



Article

Attention Networks for the Quality Enhancement of Light Field Images

Ionut Schiopu *  and Adrian Munteanu 

Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Pleinlaan 2, 1050 Brussels, Belgium; acmuntea@etrovub.be

* Correspondence: ischiopu@etrovub.be

Abstract: In this paper, we propose a novel filtering method based on deep attention networks for the quality enhancement of light field (LF) images captured by plenoptic cameras and compressed using the High Efficiency Video Coding (HEVC) standard. The proposed architecture was built using efficient complex processing blocks and novel attention-based residual blocks. The network takes advantage of the macro-pixel (MP) structure, specific to LF images, and processes each reconstructed MP in the luminance (Y) channel. The input patch is represented as a tensor that collects, from an MP neighbourhood, four Epipolar Plane Images (EPIs) at four different angles. The experimental results on a common LF image database showed high improvements over HEVC in terms of the structural similarity index (SSIM), with an average Y-Bjontegaard Delta (BD)-rate savings of 36.57%, and an average Y-BD-PSNR improvement of 2.301 dB. Increased performance was achieved when the HEVC built-in filtering methods were skipped. The visual results illustrate that the enhanced image contains sharper edges and more texture details. The ablation study provides two robust solutions to reduce the inference time by 44.6% and the network complexity by 74.7%. The results demonstrate the potential of attention networks for the quality enhancement of LF images encoded by HEVC.

Keywords: attention network; quality enhancement; light field images; video coding



Citation: Schiopu, I.; Munteanu, A. Attention Networks for the Quality Enhancement of Light Field Images. *Sensors* **2021**, *21*, 3246. <https://doi.org/10.3390/s21093246>

Academic Editor: Yun Zhang

Received: 6 April 2021

Accepted: 1 May 2021

Published: 7 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 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

In recent years, the technological breakthroughs in the sensor domain have made possible the development of new camera systems with steadily increasing resolutions and affordable prices for users. In contrast to conventional Red-Green-Blue (RGB) cameras, which only capture light intensity, plenoptic cameras provide the unique ability of distinguishing between the light rays that hit the camera sensor from different directions using microlens technology. To this end, the main lens of plenoptic cameras focus light rays onto a microlens plane, and each microlens captures the incoming light rays from different angles and directs them onto the camera sensor.

For each microlens, a camera sensor produces a so-called Macro-Pixel (MP). The raw LF image contains the entire information captured by the camera sensor, where the array of microlenses generates a corresponding array of MPs, a structure also known as lenslet images. Since each pixel in the MP corresponds to a specific direction of the incoming light, the lenslet image is typically arranged as an array of SubAperture Images (SAIs), where each SAI collects, from all MPs, one pixel at a specific position corresponding to a specific direction of the incoming light. The captured LF image can, thus, be represented as an array of SAIs corresponding to a camera array with a narrow baseline.

LF cameras have proven to be efficient passive devices for depth estimation. A broad variety of depth estimation techniques based on LF cameras have been proposed in the literature, including multi-stereo techniques [1,2], artificial intelligence-based methods [3] as well as combinations of multi-stereo and artificial intelligence-based techniques [4]. Accurately estimating depth is of paramount importance in view synthesis [5] and 3D reconstruction [6,7].

The LF domain was intensively studied during recent decades, and many solutions were proposed for each module in the LF processing pipeline, such as LF acquisition, representation, rendering, display, and LF coding. The LF coding approaches are usually divided into two major classes, including transform-based approaches and predictive-based approaches, depending on which module in the image or video codec is responsible for exploiting the LF correlations.

The transform-based approaches are designed to apply a specific type of transform, such as Discrete Cosine Transform [8,9], Discrete Wavelet Transform [10,11], Karhunen Loève Transform [12,13], or Graph Fourier Transform [14,15], to exploit the LF correlations.

However, the predictive-based approaches received more attention as they propose a more straightforward solution where different prediction methods are proposed to take advantage of the LF structure. These approaches propose to exploit the correlations between the SAIs using the coding tools in the High Efficiency Video Coding (HEVC) standard [16].

The pseudo-video-sequence-based approach proposes to select a set of evenly distributed SAIs as intra-coded frames and the remaining SAIs as inter-coded frames, e.g., [17,18]. In [19,20], the non-local spatial correlation is exploited when using the lenslet representation. The view-synthesis-based approach proposes to encode only a sparse set of reference SAIs and additional geometry information and then to synthesize the remaining SAIs at the decoder side [21,22]. In this work, we first employ HEVC [16] to encode the SAI video sequence and then to enhance the reconstructed lenslet image. The proposed Convolutional Neural Network (CNN)-based filtering method can be used to post-process any HEVC-based solution.

The attention mechanism was first proposed in the machine translation domain [23]. The main idea is that instead of building a single context vector, it is better to create weighted shortcuts between the context vector and the entire source input. This revolutionary concept now provides outstanding improvements in different domains, such as hyperspectral image classification [24], deblurring [25], image super-resolution [26], traffic sign recognition [27], and small object detection [28], to name a few. Many different network architectures have leveraged the attention mechanism to significantly improve over the state-of-the-art. In this work, an attention-based residual block is introduced to help the architecture learn and focus more on the most important information in the current MP context.

In our prior work, research efforts were invested to provide innovative solutions for LF coding based on efficient Deep-Learning (DL)-based prediction methods [20,29–32] and CNN-based filtering methods for quality enhancement [33,34]. In [29], we introduced a lossless codec for LF images based on context modeling of SAI images. In [30], we proposed an MP prediction method based on neural networks for the lossless compression of LF images.

In [31], we proposed to employ a DL-based method to synthesize an entire LF image based on different configurations of reference SAIs and then to employ an MP-wise prediction method to losslessly encode the remaining views. In [32], we proposed a residual-error prediction method based on deep learning and a context-tree based bit-plane codec, where the experimental evaluation was carried out on photographic images, LF images, and video sequences. In [20], the MP was used as an elementary coding unit instead of HEVC's traditional block-based coding structure for lossy compression of LF images. In recent work, we focused on researching novel CNN-based filtering methods.

In [33], we proposed a frame-wise CNN-based filtering method for enhancing the quality of HEVC-decoded videos. In [34], we proposed an MP-wise CNN-based filtering method for the quality enhancement of LF images. The goal of this paper is to further advance our findings in [34] by introducing a novel filtering method based on attention networks, where the proposed architecture is built based on efficient processing blocks and attention-based residual blocks and operates on Epipolar Plane Images (EPI)-based input patches.

In summary, the novel contributions of this paper are as follows:

- (1) A novel CNN-based filtering method is proposed for enhancing the quality of LF images encoded using HEVC [16].
- (2) A novel neural network architecture design for the quality enhancement of LF images is proposed using an efficient complex Processing Block (PB) and a novel Attention-based Residual Block (ARB).
- (3) The proposed CNN-based filtering method follows an MP-wise filtering approach to take advantage of the specific LF structure.
- (4) The input patch is designed as a tensor of four MP volumes corresponding to four EPs at four different angles (0° , 45° , 90° , and 135°).
- (5) The elaborated experimental validation carried out on the EPFL LF dataset [35] demonstrates the potential of attention networks for the quality enhancement of LF images.

The remainder of this paper is organized as follows. Section 2 presents an overview of the state-of-the-art methods for quality enhancement. In Section 3, we describe the proposed CNN-based filtering method. Section 4 presents the experimental validation on LF images. Finally, in Section 5, we draw our conclusions from this work.

2. Related Work

In recent years, many coding solutions based on machine learning techniques have rapidly gained popularity by proposing to simply replace specific task-oriented coding tools in the HEVC coding framework [16] with powerful DL-based equivalents. The filtering task was widely studied, and many DL-based filtering tools for quality enhancement were introduced to reduce the effects of coding artifacts in the reconstructed video.

The first DL-based quality enhancement tools were proposed for image post-filtering. In [36], the Artifact Reduction CNN (AR-CNN) architecture was proposed to reduce the effect of the coding artifacts in JPEG compressed images. In [37], a more complex architecture with hierarchical skip connections was proposed. A dual (pixel and transform) domain-based filtering method was proposed in [38]. A discriminator loss, as in Generative Adversarial Networks (GANs), was proposed in [39]. An iterative post-filtering method based on a recurrent neural network was proposed in [40].

Inspired by AR-CNN [36], the Variable-filter-size Residue-learning CNN (VRCNN) architecture was proposed in [41]. The inter-picture correlation is used by processing multiple neighboring frames to enhance one frame using a CNN [42]. In [43], the authors proposed to make use of mean- and boundary-based masks generated by HEVC partitioning. In [44], a CNN processes the intra prediction signal and the decoded residual signal. In [45], a CNN processes the QP value and the decoded frame. In [46], the CNN operates on input patches designed based on additional information extracted from the HEVC decoder, which specifies the current QP value and the CU partitioning maps.

In another approach, the authors proposed to replace the HEVC built-in in-loop filtering, the Deblocking Filter (DBF) [47], and the Sample Adaptive Offset (SAO) [48]. This is a more demanding task as, in this case, the filtered frame enters the coding loop and serves as a reference to other frames. In [49], a CNN was used to replace the SAO filter. Similarly, in [50], a deep CNN was applied after SAO and was controlled by the frame- and coding tree unit (CTU)-level flags.

In [51], the authors used a deep residual network to estimate the lost details. In [52], the Multistage Attention CNN (MACNN) architecture was introduced to replace the HEVC in-loop filters. Other coding solutions focus on inserting new filtering blocks in the HEVC framework. In [53], an adaptive, in-loop filtering algorithm was proposed using an image nonlocal prior, which collaborates with the existing DBF and SAO in HEVC. In [54], a residual highway CNN (RHCNN) was applied after the SAO filter. In [55], a content-aware CNN-based in-loop filtering method was integrated in HEVC after the SAO built-in filter.

In this work, we propose to employ the attention mechanism for the quality enhancement of LF images (represented as lenslet images) by following an MP-wise filtering

approach. Our experiments show that an increased coding performance was achieved when the SAI video sequence was encoded by running HEVC without its built-in filtering methods, DBF [47] and SAO [48].

