Light Field Image Coding Using VVC standard and View Synthesis based on Dual Discriminator GAN

Nader Bakir, Wassim Hamidouche, Member, IEEE, Sid Ahmed Fezza, Khouloud Samrouth, and Olivier Déforges

Abstract—Light field (LF) technology is considered as a promising way for providing a high-quality virtual reality (VR) content. However, such an imaging technology produces a large amount of data requiring efficient LF image compression solutions. In this paper, we propose a LF image coding method based on a view synthesis and view quality enhancement techniques. Instead of transmitting all the LF views, only a sparse set of reference views are encoded and transmitted, while the remaining views are synthesized at the decoder side. The transmitted views are encoded using the versatile video coding (VVC) standard and are used as reference views to synthesize the dropped views. The selection of non-reference dropped views is performed using a rate-distortion optimization based on the VVC temporal scalability. The dropped views are reconstructed using the LF dual discriminator GAN (LF-D2GAN) model. In addition, to ensure that the quality of the views is consistent, at the decoder, a quality enhancement procedure is performed on the reconstructed views allowing smooth navigation across views. Experimental results show that the proposed method provides the reconstructed views allowing smooth navigation across views.

I. INTRODUCTION

The idea of the light flows through environment interpreted as a field was first established by Michael Faraday in 1846. The mathematical formalisation was proposed 28 years later by James Clerk Maxwell with his famous equations. The concept of light field (LF) was then first defined in Arun Gershun’s paper [1] as the amount of light traveling in every direction through every point in 3D space. This amount of light is radiance, denoted by \( L \), is measured in watts per steradian per meter squared. The plenoptic function gives the radiance along all such arrays in a scene of 3D space with constant illumination

\[
P(x, y, z, \theta, \phi, \lambda, t),
\]

the rays in space are parameterized by 3D coordinates \((x, y, z)\), two angles \((\theta, \phi)\), wavelength \(\lambda\) and time \(t\). This 7 dimensional (7D) plenoptic function can be simplified into a 5D function where the time is sampled to the device frame rate and the wavelength is composed of 3 Red-Green-Blue (RGB) components. Assuming that the air around the object does not reflect or absorb the light and all ray intensities remain constant along their path, each ray is described by its intersection with two parallel planes denoted in this paper by \((u, v)\) and \((x, y)\) as illustrated in Fig. 1 for angular and spacial coordinates, respectively. This 4D [2] Light Field function \( L(x, y, u, v) \) can be represented as a collection of perspective images of the \((x, y)\) plane viewed from a position on the \((u, v)\) plane.

The LF acquisition is performed by sampling both spatial and angular dimensions. The acquisition devices fall into two main categories depending on whether camera arrays or plenoptic camera acquisition technology is used. The camera arrays are matrices of synchronized cameras arranged in a plane often at regular interval, where each camera represents an angular sample and each image to spatial samples. Plenoptic camera relies on microlenses to capture lights coming from different directions. The spatial resolution is determined by the number of microlenses, while the angular resolution depends on the number of pixels behind each microlens. The resulting LF image from plenoptic camera is then a collection of microlens images. This latter representation, called micro-image (MI), can be de-multiplexed in order to obtain subaperture-images (SAIs), where each SAI gathers pixels with the same relative position in the microlens image. The baseline of the LF image captured by plenoptic camera is smaller compared to the one captured by camera arrays.

The LF image records important information about the scene geometry that can be leveraged in many applications. It enables for instance to simulate a change of a viewpoint for static or dynamic observer which can also enhance the viewing experience in virtual reality (VR) applications [3]. The dense LF can also enable high-quality depth map estimation [4], [5] that can be used in the construction of an accurate point...
clouds [6], [7] and image rendering with varying depth of field and focus plane post-acquisition [8].

