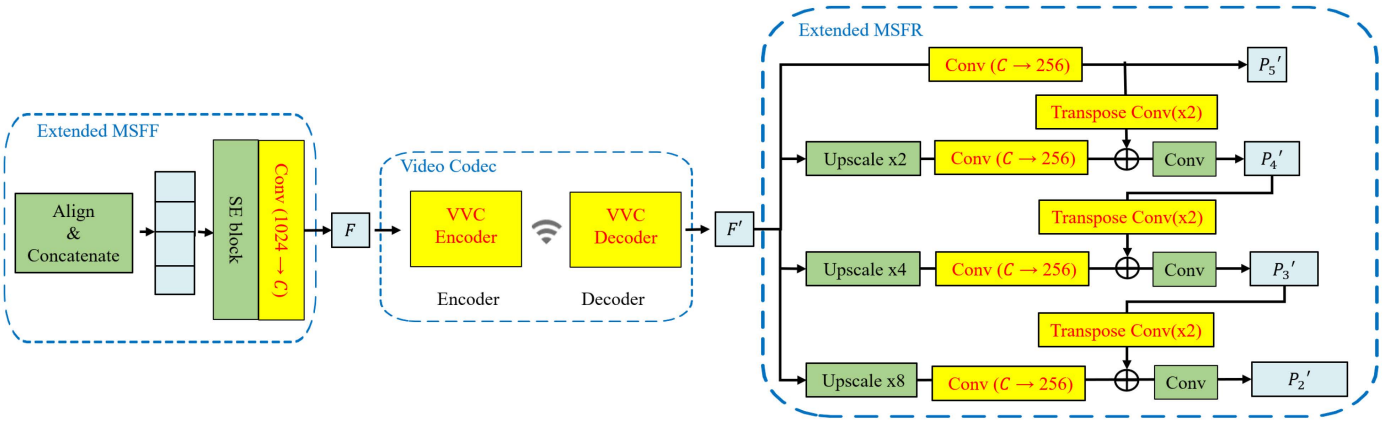


Figure 5. Structure of the proposed enhanced MSFC (E-MSFC) [9] (the extended components are indicated by yellow blocks).

3.1. Extension of MSFF and MSFR

Considering the trade-off between the bitrate and task performance, it is essential to appropriately determine the size of the single feature map to be compressed by VVC. The bitrate depends on the size of the single-frame feature map, which is determined by the number of feature channels constituting the single feature map. In this aspect, the E-MSFC reduces the number of feature channels constituting the single feature map, F, from 256 channels by extending the existing MSFF and MSFR structures to enhance the bitrate-task performance.

As shown in Figure 2, the number of the reweighted feature channels to be included in a single feature map in the existing MSFF is reduced from 1024 to 256 channels using a convolutional layer. In the extended MSFF, to reduce the size of the single feature map to be encoded using VVC, further channel-wise reduction is performed using a convolutional layer with smaller output channels than that of the existing MSFF's convolutional layer. In addition, the architecture of the existing MSFR is modified to reconstruct a feature pyramid from a single feature map generated in the extended MSFF with the reduced number of feature channels.

The architectures of the extended MSFF and the extended MSFR for further channel-wise reduction and its reconstruction in the E-MSFC are shown in Figure 5 [9]. The number of feature channels to be included in the single feature map, F, was reduced from 256 to

$C \in \{192, 144, 64\}$. In the extended MSFR module, the feature map, F', decoded by VVC is reconstructed to a feature pyramid, wherein each layer consists of 256 feature channels by the addition of a convolutional layer to each layer as follows. F' is reconstructed to P'_5 using a convolutional layer that restores the reduced feature channels to 256 feature channels. In addition, for the reconstruction of $P'_x(x = 2, 3, 4)$, a convolutional layer is added to restore the reduced channel of the upscaled feature, F', to 256 feature channels for each layer. Then, a transpose convolutional layer is used to upscale the feature map in the top-down architecture instead of using the nearest interpolation and convolutional layers as used in the existing MSFR. This ensures the reconstruction of the decoded single feature map, F', with the reduced feature channels to a feature pyramid with 256 feature channels. The process of aligning and concatenating proceeds in the same way as the existing MSFF, as described in Section 2.1.