3. Proposed Method

In the literature, the LF image is usually represented as a 5D structure denoted by $LF(p, q, x, y, c)$, where the (p, q) pair denotes the pixel location in an MP matrix, usually of $N \times N$ resolution; the (x, y) pair denotes the pixel location in an SAI matrix of size $W \times H$; and c denotes the primary color channel, $c = 1, 2, 3$. Let us denote $MP_{x,y} = LF(:, :, x, y, c)$ as the MP captured by the microlens at position (x, y) in the microlens array; $SAI_{p,q} = LF(p, q, :, :, c)$ as the SAI corresponding to view (p, q) in the SAI stack; and LL as the lenslet image of size $NH \times NW$, which is defined as follows:

$$LL((x-1)N+1 : xN, (y-1)N+1 : yN, c) = MP_{x,y}, \forall x = 1 : W, \forall y = 1 : H. \quad (1)$$

The experiments were conducted using the EPFL LF dataset [35] where $N = 15$ and $W \times H = 625 \times 434$. The LF images were first color-transformed from the RGB color-space to the YUV color-space, and only the Y (luminance) channel was enhanced. Therefore, $c = 1$ and $MP_{x,y}$ were of size 15×15 .

In this paper, a novel CNN-based filtering method is proposed to enhance the quality of LF images encoded using the HEVC video coding standard [16]. Figure 1 depicts the proposed CNN-based filtering scheme. The LF image, represented as an array of SAIs, is first arranged as an SAI video sequence and then encoded by the reference software implementation of HEVC called HM (HEVC Test Model) [56] under the All Intra (AI) profile [57]. Any profile can be used to encode the SAI video sequence as the proposed CNN-based filtering scheme is applied to the entire SAI video sequence. Therefore, in this work, a raster scan order is used to generate the SAI video sequence, while in the literature, a spiral order starting from the center view and looping in a clockwise manner towards the edge views is used to generate the SAI video sequence. Next, the reconstructed SAI sequence is arranged as a lenslet image using Equation (1), and EPI-based input patches were extracted from the reconstructed lenslet image, see Section 3.1.

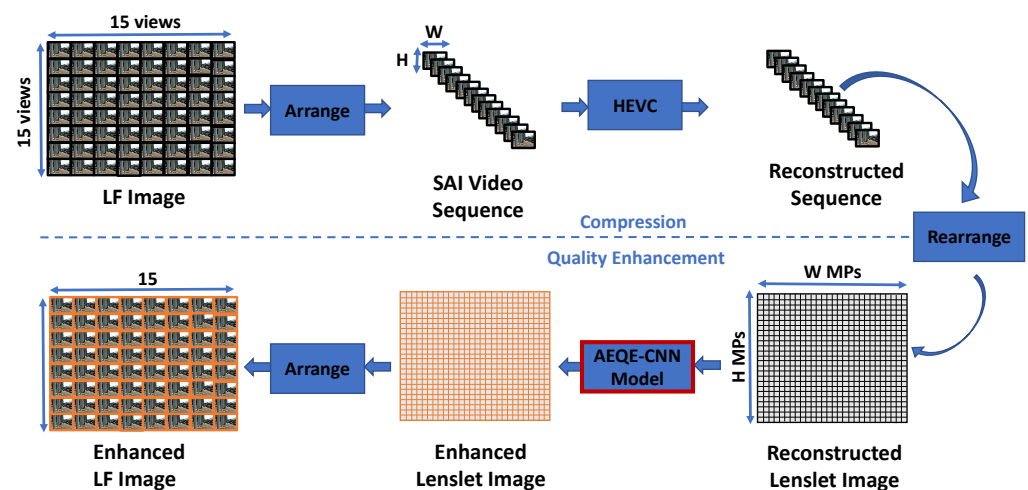


Figure 1. The proposed CNN-based filtering scheme. (Top) Compression: The LF Image (represented as an array of SAI) is arranged as a SAI video sequence and then encoded by HEVC. (Bottom) Quality Enhancement: The reconstructed sequence is arranged as a lenslet image (represented as an array of MPs) and each MP is enhanced by the proposed CNN-based filtering method using an AEQE-CNN model.

A CNN model with the proposed novel deep neural architecture called Attention-aware EPI-based Quality Enhancement Convolutional Neural Network (AEQE-CNN),

see Section 3.2, processed the input patches to enhance the MPs and obtain the enhanced lenslet image. Finally, the enhanced lenslet image is arranged as a LF image to be easily consumed by users.

Section 3.1 presents the proposed algorithm used to extract the EPI-based input patches. Section 3.2 describes in detail the network design of the proposed AEQE-CNN architecture. Section 3.3 presents the training details.

3.1. Input Patch

In this paper, input patches of size $15 \times 15 \times 9 \times 4$ were extracted from the reconstructed lenslet image. More exactly, the input patch concatenated four EPIs corresponding to 0° (horizontal EPI), 45° (first diagonal EPI), 90° (vertical EPI), and 135° (second diagonal EPI) from the MP neighbourhood of $b = 4$ MPs around the current MP, as depicted in Figure 2. Let us denote $\mathcal{N}_{x,y}$ as the MP neighbourhood around the current MP, $MP_{x,y}$, where

$$\mathcal{N}_{x,y} = \begin{bmatrix} MP_{x-b,y-b} & \dots & MP_{x-b,y} & \dots & MP_{x-b,y+b} \\ \vdots & & \vdots & & \vdots \\ MP_{x,y-b} & \dots & MP_{x,y} & \dots & MP_{x,y+b} \\ \vdots & & \vdots & & \vdots \\ MP_{x+b,y-b} & \dots & MP_{x+b,y} & \dots & MP_{x+b,y+b} \end{bmatrix}. \quad (2)$$

Four EPIs of size $N \times N \times (2b + 1) = 15 \times 15 \times 9$ were extracted from $\mathcal{N}_{x,y}$ as follows:

- (1) The 0° EPI of MP volume: $[MP_{x,y-b} \ MP_{x,y-b+1} \ \dots \ MP_{x,y+b}]$;
- (2) The 45° EPI of MP volume: $[MP_{x-b,y-b} \ MP_{x-b+1,y-b+1} \ \dots \ MP_{x+b,y+b}]$;
- (3) The 90° EPI of MP volume: $[MP_{x-b,y} \ MP_{x-b+1,y} \ \dots \ MP_{x+b,y}]$; and
- (4) The 135° EPI of MP volume: $[MP_{x+b,y-b} \ MP_{x+(b-1),y-(b-1)} \ \dots \ MP_{x-b,y+b}]$.

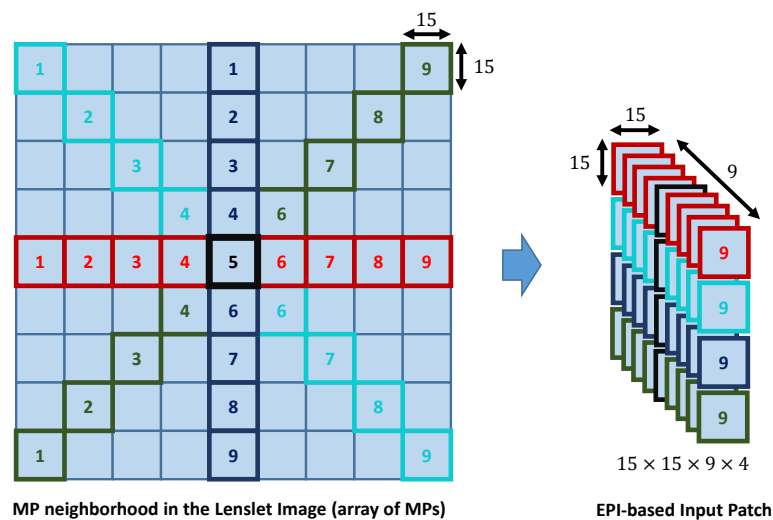


Figure 2. Extraction of the EPI-based input patch from the lenslet image represented as an array of MPs. Four EPIs are selected: 0° (horizontal) EPI marked with red, 45° (first diagonal) EPI marked with cyan, 90° (vertical) EPI marked with blue, and 135° (second diagonal) EPI marked with green. The current MP is marked with black.

The four EPIs were processed separately by the AEQE-CNN architecture as described in the following section.

3.2. Network Design

Figure 3 depicts the proposed deep neural network architecture. AEQE-CNN is designed to process the EPI-based input patches using efficient processing blocks and attention-based residual blocks. 3D Convolutional layers (Conv3D) equipped with $3 \times 3 \times 3$ kernels are used throughout the network architecture.

AEQE-CNN was built using the following types of blocks depicted in Figure 4: (i) the Convolutional Block (CB) contains a sequence of one Conv3D, one batch normalization (BN) layer [58], and one Rectified Linear Unit (ReLU) activation function; (ii) the proposed Processing Block (PB) contains a two branch design with one and two CB blocks where the feature maps of the two branches are concatenated to obtain the output feature maps; (iii) the proposed Attention-based Residual Block (ARB) contains a sequence of two PB blocks and one Convolutional Block Attention Module (CBAM), see Figure 5, and a skip connection to process the current patch.

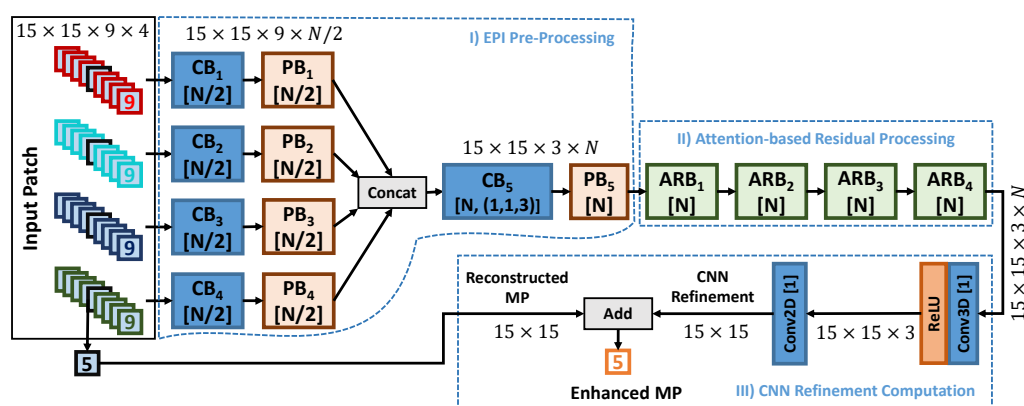


Figure 3. The proposed network architecture called Attention-aware EPI-based Quality Enhancement Convolutional Neural Network (AEQE-CNN).