Fig. 2 illustrates the processing chain for LF image deployment. After acquisition, the LF image is processed by a pre-processing block to rearrange the data in an appropriate format for coding. The coding block removes spatial and angular redundancies in the LF image to reduce the data size for efficient storage and transmission. The decoding block recovers from the bitstream the LF image which is then processed by the post-processing block. This latter may perform calibration, color correction with associated meta data or creating new interpolated views, synthetic aperture, refocusing, and extended focus for visualisation by the viewing block. The LF image creates a large amount of data raising new challenges to the compression research community to design efficient coding solutions that drastically reduce the size of the LF image while providing a high quality of experience in terms of immersion and realism offered by this technology. In response, several coding approaches have been proposed in the literature which depend on the acquisition process of LF image and its representation.

In this paper, we investigate a lossy coding of lenslet-based acquisition LF image, also known as lenslet light field (LLF) imaging. In the LF processing chain illustrated in Fig. 2, our contributions build the coding and decoding blocks that process the input LLF image and the encoded bitstream, respectively. The SAIs (referred here to as views) of the LF image are first arranged in a pseudo-video sequence, which is then encoded with the latest video coding experts group (VCEG)/motion picture experts group (MPEG) video coding standard called versatile video coding (VVC) in temporal hierarchical coding configuration (i.e., temporal scalability). Only a sparse set of reference views, illustrated in Fig. 3, are encoded at low temporal layers and are then used as reference to whether encode or synthesize the rest of views at the decoder side. We propose a LF dual discriminator generative adversarial network (LF-D2GAN) to synthesize the missing views at the decoder side. The LF-D2GAN consists of one generator and two discriminators. The generator is composed of two components to predict the disparity and colors of a missing LF view. To enhance the generator performance, the training process is guided by two discriminators combining Kullback-Leibler (LK) and reverse LK divergences into a unified objective function. Furthermore, in order to avoid large fluctuations quality across the reconstructed views, we propose a multi-view quality enhancement convolutional neural network (MV-QENet) as post-processing to propagate the quality from views decoded at high quality to other views. Fig. 4 illustrates in red and green curves the quality fluctuations in peak signal-to-noise ratio (PSNR) of the LF views reconstructed before and after performing the quality enhancement block, respectively. The performance of the proposed solution has been extensively assessed and compared with the state-of-the-art solutions. The experimental results showed the superiority of the proposed approach in terms of both coding efficiency and visual quality.

The rest of this paper is organized as follows. Section II gives a review on existing LF image coding solutions. The proposed solution is then described in Section III. The performance of the proposed solution is assessed and analyzed in Section IV in terms of both coding efficiency and complexity. Finally, Section V concludes the paper.
### II. RELATED WORK

To ensure efficient storage and transmission of LLF imaging, many coding solutions have been proposed in recent years. In this section, we will give a brief review on LLF image coding solutions available in the literature. For more exhaustive description of these solutions, the reader may also refer to two overview papers recently published in [30] and [31]. Table I summarises the main features of the covered solutions in terms of fidelity (lossy or lossless coding), data representation prior encoding, consideration of the geometry and sparse representation, the adopted coding approach, and finally the compliance with a coding standard. The geometry-related data can represent the distances of a 3D scene such as depth or disparity information, when not available at the decoder, this geometry-related data can be estimated from the decoded views (texture).

Authors in [15] proposed a BI-SS estimation and compensation to remove spacial and angular redundancies within the LF image in MI representation. The BI-SS prediction is proposed as an additional prediction mode under the high efficiency video coding (HEVC) encoder in Intra coding configuration. The BI-SS prediction performs whether unipredictive or bi-predictive coding with reference blocks from already decoded and filtered (in-loop) blocks in the same image. The bi-predictive coding performs a weighed combination of two candidate blocks as in HEVC bi-directional prediction but the reference blocks are from the same image. The best coding mode among the 35 Intra prediction modes and the two new modes is selected by the encoder through a regular rate-distortion optimization process.