To compress the single feature map, F, using VVC, the E-MSFC generates a single-frame feature map using min–max normalization. That is, each channel feature map constituting a single feature map, F, is spatially packed into a single frame using a raster-scan format in the ascending order of the channel index, as shown in Figure 3. Thereafter, each element of the packed frame is converted into a 10-bit depth format suitable for the encoding with VVC using the min–max normalization.

3.2. E-MSFC with a Bottom-Up MSFF

In addition, to improve the machine vision task performance, further extension of the MSFF module is employed in the E-MSFC. The feature pyramid of FPN consists of five multiscale feature maps: Higher-level feature maps exhibit a smaller size, as shown in Table 1. In the inference of object detection, higher-level feature maps are mainly used to detect large objects and lower-level feature maps are used to detect small objects. In this network, the feature maps, $P_x(x = 4, 5)$ and $P_x(x = 2, 3)$, can be regarded as higher-level feature maps and lower-level feature maps, respectively. According to the MSFR structures, lower-level feature maps are reconstructed from higher-level feature maps using a top-down structure. In such a top-down structure, information about lower-level features is likely to be lost. In addition, when the lower-level features, which have not been properly reconstructed, are used for the inference, small objects may not be correctly determined or even fail to be detected.

Therefore, to compensate for this shortcoming, the MSFF in the E-MSFC is further extended with a bottom-up structure to contain the information on lower-level features at higher-level features. This ensures an increase in the overall task performance via an improved detection of small objects. Therefore, we propose an MSFF extension with a bottom-up structure to contain the information on lower-level features at higher-level features [22].

As shown in Figure 6, the extended MSFF with a bottom-up structure includes additional preprocessing on the multiscale feature maps. To add the lower-level feature map information to the higher-level feature maps, the lower-level feature maps are downscaled and added to the higher-level feature maps. In detail, to add the information of the lowest-level feature map, P_2, to the upper-level feature map,P_3, the P_2 is downscaled using a convolution layer and added to P_3. Thereafter, the P'_3 is generated through a convolutional layer that fine-tunes the summed feature. Using the same process, P'_3 and P_4 are used to generate P'_4. This ensures the generation of the feature maps, P'_3 and P'_4, that contain lower-level feature map information. Lastly, the feature maps, P'_3 and P'_4, are used for the fusion of the single feature map instead of P_3 and P_4 in the bottom-up MSFF. In contrast, as P_5 contains the most important information for the inference, bottom-up processing is not applied to prevent possible information distortion. The fusion process of the single feature map mentioned above is summarized as follows:

$$P'_3 = Conv(Conv(P_2, \times 2) + P_3) \tag{6}$$

$$P'_4 = Conv\left(Conv\left(P'_3, \times 2\right) + P_4\right) \tag{7}$$

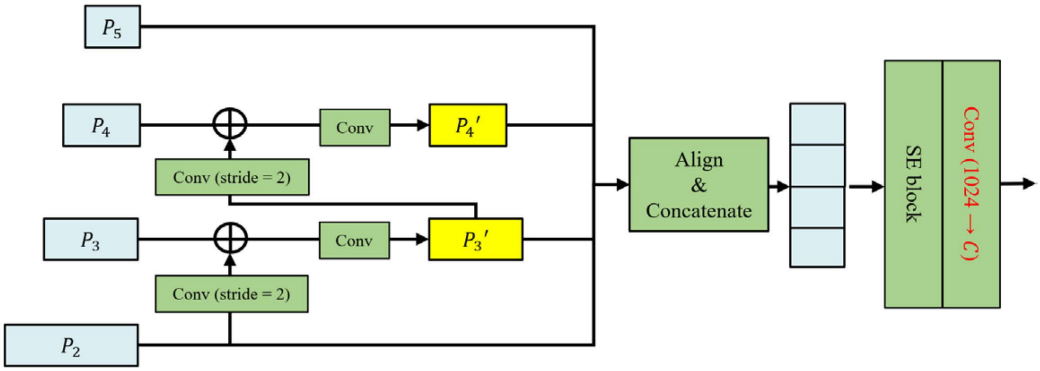


Figure 6. Structure of the bottom-up MSFF in the E-MSFC [22].