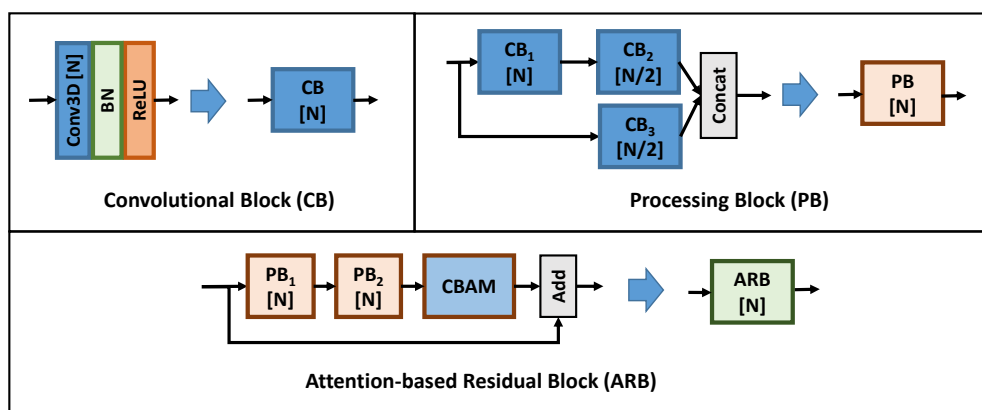


Figure 4. The layer structure of the three blocks used to build the proposed architecture: (top-left) Convolutional Block (CB); (top-right) Processing Block (PB); and (bottom) Attention-based Residual Block (ARB), where the Convolutional Block Attention Module (CBAM) was proposed [59] and modified here as depicted in Figure 5.

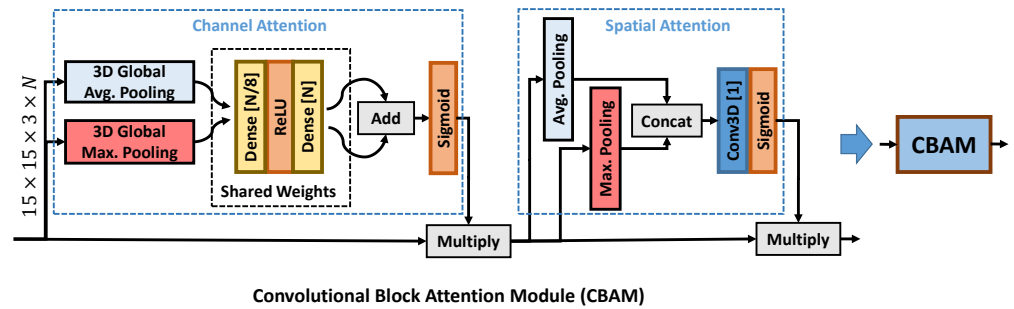


Figure 5. The layer structure of Convolutional Block Attention Module (CBAM), which uses both channel and spatial attention. The module was proposed in [59] and was modified here to compute the attention map for an MP volume.

Figure 3 shows that the AEQE-CNN architecture processes the EPI-based input patch using three stages. In the first stage, called EPI Pre-Processing, the MP volume corresponding to an EPI is processed using one CB block and one PB block, each equipped with $N/2$ filters, to extract the EPI feature maps, which are then concatenated and further processed by CB_5 and PB_5 , which are both equipped with N filters. CB_5 uses the stride $s = (1, 1, 3)$ to reduce the current patch resolution from $15 \times 15 \times 9$ to $15 \times 15 \times 3$ to decrease the inference time and to reduce the MP neighbourhood from 9 MPs to 3 MPs.

In the second stage, called Attention-based Residual Processing, a sequence of four APB blocks with N filters are used to further process the patch and extract the final feature maps of size $15 \times 15 \times N$. The final stage, called CNN Refinement Computation, is used to extract the final CNN-refinement using one Conv3D layer with ReLU activation and one Conv2D layer (equipped with a 3×3 kernel) with one filter. The CNN-refinement is then added to the currently reconstructed MP to obtain the enhanced MP.

In this paper, we propose to employ an attention-based module designed based on the CBAM module introduced in [59]. Figure 5 depicts the layer structure of CBAM. CBAM proposes the use of both channel attention and spatial attention. The channel attention uses the shared weights of two dense layers to process the two feature vectors extracted using global average pooling and global maximum pooling, respectively. The spatial attention uses a Conv3D layer to process the feature maps extracted using average pooling and maximum pooling. The two types of attention maps are obtained using a sigmoid activation layer and then applied in turn using a multiplication layer. The CBAM block was proposed in [59] for the processing of two-dimensional patches, while, here, the CBAM design was modified to be applied to MP volumes (three-dimensional patches).

3.3. Training Details

The AEQE-CNN models were trained using the Mean Squared Error (MSE) loss function equipped with an ℓ_2 regularization procedure to prevent model over-fitting. Let us denote: $\Theta_{\text{AEQE-CNN}}$ as the set of all learned parameters of the AEQE-CNN model; $\mathbf{X}^{(i)}$ as the i -th EPI-based input patch in the training set of size $15 \times 15 \times 9 \times 4$; and $\mathbf{Y}^{(i)}$ as the corresponding MP in the original LF image of size 15×15 . Let $F(\cdot)$ be the function that processes $\mathbf{X}^{(i)}$ using $\Theta_{\text{AEQE-CNN}}$ to compute the enhanced MP as $\hat{\mathbf{Y}}^{(i)} = F(\mathbf{X}^{(i)}, \Theta_{\text{AEQE-CNN}})$. The loss function is formulated as follows:

$$\mathcal{L}(\Theta_{\text{AEQE-CNN}}) = \frac{1}{L} \sum_{i=1}^L \|\text{vec}(\mathbf{Y}^{(i)}) - \text{vec}(\hat{\mathbf{Y}}^{(i)})\|_2^2 + \lambda \|\Theta_{\text{AEQE-CNN}}\|_2^2, \quad (3)$$

where L is the number of input patches, λ is the regularization term that is set empirically as $\lambda = 0.001$, and vec is the vectorization operator. Here, the Adam optimization algorithm [60] is employed.

By setting $N = 32$, the AEQE-CNN models contain 782,661 parameters that must be trained. Experiments using a more lightweight AEQE-CNN architecture were also performed, see Section 4.4. Version *HM 16.18* of the reference software implementation is used for the HEVC codec [16]. Note that other software implementations of HEVC, such as FFmpeg [61], Kvazaar [62], and OpenHEVC [63,64] are available; however, in this work, the reference software implementation of HEVC was used due to its high popularity within the research community. The proposed CNN-based filtering method trained four AEQE-CNN models, one for each of the four standard QP values, $QP = \{22, 27, 32, 37\}$.

The proposed neural network was implemented in the Python programming language using the Keras open-source deep-learning library, and was run on a machine equipped with Titan Xp Graphical Processing Units (GPUs).

In our previous work [33,34], the experimental results showed that an improved performance was obtained when HEVC was modified to skip its built-in in-loop filters, DBF [47] and SAO [48]. Therefore, here, four models were trained using EPI-based input patches extracted from reconstructed LF images obtained by running HEVC with its built-in in-loop filters, called AEQE-CNN + DBF&SAO, and four models were trained using EPI-based input patches extracted from reconstructed LF images obtained by running HEVC without its built-in in-loop filters, called AEQE-CNN. This training strategy demonstrates that the proposed CNN-based filtering method can be integrated into video coding systems where no modifications to the HEVC anchor are allowed.

The proposed AEQE-CNN architecture differs from our previous architecture design named MP-wise quality enhancement CNN (MPQE-CNN) [34] as follows. MPQE-CNN operates on MP volumes extracted from the closest 3×3 MP neighbourhood, while AEQE-CNN operates on EPI-based input patches extracted from an 9×9 MP neighbourhood. MPQE-CNN follows a multi-resolution design with simple CB blocks, while AEQE-CNN follows a design of multi-EPI branch processing and sequential residual block processing built based on more efficient PB blocks and novel attention-aware ARB blocks.

4. Experimental Validation

Section 4.1 describes the experimental setup used to compare the proposed CNN-based filtering method with the state-of-the-art methods. Section 4.2 illustrates the experimental results obtained over the test. Section 4.3 presents the visual results of the proposed CNN-based filtering method in comparison with the HEVC anchor. Finally, Section 4.4 presents an ablation study that analyses the possibility to reduce the network complexity and runtime using different approaches.

4.1. Experimental Setup

LF image Dataset. The experimental validation was carried out on the EPFL LF dataset [35], which contained 118 LF images in the RGB format, divided into 10 categories. Similar to [34], here, only the first 8 bits of the RGB color channels were encoded, and, similar to [29], 32 corner SAIs (8 from each corner) were dropped from the array of SAIs as they contained sparse information due to the shape of the microlens used by the plenoptic camera. Since the SAIs were color-transformed to the YUV format and only the Y channel was enhanced, the SAI video sequence contained 193 Y-frames. The closest frame resolution that HEVC [16] accepted as input was $W \times H = 632 \times 440$.

For a fair comparison with MPQE-CNN [34], the experiments were carried out on the same Training set (10 LF images) and Test set (108 LF images) as defined in [34], i.e., the Training set contained the following LF images: *Black_Fence*, *Chain_link_fence_1*, *ISO_chart_1*, *Houses_&_lake*, *Backlight_1*, *Broken_mirror*, *Bush*, *Fountain_&_Vincent_1*, *Ankylosaurus_&_Diplodocus_1*, and *Bench_in_Paris*. A total number of $625 \times 434 \times 10 = 2,712,500$ EPI-based input patches were collected from the 10 training images, and a 90%–10% ratio was used for splitting the training set into training–validation data. A batch size of 350 EPI-based input patches was used.

Comparison with the state-of-the-art methods. The two proposed methods, AEQE-CNN + DBF&SAO and AEQE-CNN, were compared with (i) the HEVC [16] anchor, denoted by HEVC + DBF&SAO; (ii) the FQE-CNN architecture from [33] where each SAI in the LF image was enhanced in turn; and (iii) the MPQE-CNN architecture from [34] based on a similar MP-wise filtering approach. The distortion was measured using the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) [65]. The standard Bjøntegaard delta bitrate (BD-rate) savings and Bjøntegaard delta PSNR (BD-PSNR) improvement [66] were computed using the four standard QP values: $QP = \{22, 27, 32, 37\}$.