A two streams coding scheme has been proposed by Viola et al. in [9] for LF image in subaperture representation. The LF views are first split in two sets (streams). The views of the first set (reference views) are arranged in a pseudo-video sequence which is then encoded with HEVC encoder in Inter configuration (low delay configuration). The encoder also estimates the graph that models the dependencies among the LF image views. The graph weights are quantized and transmitted to the decoder. This latter decodes the reference views and with the decoded graph, it solves an optimization problem to recover the rest of views, and at low bitrate, enhance the quality of the reference views.

A 4D separable transform through the 4D-discrete cosine transform (DCT) is used in [10] to decorrelate the LF image and concentrate its energy in few coefficients. These coefficients are then clustered using hexadec tree structure, where each node corresponds to a 4D block of transform coefficients in a specific sub-band. Each node in this tree structure can be further sub-divided into sixteen children (sub-regions) and a binary symbol \( I \) is encoded otherwise \( \bar{I} \) is encoded for no further split. The decision to terminate the recursive split process is taken when a sub-region contains only zero coefficients or a single non-zero coefficient. The binary symbols of the constructed hexadec tree with the quantized direct current (DC) and alternating current (AC) DCT coefficients are encoded by a context-adaptive binary arithmetic coding (CABAC) with three contexts. One binary context is used for segmentation flags and two non-binary contexts for DC and AC coefficients in each sub-band. This solution has been further enhanced in terms of both coding efficiency and random access feature. The solution was adopted by the JPEG Pleno standard as the 4D transform mode.

Authors in [11] proposed a \( N \)-layer hierarchical coding scheme for LF image in subaperture representation. The LF views are first arranged in \( N \)-layer structure, where \( N \) is set to 6. The \( N - 1 \) first layers are called reference layers since the associated views are used as reference to encode views at higher layers. The views of the first layer with the corresponding inverse depth map\(^2\) are encoded with JPEG 2000 [32]. The inverse depth map of a view at higher layer is synthesized from the reference inverse depth map with a simple pixel-wise wrapping operation. The reference view candidate is then wrapped to the location of the view to generate the so-called wrapped reference view candidate. The wrapped reference views are fused in a single high quality reference view. The fusion is performed by a simple least-

| Solution            | Fidelity | Representation | Geometry | Sparsity | Coding approach                                                                 | Standard-compl. |
|---------------------|----------|----------------|----------|----------|---------------------------------------------------------------------------------|-----------------|
| Viola et al. [9]    | Lossy    | SAIs           | ✓        | ✓        | HEVC & Graph-based representation and coding                                    |                 |
| De Carvalho et al.  | Lossy    | SAIs           | ✓        | ✓        | 4D-DCT transform & Hexadeca-trees                                                |                 |
| Astola et al. [11]  | Lossy    | SAIs           | ✓        | ✓        | JPEG 2000 & Wrapping and Sparse prediction                                       | JPEG Pleno      |
| R.A. Farrugia et al.| Lossless | SAIs           | ✓        | ✓        | 4 wavelet transform                                                              |                 |
| Ahmad et al. [13]   | Lossy    | MI             | ✓        | ✓        | Multi-View HEVC                                                                  | MV-HEVC         |
| Conti et al. [17]   | Lossy    | MI             | ✓        | ✓        | HEVC & bi-prediction self-similarity (BI-SS)                                     |                 |
| Liu et al. [18]     | Lossy    | MI             | ✓        | ✓        | HEVC & Gaussian Process Regression (GPR)-based prediction                       |                 |
| Jiang et al. [19]   | Lossy    | MI             | ✓        | ✓        | Homography-based Low Rank Approximation                                           |                 |
| Dib et al. [20]     | Lossy    | SAIs           | ✓        | ✓        | Super-Ray Based Low Rank Approximation                                           |                 |
| Zhao et al. [21]    | Lossy    | SAIs           | ✓        | ✓        | Pseudo-video sequence & JEM codec                                               |                 |
| Liu et al. [22]     | Lossy    | SAIs           | ✓        | ✓        | Pseudo-video sequence & JEM codec                                               |                 |
| Hou et al. [23]     | Lossy    | SAIs           | ✓        | ✓        | HEVC & CNN-based angular super-resolution                                       |                 |
| Jia et al. [24]     | Lossy    | SAIs           | ✓        | ✓        | HEVC & LF-GAN                                                                   |                 |
| Zhao et al. [25]    | Lossy    | SAIs           | ✓        | ✓        | HEVC & Linear Approximation                                                     |                 |
| Bakir et al. [26]   | Lossy    | SAIs           | ✓        | ✓        | HEVC, Linear Approximation and CNN                                               |                 |
| Wang et al. [27]    | Lossy    | SAIs           | ✓        | ✓        | HEVC & Multibranch Spatial Transformer Networks                                  |                 |
| Komatsu et al. [28] | Lossy    | SAIs           | ✓        | ✓        | Binary images representation of the LF                                          |                 |
| Chen et al. [29]    | Lossy    | SAIs           | ✓        | ✓        | HEVC & Global Multiscale Representation                                          | HEVC           |
| Conti et al. [30]   | Lossy    | SAIs           | ✓        | ✓        | HEVC & field-of-view scalability                                                |                 |
| Zhao et al. [31]    | Lossy    | SAIs           | ✓        | ✓        | MV-HEVC & Super Resolution CNN                                                   |                 |
| Proposed            | Lossy    | SAIs           | ✓        | ✓        | VVC, LF-D2GAN and MV-QENet                                                      | MV-HEVC         |