4. Experimental Results

As mentioned previously, in the proposed E-MSFC method, a feature pyramid of multiscale features, $P_x(x = 2, 3, 4, 5)$, extracted from FPN is compressed using VVC instead of the SSFC. To train the extended MSFF and MSFR modules of the E-MSFC, the E-MSFC was integrated into a Faster R-CNN X101-FPN of Detectron2 [23]. In the training, only the extended MSFF and MSFR modules were trained while freezing the parameters of the Faster R-CNN. The modules were trained using the COCO train2017 dataset [24] and an OpenImage V6 [25] validation dataset of 5 K images was used for the evaluation according to the CTCs of VCM [21]. The initial learning rate was set to 0.0005 and the training was iterated 300,000 times with a batch size of two. The extended MSFF and MSFR modules were separately trained according to the channel number, C, of the single feature map, F.

As mentioned previously, the single feature map to be compressed by VVC is a 10-bit depth format utilizing the min–max normalization. Therefore, the minimum and maximum values of F, each of which is a 32-bit floating-point value, should be transmitted to the decoder side to reconstruct the feature pyramid. Consequently, the data size of min/max values should be included in the bitrate calculation measured in bits per pixels (BPP). Figure 7 shows examples of the single-frame feature maps packing the fused single feature channels with different channel-wise reductions. After the channel-wise reduction, the size of the single-frame feature maps to be compressed decreases as the feature channel number decreases.

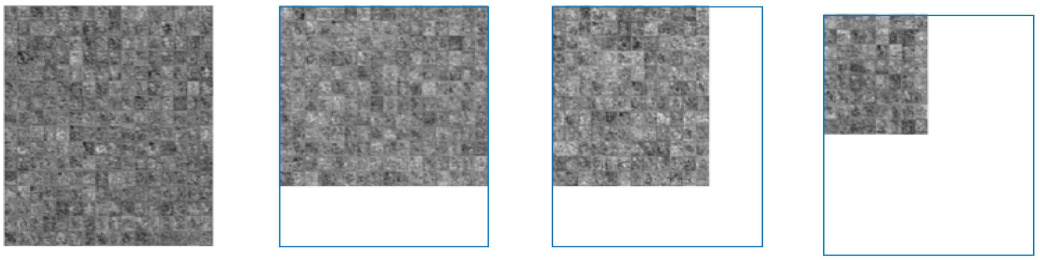


Figure 7. Example of single-frame maps packing a single feature with different channel-wise reductions in the E-MSFC. (**a**) 256. (**b**) 192. (**c**) 144. (**d**) 64.

According to the VCM CTCs, the VVC test model (VTM)-12.0 [26] was used as a video codec for the single-frame feature map compression. The experimental results of the overall performance were measured as mean average precision (mAP), which is the accuracy of the object detection task, at the given bitrate in BPP, that is, the BPP-mAP performance [21]. Table 2 shows the experimental results of the BPP-mAP performances for a set of quantization parameters (QPs) {22, 27, 32, 37, 42, 47} according to the channel numbers $C \in$ {256, 192, 144, 64} reduced from 256 using the proposed channel-wise reduction. The BD-rate gains of the proposed E-MSFC over the image anchor of VCM [4] are shown in the last row of Table 2. Figure 8 shows the BPP-mAP curves illustrating

the performance results shown in Table 2. Compared to the image anchor, the proposed method exhibited BD-rate gains of 43.17, 59.20, 65.40 and 84.98% when the number of the single feature channels is 256, 192, 144 and 64 channels, respectively. As shown in Figure 8, the overall compression efficiency improves as the number of channels decreases. However, as the maximum mAP over the entire bit rate range varied slightly with a change in the feature channel number, there can still be a considerable amount of redundancy within a single feature map. Therefore, the number of channels that can be reduced can be estimated from the performance change through the decrease in the number of channels.

Table 2. Experimental results of the BPP-mAP performances of the E-MFSC according to the channel number of the single feature map.