4.2. Experimental Results

Figure 6 shows the compression results over the test set (108 LF images) for the rate-distortion curves computed as Y-PSNR-vs.-bitrate and SSIM-vs.-bitrate. Figure 7 shows the Y-BD-PSNR and Y-BD-rate values computed for each LF image in the test set. The proposed methods provide an improved performance compared with HEVC [16] + DBF&SAO, FQE-CNN [33], and MPQE-CNN [34] at both low and high bitrates. The results show that AEQE-CNN provided a small improvement over AEQE-CNN + DBF&SAO. The proposed CNN-based filtering method was able to provide a large improvement even when no modification was applied to the HEVC video codec.

Table 1 shows the average results obtained over the test set. AEQE-CNN provided Y-BD-rate savings of 36.57% and Y-BD-PSNR improvements of 2.301 dB over HEVC [16], i.e., a more than 40% improvement was achieved compared with MPQE-CNN [33].

Table 1. Average results obtained over the test set.

Method	Bjøntegaard Metric	
	Y-BD-PSNR (dB)	Y-BD-Rate (%)
FQE-CNN [33]	0.4515	−9.1921
MPQE-CNN [34]	1.5478	−25.5285
AEQE-CNN + DBF&SAO	2.2044	−35.3142
AEQE-CNN	2.3006	−36.5713

Figure 8 shows the Rate-Distortion (RD) results for three randomly selected LF images in the test set, *Chain_link_fence_2*, *Flowers*, and *Palais_du_Luxembourg*. AEQE-CNN provided an Y-BD-PSNR improvement of around 2 dB at both low and high bitrates. The SSIM-vs.-bitrate results show that the visual quality at low bitrates was highly improved of around 0.08.

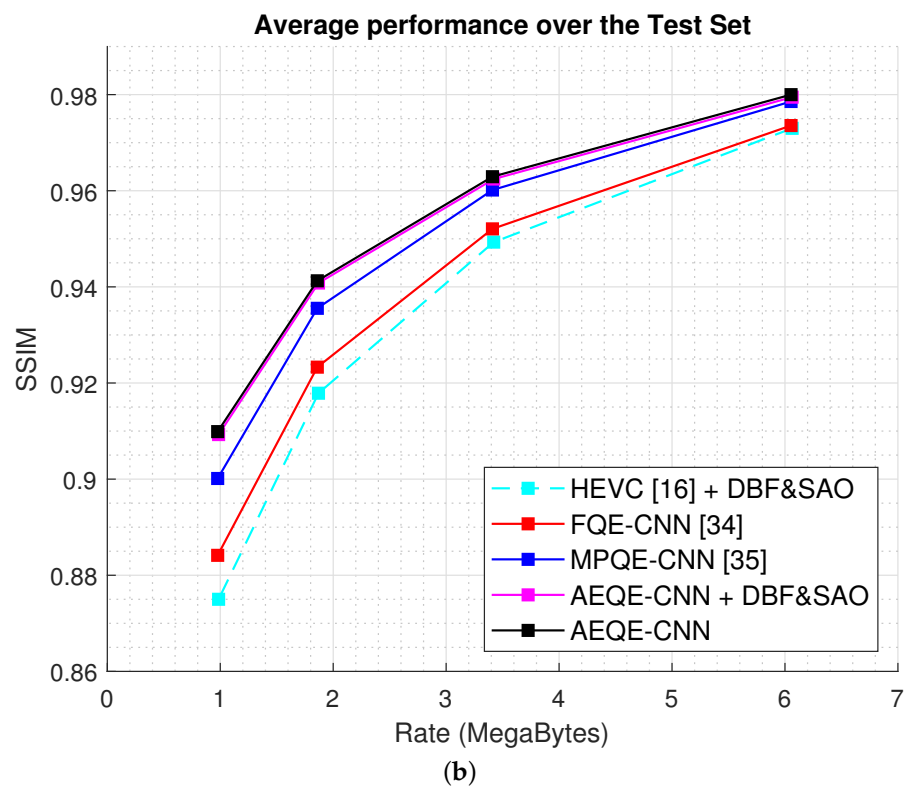
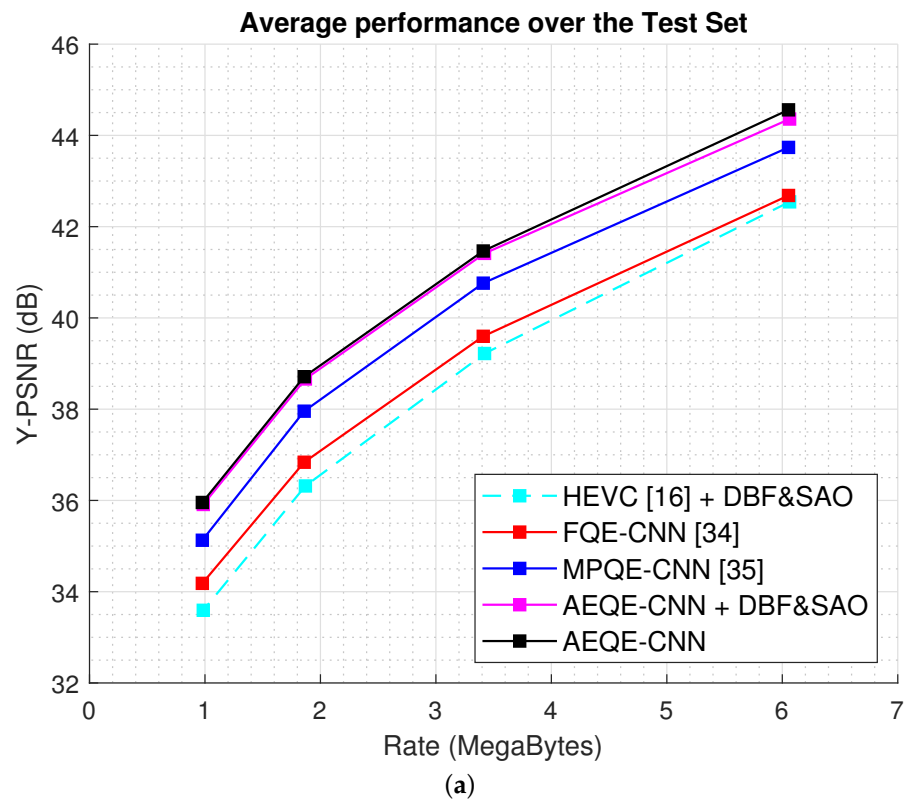
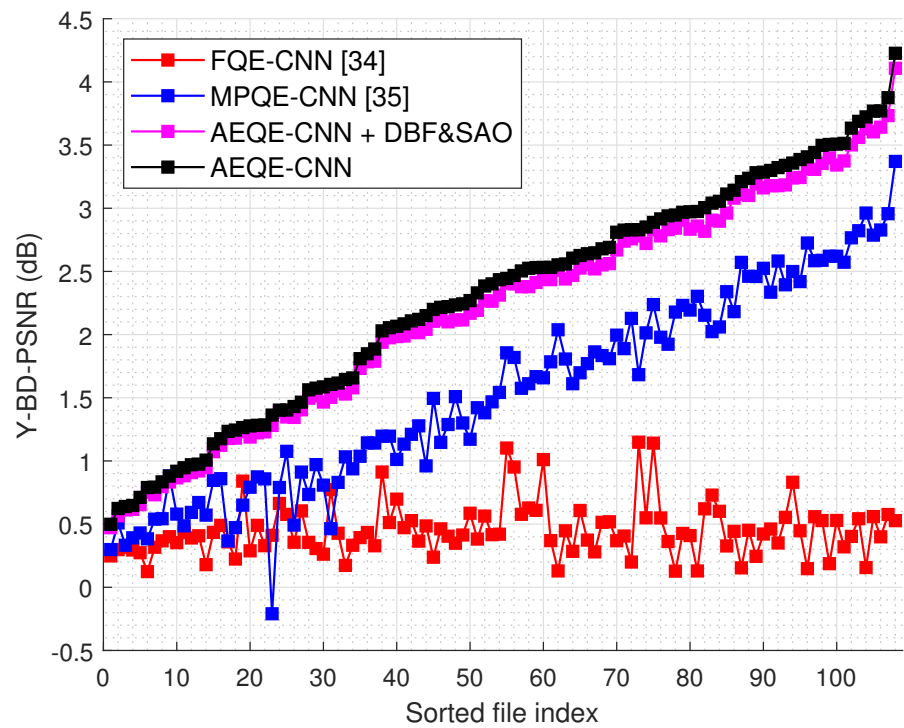
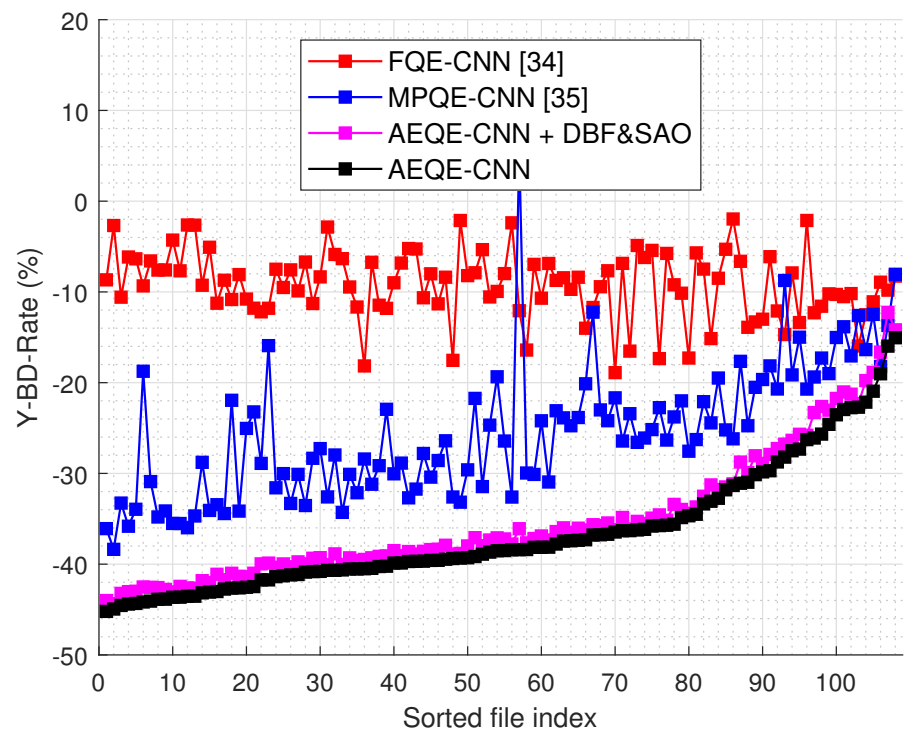


Figure 6. The Rate-Distortion results over the test set. (a) Y-PSNR-vs.-bitrate. (b) SSIM-vs.-bitrate.