\(^2\)The inverse depth map corresponds to the ratio between the camera focal and the pixel depth value.
squares regression technique. The coefficients derived from this latter step are quantized in 16 bits and sent to the decoder. A least-squares minimization method is also used to predict the encoded view from the constructed reference view and the resulting non-zero coefficients for each color component are encoded with an arithmetic encoder. Finally, the prediction residue, which is the difference between the view to encode and the merged reference view, is encoded with JPEG 2000. This solution has also been adopted by JPEG Pleno standard as the 4D predictive mode.

Ahmad et al. [13] proposed to arrange the LF image as a multi-view sequence that is encoded by the Multi-View extension of HEVC standard (MV-HEVC) [33]. A row of SAIs as shown in Fig. 3 corresponds to a single view in the multi-view sequence. Temporal and multi-view predictions are used to efficiently leverage the spatial and angular correlations in the LF image. Four hierarchical predictions levels with specific quantization parameters (QPs) were defined in horizontal (i.e., temporal) and vertical (i.e., views) directions to perform efficient prediction of the SAIs. This solution has been described in more details and assessed under the JPEG Pleno test conditions in [34]. The MV-HEVC extension has also been used in [14] to encode the SAIs arranged in four quadrants. All views are encoded one quadrant after another to reduce the reference buffer size. Under each quadrant, a hierarchical coding configuration is used to leverage angular and spatial correlations within the views.

Jiang et al. [17] proposed a coding method called homography-based low rank approximation (HLRA). This method jointly optimizes global or multiple homographies that align the LF views and low rank approximation matrices. Global or multiple homographies configuration is selected depending on the variation of the disparity across the views. The low-rank representation of the LF image is then encoded with HEVC. Dib et al. [18] proposed a compression scheme for LF image using super-ray based local low rank models. A novel method for disparity estimation and compensation was proposed so that the super-rays are constructed to yield the lowest approximation error for a given rank. This representation is based on two low rank models, one for the central view pixels that are visible in all views while the other is used for occlusions. Authors in [20] proposed a new coding concept for 4D LF relying on a new representation of the LF with \( N \) binary images and the corresponding weights. A set of binary basis images is selected to capture a common structure among all viewpoints, and the difference among the viewpoints are represented with pixel-independent weight. A least squares problem is solved to derive the \( N \) binary images and the corresponding weights. These images can then be encoded with an arithmetic encoder.