	E-MSFC 256 ch		**E-MSFC 192 ch**		**E-MSFC 144 ch**		**E-MSFC 64 ch**		**Image Anchor**	
	BPP	**mAP**	**BPP**	**mAP**	**BPP**	**mAP**	**BPP**	**mAP**	**BPP**	**mAP**
QP 22	0.657	78.828	0.521	78.804	0.396	78.316	0.177	78.711	0.863	78.796
QP 27	0.421	78.747	0.341	78.685	0.262	78.428	0.117	78.554	0.509	78.326
QP 32	0.221	78.753	0.181	78.544	0.141	78.357	0.064	78.811	0.287	76.998
QP 37	0.096	75.010	0.076	75.524	0.060	75.252	0.028	76.675	0.153	74.336
QP 42	0.039	66.510	0.030	67.262	0.024	66.812	0.012	67.665	0.078	68.957
QP 47	0.015	51.614	0.011	51.977	0.009	51.935	0.005	53.202	0.037	56.547
BD-rateGain	43.17%		59.20%		65.40%		84.98%			

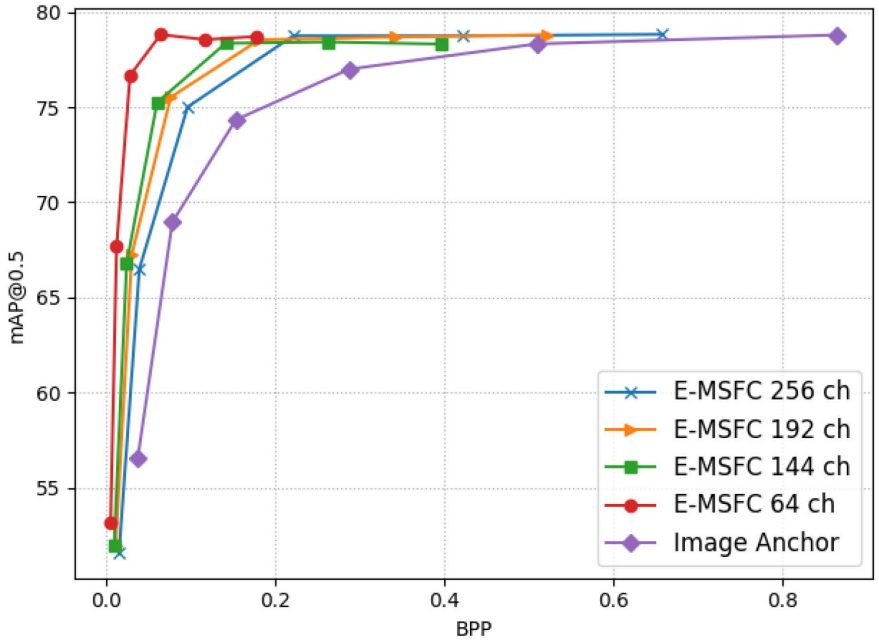


Figure 8. Experimental results of BPP-mAP performances of the E-MSFC.

The performance of the E-MFSC was also compared to those of the existing MSFC methods [7,8], which compresses 256-channel feature maps into 64-channel single feature maps and quantizes the 32-bit floating-point feature values into n-bit values ($n = 8, 4, 2$). The existing MSFC [8] was trained on the COCO train2017 dataset and evaluated using the COCO validation2017 dataset [24]. Table 3 shows the comparison results of the compression performances between the existing MSFC [8] and the proposed E-MFSC. In the comparison, the E-MSFC with a 64-channel single feature map was evaluated on the same COCO validation 2017 dataset. Figure 9 shows the BPP-mAP curves illustrating the performance results shown in Table 3. The E-MSFC significantly outperformed the existing MSFC,

indicating the high efficiency of VVC in the compression of feature maps. As shown in Table 3, even without compression, the proposed method exhibited improved performance than the existing method. This implies that the proposed E-MSFC generates feature maps that more efficiently contain the information required for machine vision tasks.

Table 3. Comparison of the single feature map compression performance between the existing MSFC and the proposed E-MSFC with COCO validation2017 dataset.