(a)



(b)

Figure 7. The Bjøntegaard metric results for every LF image in the test set: (a) Y-BD-PSNR gains (dB); (b) Y-BD-rate savings (%).

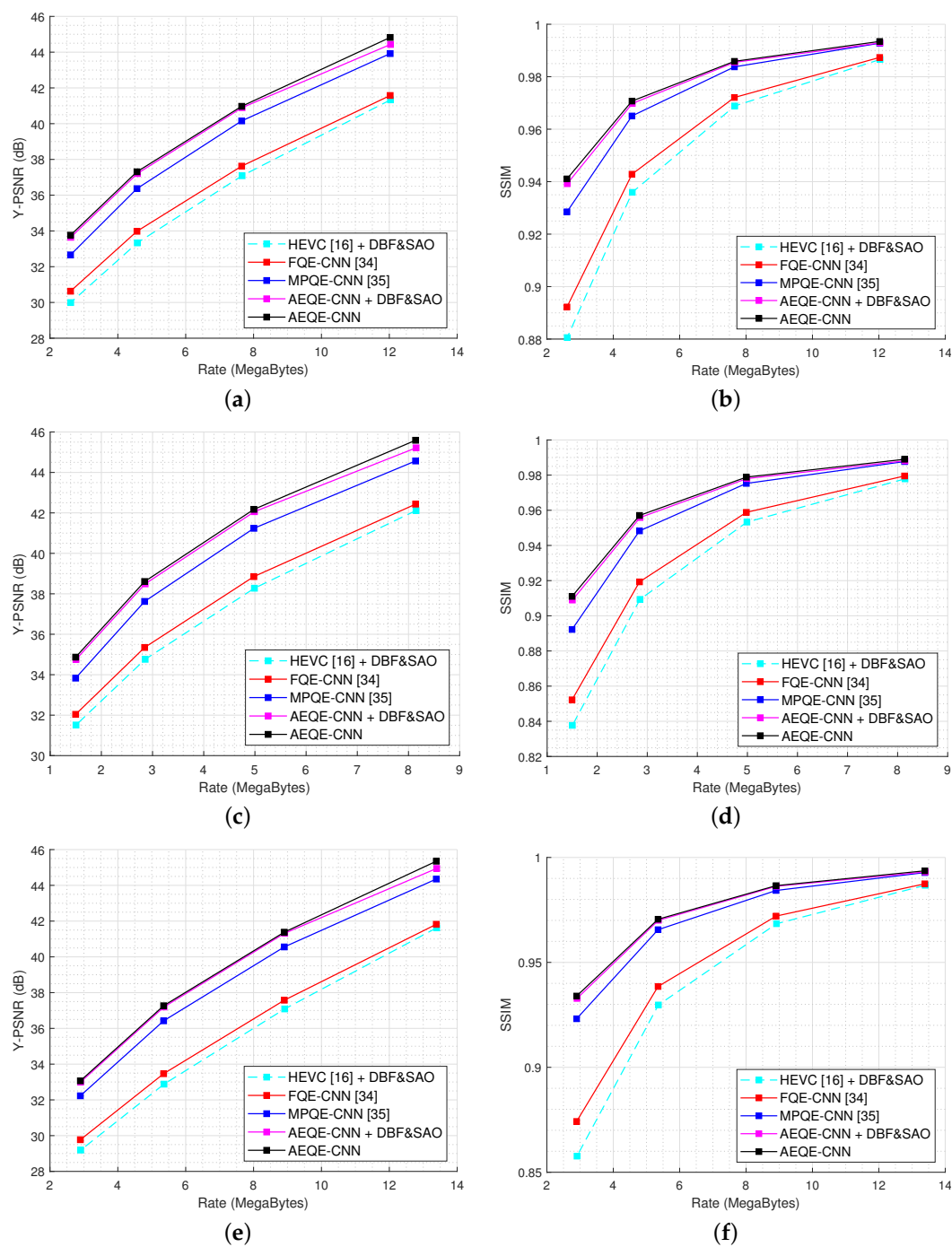
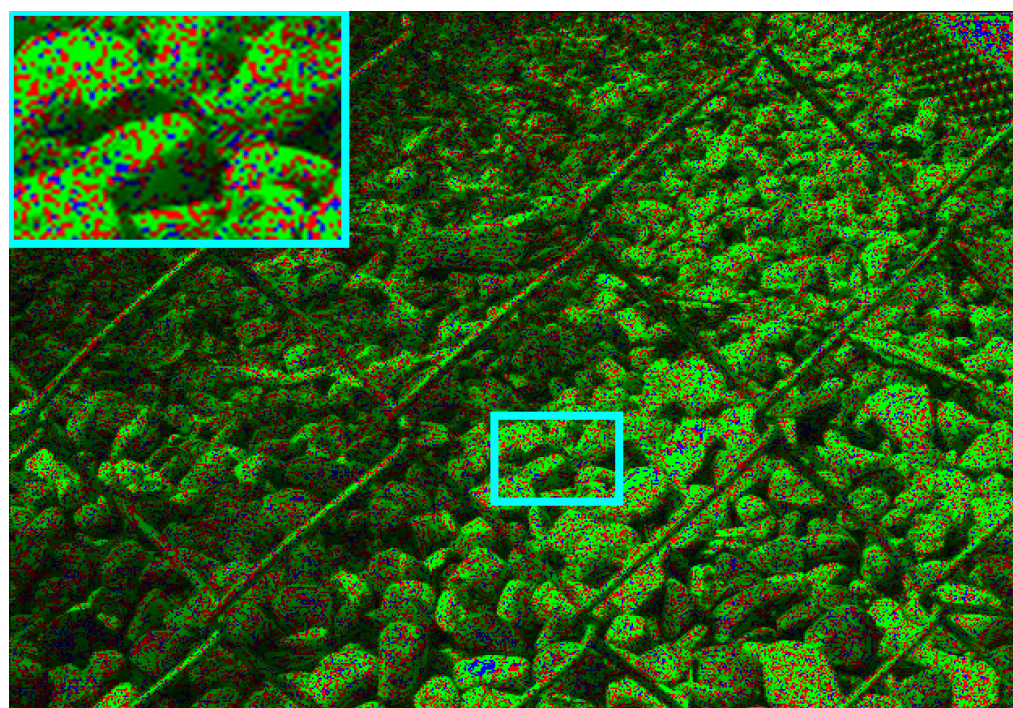


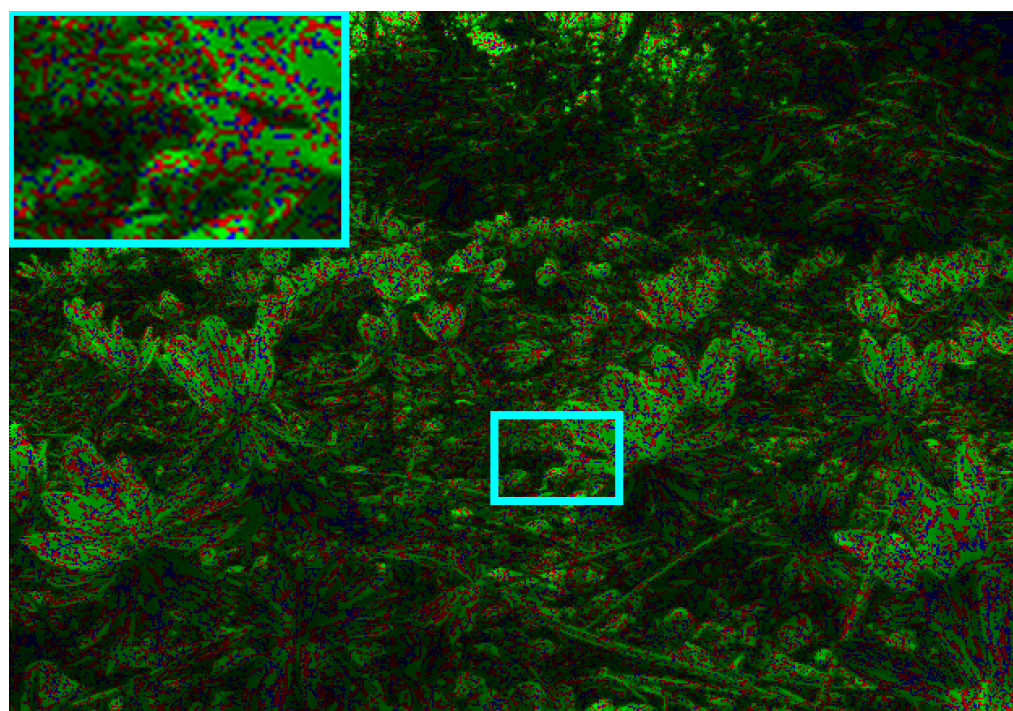
Figure 8. The Rate-Distortion results for three LF images in the test set. (a) Y-PSNR-vs.-bitrate for *Chain_link_fence_2*; (b) SSIM-vs.-bitrate for *Chain_link_fence_2*; (c) Y-PSNR-vs.-bitrate for *Flowers*; (d) SSIM-vs.-bitrate for *Flowers*; (e) Y-PSNR-vs.-bitrate for *Palais_du_Luxembourg*; (f) SSIM-vs.-bitrate for *Palais_du_Luxembourg*.