Several works [19], [20] have investigated a straightforward coding approach that organizes the LF views in a pseudo-video sequence, which is then encoded with a classical hybrid video encoder. For instance, Liu et al. [29] proposed a compression of LF image based on pseud-video sequence of SAIs. A subset of views is then arranged in a specific coding order that accounts for similarities between adjacent views and encoded using the joint exploration model (JEM) encoder.

Another approach consists in encoding a spare set of views using a video encoder, while the rest of views are synthesized at the decoder side. The latter solution has been followed by several authors [21]–[29], for instance, linear approximation has been investigated in [23] to estimate the views at the decoder from neighbour views, while a combination of linear approximation and convolutional neural network (CNN) has been proposed in [24] to synthesize missing views at the decoder side. In the same way, Jia et al. [22] proposed to use the generative adversarial network (GAN) to generate unsampled views. To enhance the coding efficiency, the authors proposed to encode and transmit the residual error between the generated uncoded views and their original versions. Hou et al. [21] proposed a method that exploits the inter- and intra-view correlations effectively by characterizing its particular geometrical structure using both learning and advanced video coding techniques. The SAIs are first partitioned into key and non-key SAIs. The key SAIs are encoded with a 2D video encoder while the non-key images are synthesized at the decoder side by a learning-based angular super-resolution approach. The residual images between the original non-key SAI and their synthesized versions are also arranged in a pseudo-video sequence and encoded with a video encoder. Wang et al. [25] proposed a novel light field image compression scheme using multibranch spatial transformer networks (MSTN) based view synthesis. First, a sparse subset of views, arranged into a pseudo-video sequence, are encoded by a video encoder. Then, the rest of views are synthesized based on the similarity between neighboring views with the MSTN block. This latter enables better characterization of the non-linear relationship between the sub-views with adaptive learning of the affine transformations between the neighboring views, which are used to warp the input views to generate accurate high-order approximation of the missing views. In [27], the authors proposed to encode a set of reference views with HEVC while the rest of views are estimated at the decoder side in two steps. The first step predicts a disparity-based global representation and then the prediction is performed as a second step based on multiplane as the form of this global representation. The reference views are encoded in [29] with the MV-HEVC standard. The quality of these views is first enhanced at the decoder side with a quality enhancement CNN. The resulting enhanced views are then used to as input to predict the reset of views with two super-resolution CNNs.

All these latest LF coding solutions based on view synthesis exploit a sparse representation of the LF image and differ in the process of selecting the key of views and especially in the view synthesis algorithm that can be [31]: 1) an Depth Image Based Rendering (DIBR) based method, 2) a transform-assisted method and 3) a learning-based method.

The LF coding solutions using learning-based view synthesis obtained the highest coding performance compared to other view synthesis techniques. However, these solutions suffer from three major drawbacks: first, the used synthesis module is based on a learning approach generally relying on a variant of CNN to synthesize the discarded views. Hence, a learning-based approach shows a large variation in performance for a set of LF images with different color, spatial and occlusion
characteristics. Second, in most of the proposed methods, the number of encoded/dropped views is predetermined and set manually, which does not reflect the best selection choice that leads to the highest coding efficiency. Finally, we noticed a large visual quality fluctuation across the reconstructed views at the decoder side. This quality fluctuation may result in lower viewing experience in many LF applications.

Our proposed solution belongs to the learning-based view synthesis method class and has been designed to overcome the mentioned drawbacks with the following main contributions:

- A use of VVC’s temporal scalability structure to drop views without impacting the decoding of other ones and without signaling them in the bitstream, thus keeping the bitstream compliant with the VVC codec.
- A learning-based view synthesis method called LF-D2GAN which is based on two CNNs for color and disparity estimations that are simultaneously trained with two adversarial discriminators.
- A rate-distortion optimization algorithm that selects at the encoder side whether the non-reference views should be encoded with VVC or synthesized at the decoder side by the LF-D2GAN block.
- A novel MV-QENet method which is applied as a post-processing to improve the quality of the non-reference views. This MV-QENet block propagates the quality to the target non-reference view from two carefully selected reference views. The MV-QENet block allows reducing the quality fluctuation across the decoded LF views and thus increases the viewing experience.