Existing MSFC			Proposed E-MSFC			Image Anchor	
	BPP	**mAP**		**BPP**	**mAP**	**BPP**	**mAP**
w/o compression	-	57.403	w/o compression	-	61.778	-	63.651
8-bit quantization	1.6581	57.388	QP 22	0.557	61.464	2.165	60.437
4-bit quantization	0.8291	56.831	QP 27	0.382	61.227	1.371	59.919
2-bit quantization	0.4145	44.940	QP 32	0.221	60.459	0.810	58.762
-	-	-	QP 37	0.104	54.218	0.444	55.289
-	-	-	QP 42	0.045	36.998	0.225	47.341
-	-	-	QP 47	0.022	22.712	0.103	34.039

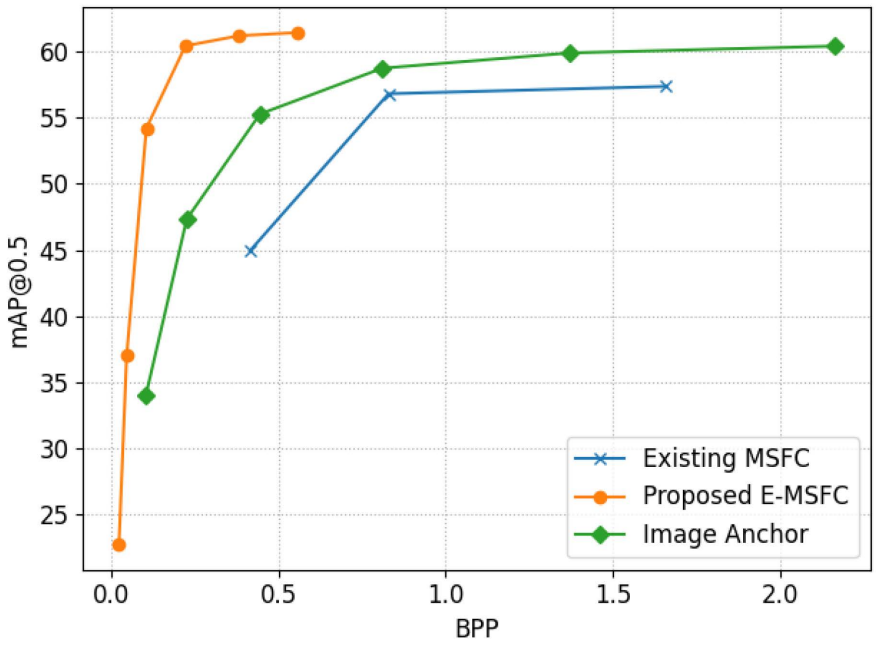


Figure 9. Comparison of the single feature map compression performance between the existing MSFC and the proposed method with COCO validation2017 dataset.

The E-MSFF with the bottom-up MSFF structure was trained under the same conditions used to train the E-MSFC. The performance of the E-MSFC with the bottom-up MSFF was compared to the E-MSFC where a single feature map contains 192 or 64 feature channels. As shown in Table 4, the E-MSFC with the bottom-up MSFF provides an additional BD-rate gain of 2.72% and 0.96% compared to the E-MFSC for the 192-channel and 64-channel feature maps, respectively. Figure 10 shows the BPP-mAP curve for each case of the feature map with 192 or 64 channels.

Table 4. Experimental results of the BPP-mAP performance of the E-MSFC with the bottom-up MSFF.

	E-MSFC 192 ch		**E-MSFC with Bottom-Up MSFC 192 ch**		**E-MSFC 64 ch**		**E-MSFC with Bottom-Up MSFC 64 ch**	
	BPP	**mAP**	**BPP**	**mAP**	**BPP**	**mAP**	**BPP**	**mAP**
QP 22	0.521	78.804	0.519	79.307	0.177	78.711	0.175	78.971
QP 27	0.341	78.685	0.340	79.409	0.117	78.554	0.116	78.777
QP 32	0.181	78.544	0.182	79.165	0.064	78.811	0.063	78.997
QP 37	0.076	75.524	0.077	76.469	0.028	76.675	0.027	76.862
QP 42	0.030	67.262	0.030	67.977	0.012	67.665	0.011	68.359
QP 47	0.011	51.977	0.011	52.388	0.005	53.202	0.004	53.545
BD-rate Gain	59.20%		61.92%		84.98%		85.94%	
	2.72%				0.96%			