4.3. Visual Results

Figure 9 shows the pseudo-coloured image comparison between AEQE-CNN and HEVC [16] + DBF&SAO for two LF images in the test set, *Chain_link_fence_2* and *Flowers*. The green, blue, and red pixels mark the positions where AEQE-CNN provided an improved, similar, and worse performance, respectively, compared with HEVC [16] + DBF&SAO anchor. Green is the dominant color, which shows that AEQE-CNN enhanced the quality of almost all pixels in the LF image.



(a)



(b)

Figure 9. Pseudo-coloured image comparison between AEQE-CNN and HEVC [16] + DBF&SAO based on the absolute reconstruction error for the center SAI at position $(p, q) = (8, 8)$, and for $QP = 37$. Green marks the pixel positions where AEQE-CNN achieved better performance. Blue marks the pixel positions where the two methods had the same performance. Red marks pixels where HEVC [16] + DBF&SAO achieved better performance. The cyan rectangle marks an image area shown zoomed-in at the top-left corner and the corresponding Y channel in Figure 10. The results for two LF images in the test set: (a) *Chain_link_fence_2*; (b) *Flowers*.

Figure 10 shows the visual result comparison between AEQE-CNN and HEVC [16] + DBF&SAO for the corresponding Y channel of the two zoomed-in image areas marked by cyan rectangles in Figure 9. AEQE-CNN provided much sharper image edges and added more details to the image textures.

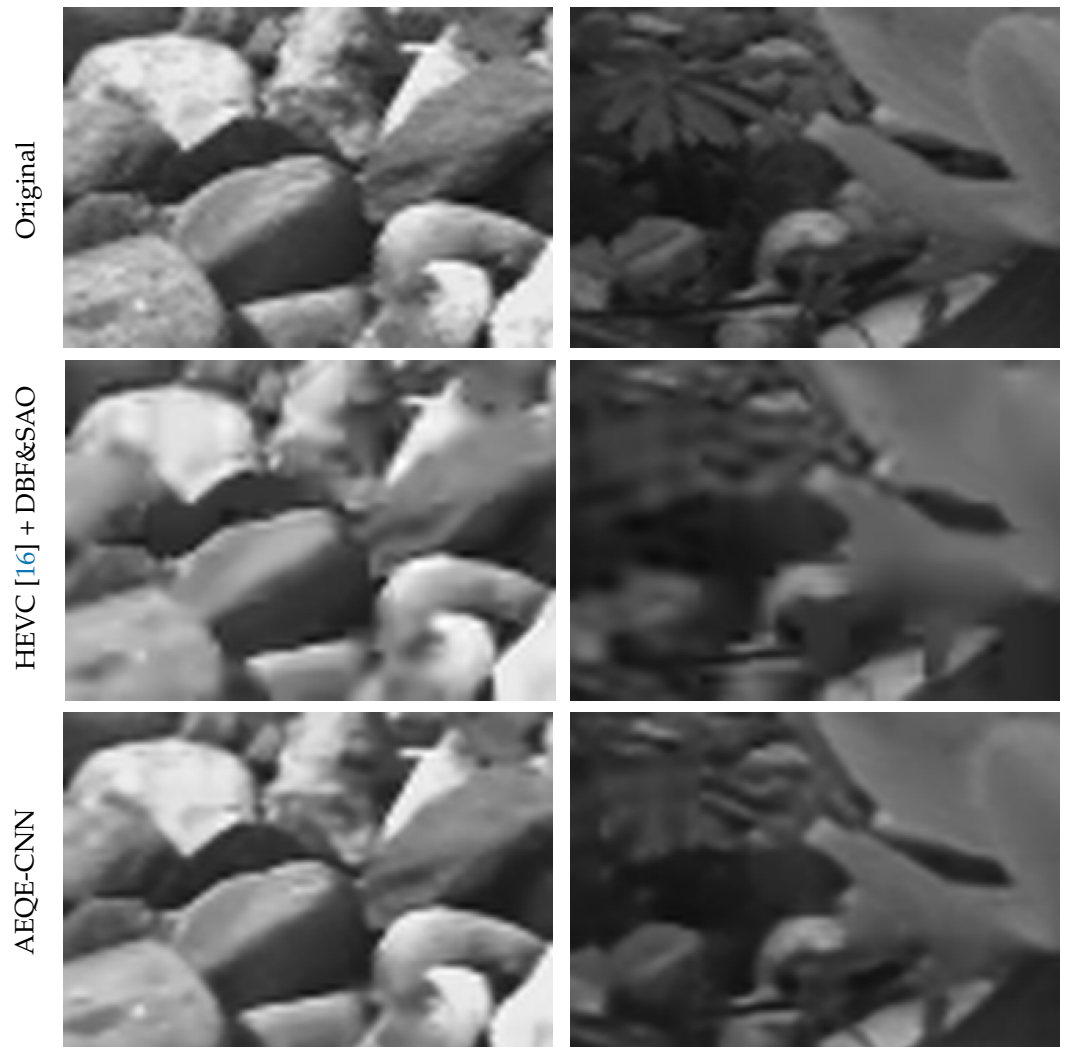


Figure 10. Visual comparison between AEQE-CNN and HEVC [16] + DBF&SAO for the Y channel of the zoomed-in image area marked by the cyan rectangle in Figure 9 above.

4.4. Ablation Study

In this work, we also studied the possibility to reduce the network complexity and runtime using two different approaches. In the first approach, an architecture variation of AEQE-CNN was generated by halving the number of channels used throughout the architecture by the 3D Convolution layers from $N = 32$ to $N = 16$. This first AEQE-CNN architecture variation is called AEQE-CNN [N=16]. In the second approach, the size of the MP neighbourhood, $\mathcal{N}_{x,y}$ (see Section 3.1), was reduced from 9×9 MPs (i.e., $b = 4$) to 3×3 MPs (i.e., $b = 1$).

More precisely, the same neighbourhood window as in [34] was used here with the goal of evaluating the influence of the size of the MP neighbourhood in the final enhancement results. In this case, the EPI volumes were of the size $15 \times 15 \times 3$; therefore, the CB_5 block in the AEQE-CNN architecture (see Figure 3) used a default stride of $s' = (1, 1, 1)$ instead of $s = (1, 1, 3)$. This second AEQE-CNN architecture variation is called AEQE-CNN [3×3].

Table 2 shows the average results obtained over the test set for the three AEQE-CNN architectures. The AEQE-CNN provided the best performance using the highest complexity

and runtime. The network variations corresponding to the two approaches for complexity reduction still provided a better performance compared with the state-of-the-art methods and a close performance to AEQE-CNN. AEQE-CNN [N=16] offered a reduction of 44.6% in the inference runtime and a reduction of 74.7% in the network complexity, with a drop in the average performance of only 8.93% in Y-BD-PSNR and 3.59% in Y-BD-Rate.

Table 2. The average results obtained over the test set for the three AEQE-CNN network variations.

Method	Bjontegaard Metric		Nr. of Trained Parameters	Inference Time Per Img.
	Y-BD-PSNR	Y-BD-Rate		
AEQE-CNN [N=16]	2.0954 dB	−35.2581%	197,661 (−74.7%)	98 s (−44.6%)
AEQE-CNN [3×3]	2.0799 dB	−35.0914%	782,661	105 s (−40.7%)
AEQE-CNN	2.3006 dB	−36.5713%	782,661	177 s

AEQE-CNN [3×3] offered a reduction of 40.7% in the inference runtime, with a drop in the average performance of only 9.6% in Y-BD-PSNR and of 4.05% in Y-BD-Rate. The ablation study demonstrate that AEQE-CNN [3×3] provided a large reduction in the network complexity and inference runtime while accepting a small performance drop compared with AEQE-CNN.

Figure 11 shows the rate-distortion curves computed over the test set for AEQE-CNN [N=16], AEQE-CNN [3×3], and AEQE-CNN. The results demonstrate again that the two network variations provided a close performance to AEQE-CNN. The performance dropped with less than 0.2 dB at low and high bitrates for the two architecture variations. The results obtained by AEQE-CNN [3×3] demonstrate that the proposed AEQE-CNN architecture, built using the PB and ARB blocks, provided an improved performance compared with the MPQE-CNN architecture [34] when operating on the same MP neighbourhood.

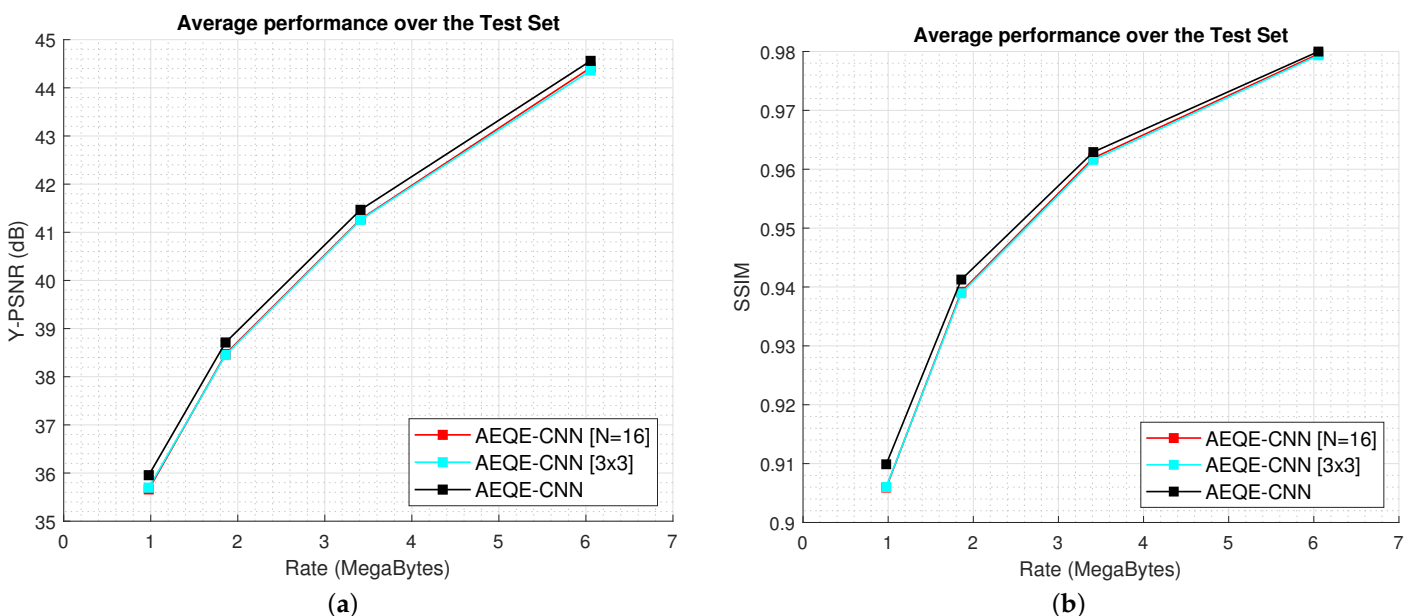


Figure 11. The Rate-Distortion results over the test set for the three network variations. (a) Y-PSNR-vs.-bitrate. (b) SSIM-vs.-bitrate.