III. PROPOSED METHOD

In this paper, we propose a novel approach to encode a LF image in subaperture representation. The $N$ SAIs (views) are first split into a sparse set of $N_R$ reference views ($V_{p_1}, \ldots, V_{p_{N_R}}$) and $N - N_R$ non-reference views ($V_{q_1}, \ldots, V_{q_{N-N_R}}$), with $p_i$ and $q_j$ are the angular $(u,v)$ positions of the reference and non-reference views, respectively. All these views are arranged in a pseudo-video sequence which is then encoded with a hybrid Intra/Inter video encoder in hierarchical coding configuration (i.e., temporal scalability).

The reference views are encoded at low temporal layers and are used as reference for encoding the non-reference views. These latter are encoded at higher temporal layers and thus are not used as reference to encode the reference views. The non-reference views are also synthesized at the encoder side with a synthesis block that takes as input the decoded reference views. The encoder then performs a rate-distortion optimization between the synthesized and decoded non-reference views and selects the one that minimizes the rate-distortion cost. The bitstream is therefore composed of reference views and a set of non-reference views encoded with a video encoder. The non-reference views for which the rate-distortion cost is lower with the synthesis block are discarded from the bitstream without impacting the decoding of the transmitted views. The decoder performs inverse encoding operation to decode the transmitted views. The non-reference views dropped by the encoder are then synthesized and, thereafter, a quality enhancement is performed on them as a post-processing to ensure consistency of quality between views.

The block diagram of the proposed LF image coding scheme is illustrated by Fig. 5. In the rest of this section we will investigate in more details the elementary blocks of our proposed approach including the 2D video encoder, view synthesis, rate-distortion optimization and post-processing quality enhancement.

A. LF pseudo-video sequence encoding

The LF image presents large angular and spatial correlations in the SAIs. These SAIs when arranged in a pseudo-video sequence can be efficiently encoded with a hybrid video encoder that leverages these correlations through Intra/Inter predictions and transform coding. The joint video exploration team (JVET), jointly established by ISO/MPEG and ITU/VCEG standardisation committees, has released in July 2020 the latest video coding standard called VVC [35]. VVC enables a bitrate saving of 35% to 50% with respect to its predecessor HEVC standard for the same visual quality [36]. This coding gain is enabled by several coding tools at different levels of the coding chain including frame partitioning, Intra/Inter predictions, transform, quantization, entropy coding and in-loop filters. In particular, VVC performs more efficient Intra
and Inter predictions than HEVC by either enhancing HEVC tools or introducing new ones [37]–[39].

The proposed approach is based on the VVC standard to encode the pseudo-video sequence in temporal scalability configuration. However, our approach is codec agnostic in that it can be used with any 2D video codec that supports hierarchical groups of pictures (GOP) structure and temporal scalability.

The advanced Intra/Inter VVC tools will take advantage of the spatial and angular redundancies of the LF image. VVC supports by design temporal scalability through the random access coding configuration. This latter, illustrated in Fig. 6 enables different temporal layers and each temporal layer uses as reference only frames from lower temporal resolution, i.e., lower layer. Therefore, frames of each temporal layer $t_i$ can be dropped without impacting the decoding of frames at lower temporal resolution $t_j$ with $t_i > t_j$. In the proposed coding approach, we leverage the concept of temporal resolution to drop views at the encoder without impacting the decoding process and thus performing the best rate-distortion performance.