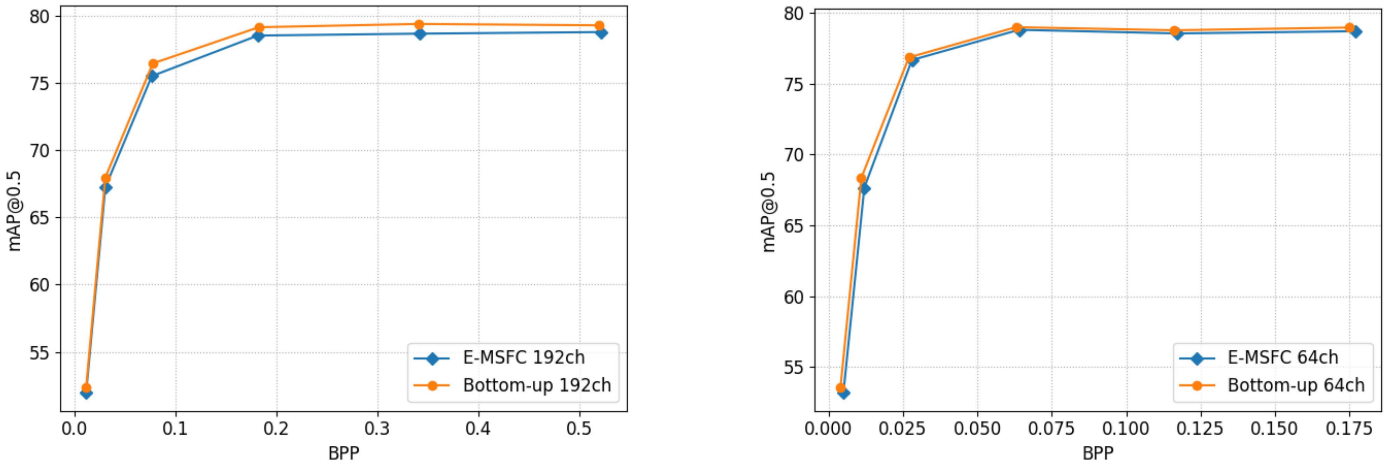


Figure 10. Experimental results of the BPP-mAP performance of the E-MSFC with the bottom-up MSFF. (**a**) 192-channel feature map. (**b**) 64-channel feature map.

Figure 11 shows the inference results of object detection when the E-MSFC and the E-MSFC with the bottom-up MSFF were applied to the original image. In the figure, the red circles indicate additionally detected objects compared to the E-MSFC. For instance, in Figure 11a, the E-MSFC initially fails to detect small birds, but after applying the bottom-up MSFF module, detection is successful. Similarly, in Figure 11b, the E-MSFC initially fails to detect persons who appear small due to distance, but detection is successful in the bottom-up MSFC. As a result, as shown in Figure 11a,b, in the case of E-MSFC, 12 objects are detected, whereas 15 objects are detected when the bottom-up MSFF is applied. The E-MSFC with the bottom-up MSFF, which embeds lower-level feature information in higher-level features, enabled the additional detection of smaller objects compared to the E-MSFC. Therefore, the E-MSFC with the bottom-up MSFF exhibits a better inference performance than the E-MSFC at the same bitrate.



Figure 11. Comparison of object detection inference results on COCO validation 2017 images ((**a**): 000000182441.jpg [width: 640, height: 399], (**b**): 000000486104.jpg [width: 500, height: 375]).

5. Conclusions

In this paper, we proposed an E-MSFC framework to efficiently compress multiscale features extracted from the feature pyramid network (FPN) and compressed using VVC. In the E-MSFC, multiscale features of a feature pyramid are fused and packed into a single feature map with further channel-wise reduction. The single feature map is compressed using VVC with min–max normalization rather than the existing SSFC. Thereafter, the compressed single feature map is transmitted, decoded and reconstructed to the feature pyramid with multiscale features to perform the object detection task. The structures of MSFR and MSFF modules in the existing MSFC were extended for channel-wise reduction and reconstruction. In addition, the MSFF module of the E-MSFC is further extended with a bottom-up structure to enhance the BPP-mAP performance by preserving information on lower-level features at higher-level features in the single feature map generation.

Experimental results revealed that the proposed E-MSFC significantly outperformed the VCM image anchor over a wider bitrate range with a BD-rate gain of up to 84.98%. In addition, the bottom-up MSFF further enhanced the performance of the E-MSFC with an additional BD-rate gain of up to 2.72%.

The proposed multiscale feature compression method using VVC was evaluated according to the experimental conditions defined by the VCM evaluation framework for the object detection vision task. However, multiscale features are common in various networks for different vision tasks, allowing the proposed method to be applied to a wide range of vision tasks and networks. Therefore, it is expected that the proposed method can be a potential candidate approach for the feature compression of VCM. The proposed method can be further enhanced by extracting more appropriate features to be compressed based on the recent works on compressed feature representation [27,28].