Figure 12 shows the results of the Bjontegaard metrics, Y-BD-PSNR and Y-BD-rate, computed for each LF image in the test set. The results demonstrate again that the two network variations provided a close performance to AEQE-CNN.

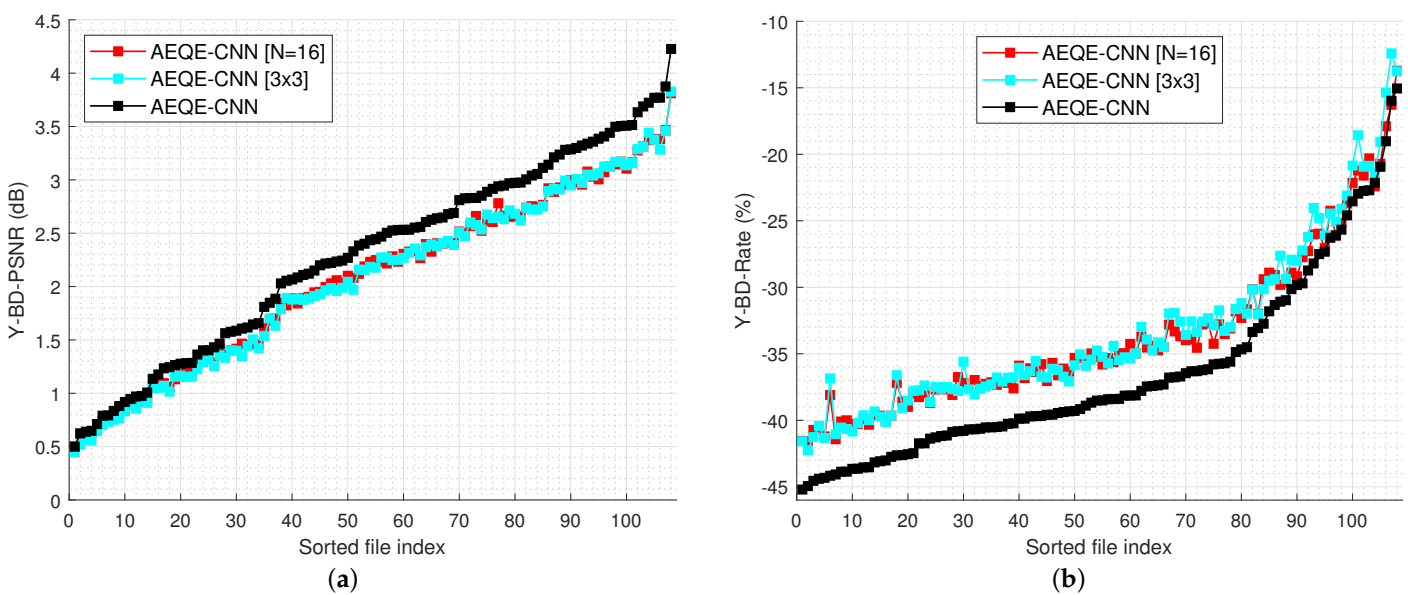


Figure 12. Bjøntegaard metrics results for every LF image in test set for the three network variations: (a) Y-BD-PSNR gains; (b) Y-BD-Rate savings.

5. Conclusions

In this paper, we proposed a novel CNN-based filtering method for the quality enhancement of LF images compressed by HEVC. The proposed architecture, AEQE-CNN, was built using novel layer structure blocks, such as complex processing blocks and attention-based residual blocks. AEQE-CNN operated on an EPI-based input patch extracted from an MP neighbourhood of 9×9 MPs and followed an MP-wise filtering approach that was specific to LF images. Similar to previous research works, the proposed AEQE-CNN filtering method provided an increased performance when the conventional HEVC built-in filtering methods were skipped. The results demonstrate the high potential of attention networks for the quality enhancement of LF images.

In our future work, we plan to study different strategies to reduce the inference runtime using lightweight neural network architectures, and to employ the CNN-based filtering method to enhance the quality of the light field images compressed using other video codecs, such as AV1 and VVC.

Author Contributions: Conceptualization, I.S.; methodology, I.S.; software, I.S.; validation, I.S.; investigation, I.S.; resources, A.M.; writing—original draft preparation, I.S. and A.M.; writing—review and editing, I.S. and A.M.; visualization, I.S.; project administration, I.S. and A.M.; funding acquisition, I.S. and A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was funded by Innoviris within the research project DRIViNg, and by Ionut Schiopu's personal fund.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results. All authors read and approved the final manuscript.

References

1. Jeon, H.G.; Park, J.; Choe, G.; Park, J.; Bok, Y.; Tai, Y.W.; Kweon, I.S. Accurate depth map estimation from a lenslet light field camera. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1547–1555. [\[CrossRef\]](#)
2. Wang, T.C.; Efros, A.A.; Ramamoorthi, R. Depth Estimation with Occlusion Modeling Using Light-Field Cameras. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2170–2181. [\[CrossRef\]](#)
3. Schiopu, I.; Munteanu, A. Deep-learning-based depth estimation from light field images. *Electron. Lett.* **2019**, *55*, 1086–1088. [\[CrossRef\]](#)

4. Rogge, S.; Schiopu, I.; Munteanu, A. Depth Estimation for Light-Field Images Using Stereo Matching and Convolutional Neural Networks. *Sensors* **2020**, *20*, 6188. [[CrossRef](#)]
5. Flynn, J.; Broxton, M.; Debevec, P.; DuVall, M.; Fyffe, G.; Overbeck, R.; Snavely, N.; Tucker, R. DeepView: View Synthesis With Learned Gradient Descent. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2362–2371. [[CrossRef](#)]
6. Peng, J.; Xiong, Z.; Zhang, Y.; Liu, D.; Wu, F. LF-fusion: Dense and accurate 3D reconstruction from light field images. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4. [[CrossRef](#)]
7. Chen, M.; Tang, Y.; Zou, X.; Huang, K.; Li, L.; He, Y. High-accuracy multi-camera reconstruction enhanced by adaptive point cloud correction algorithm. *Opt. Lasers Eng.* **2019**, *122*, 170–183. [[CrossRef](#)]
8. Forman, M.C.; Aggoun, A.; McCormick, M. A novel coding scheme for full parallax 3D-TV pictures. In Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, 21–24 April 1997; Volume 4, pp. 2945–2947. [[CrossRef](#)]
9. de Carvalho, M.B.; Pereira, M.P.; Alves, G.; da Silva, E.A.B.; Pagliari, C.L.; Pereira, F.; Testoni, V. A 4D DCT-Based Lenslet Light Field Codec. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 435–439. [[CrossRef](#)]
10. Chang, C.-L.; Zhu, X.; Ramanathan, P.; Girod, B. Light field compression using disparity-compensated lifting and shape adaptation. *IEEE Trans. Image Process.* **2006**, *15*, 793–806. [[CrossRef](#)]
11. Rüfenacht, D.; Naman, A.T.; Mathew, R.; Taubman, D. Base-Anchored Model for Highly Scalable and Accessible Compression of Multiview Imagery. *IEEE Trans. Image Process.* **2019**, *28*, 3205–3218. [[CrossRef](#)]
12. Jang, J.S.; Yeom, S.; Javidi, B. Compression of ray information in three-dimensional integral imaging. *Opt. Eng.* **2005**, *44*, 1–10. [[CrossRef](#)]
13. Kang, H.H.; Shin, D.H.; Kim, E.S. Compression scheme of sub-images using Karhunen-Loeve transform in three-dimensional integral imaging. *Opt. Commun.* **2008**, *281*, 3640–3647. [[CrossRef](#)]
14. Elias, V.; Martins, W. On the Use of Graph Fourier Transform for Light-Field Compression. *J. Commun. Inf. Syst.* **2018**, *33*. [[CrossRef](#)]
15. Hog, M.; Sabater, N.; Guillemot, C. Superrays for Efficient Light Field Processing. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1187–1199. [[CrossRef](#)]
16. Sullivan, G.; Ohm, J.; Han, W.; Wiegand, T. Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *22*, 1649–1668. [[CrossRef](#)]
17. Ramanathan, P.; Flierl, M.; Girod, B. Multi-hypothesis prediction for disparity compensated light field compression. In Proceedings of the 2001 International Conference on Image Processing (Cat. No.01CH37205), Thessaloniki, Greece, 7–10 October 2001; Volume 2, pp. 101–104. [[CrossRef](#)]
18. Wang, G.; Xiang, W.; Pickering, M.; Chen, C.W. Light Field Multi-View Video Coding With Two-Directional Parallel Inter-View Prediction. *IEEE Trans. Image Process.* **2016**, *25*, 5104–5117. [[CrossRef](#)] [[PubMed](#)]
19. Conti, C.; Nunes, P.; Soares, L.D. New HEVC prediction modes for 3D holoscopic video coding. In Proceedings of the 2012 19th IEEE International Conference on Image Processing, Orlando, FL, USA, 30 September–3 October 2012; pp. 1325–1328. [[CrossRef](#)]
20. Zhong, R.; Schiopu, I.; Cornelis, B.; Lu, S.P.; Yuan, J.; Munteanu, A. Dictionary Learning-Based, Directional, and Optimized Prediction for Lenslet Image Coding. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 1116–1129. [[CrossRef](#)]
21. Dricot, A.; Jung, J.; Cagnazzo, M.; Pesquet, B.; Dufaux, F. Improved integral images compression based on multi-view extraction. In *Applications of Digital Image Processing XXXIX*; Tescher, A.G., Ed.; International Society for Optics and Photonics, SPIE: San Diego, CA, USA, 2016; Volume 9971, pp. 170–177. [[CrossRef](#)]
22. Astola, P.; Tabus, I. Coding of Light Fields Using Disparity-Based Sparse Prediction. *IEEE Access* **2019**, *7*, 176820–176837. [[CrossRef](#)]
23. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2016**, arXiv:1409.0473.
24. Zhu, M.; Jiao, L.; Liu, F.; Yang, S.; Wang, J. Residual Spectral–Spatial Attention Network for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote. Sens.* **2021**, *59*, 449–462. [[CrossRef](#)]
25. Wan, S.; Tang, S.; Xie, X.; Gu, J.; Huang, R.; Ma, B.; Luo, L. Deep Convolutional-Neural-Network-based Channel Attention for Single Image Dynamic Scene Blind Deblurring. *IEEE Trans. Circuits Syst. Video Technol.* **2020**. [[CrossRef](#)]
26. Fu, C.; Yin, Y. Edge-Enhanced with Feedback Attention Network for Image Super-Resolution. *Sensors* **2021**, *21*, 2064. [[CrossRef](#)]
27. Zhou, K.; Zhan, Y.; Fu, D. Learning Region-Based Attention Network for Traffic Sign Recognition. *Sensors* **2021**, *21*, 686. [[CrossRef](#)]
28. Lian, J.; Yin, Y.; Li, L.; Wang, Z.; Zhou, Y. Small Object Detection in Traffic Scenes Based on Attention Feature Fusion. *Sensors* **2021**, *21*, 3031. [[CrossRef](#)]
29. Schiopu, I.; Gabbouj, M.; Gotchev, A.; Hannuksela, M.M. Lossless compression of subaperture images using context modeling. In Proceedings of the 2017 3DTV Conf.: The True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON), Copenhagen, Denmark, 7–9 June 2017; pp. 1–4. [[CrossRef](#)]