**B. LF Dual Discriminator Generative Adversarial Nets**

Several solutions [40]–[42] have been proposed in the literature to synthesize a novel or missing LF view. The problem in the coding scheme consists in estimating a missing view $\tilde{V}_q$ from a sparse set of decoded $N_R$ reference views $\tilde{V}_{p_1}, \tilde{V}_{p_2}, \ldots, \tilde{V}_{p_{N_R}}$

$$\tilde{V}_q = f(\tilde{V}_{p_1}, \tilde{V}_{p_2}, \ldots, \tilde{V}_{p_{N_R}}, q),$$

where $p_1, p_2, \ldots, p_{N_R}$ and $q$ are the $(u, v)$ positions of the $N_R$ reference views and the estimated non-reference missing view $\tilde{V}_q$, respectively.

Inspired by the success of CNN architectures for LF view synthesis [42], we propose to use a learning-based approach to synthesize the missing views at the decoder side. The proposed synthesis block is composed of one generator $G$ and two discriminators $D_1$ and $D_2$. Similar to [42], the generator is broken-down into two CNNs for efficient estimation of disparity and color, as illustrated in Fig. 7. These two sequential components are trained simultaneously to minimize a cost function. Our contribution in this block consists in enhancing the performance of the generator by conducting unsupervised learning guided by two discriminators. The disparity CNN estimates the disparity of the missing view $D_q$ from a set of features $K$ computed from the input reference views

$$D_q = g_d(K),$$

where $g_d$ is the function that computes the relationship between the input features and the disparity of the target view. The input features $K$ consist mainly of mean and standard deviation of input reference views wrapped at different disparity levels.

Using the estimated disparity $D_q$, the reference views are then wrapped to the target view

$$\tilde{V}_{p_i}(s) = \tilde{V}_{p_i} \left[ s + (p_i - q)D_q(s) \right],$$

where $s$ is the $(x, y)$ position pixel.

The $N_R$ wrapped reference views $\tilde{V}_{p_1}, \tilde{V}_{p_2}, \ldots, \tilde{V}_{p_{N_R}}$ are provided to the color estimation CNN $g_c$ in order to estimate the color of the missing view. The color CNN estimates the missing view by using all wrapped reference views $\tilde{V}_{p_1}, \tilde{V}_{p_2}, \ldots, \tilde{V}_{p_{N_R}}$, its disparity map $D_q$ estimated by the disparity CNN and its position $q$

$$\tilde{V}_q = g_c(\tilde{V}_{p_1}, \ldots, \tilde{V}_{p_{N_R}}; D_q, q).$$

As mentioned in Section I, the proposed coding approach is based on the LF-D2GAN block. GANs are deep neural network architectures composed of two consecutive neural network models, namely generator $G$ and discriminator $D$. GAN enables to simultaneously train the two models: the generative model $G$ that captures the data distribution, and the discriminative model $D$ that estimates the probability that a sample came from the training data rather than from the generator $G$ [43]. GAN has recently achieved great success in various fields, especially in fake video generation, LF super-resolution and objects detection [44], [45].

The training of the two generators $g_d$ and $g_c$ is guided by two discriminators $D_1$ and $D_2$. Given an input data $x$ which consists here in an input data patch, the first discriminator $D_1$ rewards a high score for real data ($P_{train}$) and returns low score for data generated by the generator ($P_{fake}$). In contrast, the second discriminator $D_2$ returns low score when the input data follows the real data distribution and high score for the input data close to the model distribution. The two generators are then trained simultaneously to generate samples that fool the two discriminators in a three-player minimax optimization game

$$\min_{G} \max_{\theta_1, \theta_2} \mathcal{L}(\theta_G, \theta_1, \theta_2) = \alpha \mathbb{E}_{x \sim P_{train}} \left[ \log D_1(x) \right] + \mathbb{E}_{x \sim P_{fake}} \left[ -D_1(G(x)) \right] + \beta \mathbb{E}_{x \sim P_{train}} \left[ \log D_2(G(x)) \right],$$