Author Contributions: Conceptualization, D.-H.K. and J.-G.K.; methodology, D.-H.K., Y.-U.Y., B.T.O. and J.-G.K.; software, D.-H.K. and G.-W.H.; validation, D.-H.K. and G.-W.H.; writing—original draft preparation, D.-H.K.; writing—review and editing, Y.-U.Y., B.T.O. and J.-G.K.; supervision, J.-G.K.; project administration, J.-G.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean government (MSIT), grant number 2020-0-00011 (Video Coding for Machine) and in part by the National Standards Technology Promotion Program of the Korean Agency for Technology and Standards (KATS) grant funded by the Korean government (MOTIE), grant number 20011687.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hollmann, C.; Liu, S.; Rafie, M.; Zhang, Y. Use cases and requirements for Video Coding for Machines. Moving Picture Experts Group (MPEG) of ISO/IEC JTC 1/SC 29/WG 2, Doc. N00190. In Proceedings of the 138th MPEG Meeting, Online, 25–29 April 2022.
2. Duan, L.; Liu, J.; Yang, W.; Huang, T.; Gao, W. Video coding for machines: A paradigm of collaborative compression and intelligent analytics. *Proc. IEEE Trans. Image Process* **2020**, *29*, 8680–8695. [CrossRef] [PubMed]
3. Zhang, Y.; Dong, P. AHG report: AHG on Video Coding for Machines. Moving Picture Experts Group (MPEG) of ISO/IEC JTC 1/SC 29/WG 2, Doc. m49944. In Proceedings of the 128th MPEG Meeting, Geneva, Switzerland, 12–16 October 2019.
4. Moving Picture Experts Group (MPEG). AHG report: Evaluation framework for Video coding for Machines. ISO/IEC JTC 1/SC 29/WG 2, Doc. N00162. In Proceedings of the 137th MPEG Meeting, Online, 17–21 January 2022.
5. Yu, L.; Pan, Y.; Rosewarne, C.; Gan, J.; Zhang, Y.; Wang, H.; Kim, Y.; Jeong, S.; Lee, J.; Do, J.; et al. AHG report: Draft description of exploration experiments on feature compression for VCM. Moving Picture Experts Group (MPEG) of ISO/IEC JTC 1/SC 29/WG 2, Doc. m58290. In Proceedings of the 136th MPEG Meeting, Online, 11–15 October 2021.
6. Bross, B.; Chen, J.; Ohm, J.-R.; Sullivan, G.J.; Wang, Y.-K. Developments in International Video Coding Standardization After AVC, With an Overview of Versatile Video Coding (VVC). *Proc. IEEE* **2021**, *109*, 1463–1493. [CrossRef]
7. Zhang, Z.; Wang, M.; Ma, M.; Li, J.; Fan, X. MSFC: Deep feature compression in multi-task network. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME) 2021, Shenzhen, China, 5–9 July 2021.
8. Han, H.; Choi, H.; Jung, S.; Kwak, S.; Yun, J.; Cheong, W.; Seo, J. AHG report: [VCM] investigation on deep feature compression framework for multi-task. Moving Picture Experts Group (MPEG) of ISO/IEC JTC 1/SC 29/WG 2, Doc. m58772. In Proceedings of the 137th MPEG Meeting, Online, 17–21 January 2022.
9. Kim, D.; Yoon, Y.-U.; Kim, J.-G.; Lee, J.; Kim, Y.; Jeong, S. AHG report: [VCM Track1] Compression of FPN multi-scale features for object detection using VVC. Moving Picture Experts Group (MPEG) of ISO/IEC JTC 1/SC 29/WG 2, Doc. m59562. In Proceedings of the 138th MPEG Meeting, Online, 25–29 April 2022.
10. Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the Computer Vision Pattern Recognition (CVPR), Honolulu, HI, USA, 19–21 July 2017.
11. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 7–12 December 2015.
12. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the Computer Vision Pattern Recognition (CVPR), Salt Lake City, UT, USA, 19–21 June 2018.
13. Do, J.; Lee, J.; Kim, Y.; Jeong, S.; Choi, J. AHG report: [VCM] Experimental results of feature compression using CompressAI. Moving Picture Experts Group (MPEG) of ISO/IEC JTC 1/SC 29/WG 2, Doc. m56716. In Proceedings of the 134th MPEG Meeting, Online, 26–30 April 2021.
14. Shao, Y.; Yu, L. [VCM] Coding Experiments of End-to-end Compression Network in VCM of ISO/IEC JTC 1/SC 29/WG 2, Doc. m54366. In Proceedings of the 131th MPEG Meeting, Online, 29 June–3 July 2020.
15. Yoon, Y.-U.; Kim, D.; Kim, J.-G.; Lee, J.; Do, J.; Jeong, S. [VCM] An approach of end-to-end feature compression network for object detection of ISO/IEC JTC 1/SC 29/WG 2, Doc. m58033. In Proceedings of the 136th MPEG Meeting, Online, 11–15 October 2021.
16. Kim, D.; Yoon, Y.-U.; Kim, J.-G. AHG report: [VCM] Compression of reordered feature sequences based on channel means for object detection, Moving Picture Experts Group (MPEG) of ISO/IEC JTC 1/SC 29/WG 2, Doc. m57497. In Proceedings of the 135th MPEG Meeting, Online, 12–16 July 2021.
17. Yoon, Y.-U.; Park, D.; Kim, J.; Chun, S.; Kim, J.-G. [VCM] Results of feature map coding for object segmentation on Cityscapes datasets of ISO/IEC JTC 1/SC 29/WG 2, Doc. m55152. In Proceedings of the 132th MPEG Meeting, Online, 12–16 October 2020.
18. Son, E.; Kim, C. [VCM] CNN Intermediate feature coding for object detection of ISO/IEC JTC 1/SC 29/WG 2, Doc. m54307. In Proceedings of the 131th MPEG Meeting, Online, 29 June–3 July 2020.
19. Wang, S.; Wang, Z.; Ye, Y.; Wang, S. [VCM] Image or video format of feature map compression for object detection of ISO/IEC JTC 1/SC 29/WG 2, Doc. m55786. In Proceedings of the 133th MPEG Meeting, Online, 11–15 January 2021.
20. Han, H.; Choi, H.; Kwak, S.; Yun, J.; Cheong, W.-S.; Seo, J. [VCM] Investigation on feature map channel reordering and compression for object detection of ISO/IEC JTC 1/SC 29/WG 2, Doc. m56653. In Proceedings of the 134th MPEG Meeting, Online, 26–30 April 2021.
21. Moving Picture Experts Group (MPEG). AHG report: Common test conditions and evaluation methodology for Video Coding for Machines. ISO/IEC JTC 1/SC 29/WG 2, Doc. N00192. In Proceedings of the 138th MPEG Meeting, Online, 25–29 April 2022.