30. Schiopu, I.; Munteanu, A. Macro-Pixel Prediction Based on Convolutional Neural Networks for Lossless Compression of Light Field Images. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athene, Greece, 7–10 October 2018; pp. 445–449. [\[CrossRef\]](#)
31. Schiopu, I.; Munteanu, A. Deep-learning-based macro-pixel synthesis and lossless coding of light field images. *Apsipa Trans. Signal Inf. Process.* **2019**, *8*, e20. [\[CrossRef\]](#)
32. Schiopu, I.; Munteanu, A. Deep-Learning-Based Lossless Image Coding. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 1829–1842. [\[CrossRef\]](#)
33. Huang, H.; Schiopu, I.; Munteanu, A. Frame-wise CNN-based Filtering for Intra-Frame Quality Enhancement of HEVC Videos. *IEEE Trans. Circuits Syst. Video Technol.* **2020**. [\[CrossRef\]](#)
34. Huang, H.; Schiopu, I.; Munteanu, A. Macro-pixel-wise CNN-based filtering for quality enhancement of light field images. *Electron. Lett.* **2020**, *56*, 1413–1416. [\[CrossRef\]](#)
35. Rerabek, M.; Ebrahimi, T. New Light Field Image Dataset. Proc. Int. Conf. Qual. Multimedia Experience (QoMEX). 2016; pp. 1–2. Available online: https://infoscience.epfl.ch/record/218363/files/Qomex2016_shortpaper.pdf?version=1 (accessed on 1 July 2017).
36. Dong, C.; Deng, Y.; Loy, C.C.; Tang, X. Compression Artifacts Reduction by a Deep Convolutional Network. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 576–584. [\[CrossRef\]](#)
37. Cavigelli, L.; Hager, P.; Benini, L. CAS-CNN: A deep convolutional neural network for image compression artifact suppression. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 752–759. [\[CrossRef\]](#)
38. Wang, Z.; Liu, D.; Chang, S.; Ling, Q.; Yang, Y.; Huang, T.S. D3: Deep Dual-Domain Based Fast Restoration of JPEG-Compressed Images. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2764–2772. [\[CrossRef\]](#)
39. Galteri, L.; Seidenari, L.; Bertini, M.; Bimbo, A.D. Deep Generative Adversarial Compression Artifact Removal. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4836–4845. [\[CrossRef\]](#)
40. Ororbia, A.G.; Mali, A.; Wu, J.; O’Connell, S.; Dreese, W.; Miller, D.; Giles, C.L. Learned Neural Iterative Decoding for Lossy Image Compression Systems. In Proceedings of the 2019 Data Compression Conference (DCC), Snowbird, UT, USA, 26–29 March 2019; pp. 3–12. [\[CrossRef\]](#)
41. Dai, Y.; Liu, D.; Wu, F. A Convolutional Neural Network Approach for Post-Processing in HEVC Intra Coding. *Lect. Notes Comput. Sci.* **2016**, 28–39. [\[CrossRef\]](#)
42. Yang, R.; Xu, M.; Wang, Z.; Li, T. Multi-frame Quality Enhancement for Compressed Video. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. [\[CrossRef\]](#)
43. He, X.; Hu, Q.; Zhang, X.; Zhang, C.; Lin, W.; Han, X. Enhancing HEVC Compressed Videos with a Partition-Masked Convolutional Neural Network. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athene, Greece, 7–10 October 2018; pp. 216–220. [\[CrossRef\]](#)
44. Ma, C.; Liu, D.; Peng, X.; Wu, F. Convolutional Neural Network-Based Arithmetic Coding of DC Coefficients for HEVC Intra Coding. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athene, Greece, 7–10 October 2018; pp. 1772–1776. [\[CrossRef\]](#)
45. Song, X.; Yao, J.; Zhou, L.; Wang, L.; Wu, X.; Xie, D.; Pu, S. A Practical Convolutional Neural Network as Loop Filter for Intra Frame. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athene, Greece, 7–10 October 2018; pp. 1133–1137. [\[CrossRef\]](#)
46. Wan, S. CE13-Related: Integrated in-Loop Filter Based on CNN. JVET Document, JVET-N0133-v2, 2019. Available online: https://www.itu.int/wftp3/av-arch/jvet-site/2019_03_N_Geneva/JVET-N_Notes_d2.docx (accessed on 1 July 2020).
47. Norkin, A.; Bjontegaard, G.; Fuldseth, A.; Narroschke, M.; Ikeda, M.; Andersson, K.; Zhou, M.; Van der Auwera, G. HEVC Deblocking Filter. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *22*, 1746–1754. [\[CrossRef\]](#)
48. Fu, C.; Alshina, E.; Alshin, A.; Huang, Y.; Chen, C.; Tsai, C.; Hsu, C.; Lei, S.; Park, J.; Han, W. Sample Adaptive Offset in the HEVC Standard. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *22*, 1755–1764. [\[CrossRef\]](#)
49. Park, W.; Kim, M. CNN-based in-loop filtering for coding efficiency improvement. In Proceedings of the 2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), Bordeaux, France, 11–12 July 2016; pp. 1–5. [\[CrossRef\]](#)
50. Zhang, Z.; Chen, Z.; Lin, J.; Li, W. Learned Scalable Image Compression with Bidirectional Context Disentanglement Network. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 1438–1443. [\[CrossRef\]](#)
51. Li, F.; Tan, W.; Yan, B. Deep Residual Network for Enhancing Quality of the Decoded Intra Frames of Hvc. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athene, Greece, 7–10 October 2018; pp. 3918–3922. [\[CrossRef\]](#)
52. Lai, P.; Wang, J. Multi-stage Attention Convolutional Neural Networks for HEVC In-Loop Filtering. In Proceedings of the 2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), Genova, Italy, 31 August–2 September 2020; pp. 173–177. [\[CrossRef\]](#)
53. Zhang, X.; Xiong, R.; Lin, W.; Zhang, J.; Wang, S.; Ma, S.; Gao, W. Low-Rank-Based Nonlocal Adaptive Loop Filter for High-Efficiency Video Compression. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 2177–2188. [\[CrossRef\]](#)

54. Zhang, Y.; Shen, T.; Ji, X.; Zhang, Y.; Xiong, R.; Dai, Q. Residual Highway Convolutional Neural Networks for in-loop Filtering in HEVC. *IEEE Trans. Image Process.* **2018**, *27*, 3827–3841. [[CrossRef](#)] [[PubMed](#)]
55. Jia, C.; Wang, S.; Zhang, X.; Wang, S.; Liu, J.; Pu, S.; Ma, S. Content-Aware Convolutional Neural Network for In-Loop Filtering in High Efficiency Video Coding. *IEEE Trans. Image Process.* **2019**, *28*, 3343–3356. [[CrossRef](#)]
56. Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute (HHI). HEVC Reference Software. Available online: hevc.hhi.fraunhofer.de (accessed on 1 July 2019)
57. Bossen, F. Common HM Test Conditions and Software Reference Configurations. JCT-VC Document, JCTVC-G1100, 2012. Available online: https://www.itu.int/wftp3/av-arch/jctvc-site/2012_02_H_SanJose/JCTVC-H_Notes_dI.doc (accessed on 1 July 2017).
58. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**, arxiv:1502.03167.
59. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
60. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arxiv:1412.6980.
61. FFmpeg. Libx265 Implementation of HEVC. Available online: <http://ffmpeg.org> (accessed on 1 April 2021).
62. Viitanen, M.; Koivula, A.; Lemmetti, A.; Ylä-Outinen, A.; Vanne, J.; Hämäläinen, T.D. Kvazaar: Open-Source HEVC/H.265 Encoder. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 1179–1182. [[CrossRef](#)]
63. Hamidouche, W.; Raulet, M.; Déforges, O. 4K Real-Time and Parallel Software Video Decoder for Multilayer HEVC Extensions. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *26*, 169–180. [[CrossRef](#)]
64. Pescador, F.; Chavarrías, M.; Garrido, M.; Malagón, J.; Sanz, C. Real-time HEVC decoding with OpenHEVC and OpenMP. In Proceedings of the 2017 IEEE International Conference on Consumer Electronics (ICCE), Berlin, Germany, 3–6 September 2017; pp. 370–371. [[CrossRef](#)]
65. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]
66. Bjøntegaard, G. Calculation of average PSNR differences between RD-curves. In Proceedings of the ITU-T Video Coding Experts Group (VCEG) 13th Meeting, Austin, TX, USA, 2–4 April 2001; pp. 2–4.