22. Kim, D.; Yoon, Y.-U.; Kim, J.-G.; LEE, J.; Jeong, S. AHG report: [VCM-Track1] Performance of the enhanced MSFC with bottom-up MSFF. Moving Picture Experts Group (MPEG) of ISO/IEC JTC 1/SC 29/WG 2, Doc. m60197. In Proceedings of the 139th MPEG Meeting, Online, 18–22 July 2022.
23. Detectron2. Available online: https://github.com/facebookresearch/detectron2 (accessed on 10 November 2021).
24. COCOdataset 2017. Available online: https://cocodataset.org/#download (accessed on 15 September 2017).
25. OpenImages V6. Available online: https://storage.googleapis.com.openimages/web/index.html (accessed on 20 February 2020).
26. VVC Reference Software Version 12.0. Available online: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tags/VTM-12.0 (accessed on 23 February 2021).
27. Chu, H.; Wang, W.; Deng, L. Tiny-Crack-Net: A multiscale feature fusion network with attention mechanisms for segmentation of tiny cracks. *CACIE* **2022**, *37*, 1914–1931. [CrossRef]
28. Gong, R.; He, S.; Tian., T.; Chen, J.; Hao, Y.; Qiao, C. FRCNN-AA-CIF: An automatic detection model of colon polyps based on attention awareness and context information fusion. *Comput. Biol. Med.* **2023**, *158*, 106787. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.