where $\mathbb{E}$ represents expected value, $x$ is the real data, $\tilde{x}$ is the generated data, $P$ represents the probability distribution, $\alpha$ and $\beta$ are two hyper-parameters ($0 < \alpha, \beta \leq 1$) to stabilize the learning of the model and control the effect of $\text{KL}$ and reverse $\text{KL}$ divergences on the optimization problem [46]. The models are trained by alternatively updating discriminators parameters $\theta_{D_1}, \theta_{D_2}$ and the generator parameters $\theta_G$ by solving a minimax optimization game.

Three cost functions defined in [7], [8] and [9] are computed to obtain the error that should be transmitted respectively to $D_1$, $D_2$ and $G$ for their backward weights updating, as shown in Fig. 7 (dash lines). Thus, (7) and (8) are used to update
the weights of the discriminators $D_1$ and $D_2$, respectively, by ascending the obtained stochastic gradient.

$$\nabla_{\theta_{D_1}} \frac{1}{M} \sum_{m=1}^{M} \alpha \log D_1(x^{(m)}) - D_1(G(z^{(m)}))$$ (7)

$$\nabla_{\theta_{D_2}} \frac{1}{M} \sum_{m=1}^{M} \beta \log D_2(G(z^{(m)})) - D_2(x^{(m)})$$ (8)

while (7) represents the cost function to gain the error that should be given to the generator $G$ for its weight updating.

$$\nabla_{\theta_G} \frac{1}{M} \sum_{m=1}^{M} \beta \log D_2(G(z^{(m)})) - D_1(G(z^{(m)}))$$ (9)

C. Rate-distortion optimization

Instead of fixing the number of dropped views, in our approach this is done adaptively on the basis of a rate-distortion optimization (RDO) process. At the encoder side, first, LF subaperture views are organized into groups of 16 views that form GOP, as illustrated in Fig. 6. Next, in each GOP, the images at the remaining levels 3 and 4 are either encoded using the VVC codec or dropped. For these decisions, we propose a RDO algorithm to select whether the VVC coding scheme to encode reference views. Thus, the proposed RDO block can decide which views from the upper levels exclusively to the RDO block

Algorithm 1 RDO block based Lagrangian optimization

Require: $J \gets \{ \forall v \in TL[3 \text{ or } 4], \forall m \in \{VVC, LF - D2GAN\}, J = D + \lambda R \}$

for all $v \in TL[4]$ do
  if $J(VVC) < J(LF-D2GAN)$ then
    Encode $v$ by VVC
    Send($v$)
  else
    Generate $v$ by LF-D2GAN
  end if
end for

for all $v \in TL[3]$ do
  if $J(VVC) > J(LF-D2GAN) \text{ and all dependent}$ views are synthesized by LF-D2GAN then
    Generate $v$ by LF-D2GAN
  else
    Encode $v$ by VVC
    Send($v$)
  end if
end for

are, firstly, after an extensive study, we found that these levels together represent around 28% of the total bitrate. Second, the views at the upper levels are not used as references in the VVC coding scheme to encode reference views. Thus, the proposed RDO block can decide which views from the upper level can be encoded using VVC or dropped and synthesized using LF-D2GAN. To reach this goal, the encoder computes the rate-distortion (RD) cost function $J$ given by (10) for both the VVC decoded view and the one synthesized by the LF-D2GAN.

$$J = D + \lambda R,$$ (10)

where $\lambda$ is the Lagrangian multiplier, $D$ is the distortion and $R$ is the rate in bits per pixel (bpp). To set the Lagrangian multiplier $\lambda$, we empirically determine its value by testing a large set of LF images. We found that the value of 0.1 for $\lambda$ is

Fig. 7: Architecture of the proposed LF dual discriminator generative adversarial network (LF-D2GAN).

| Convolution Layer | ReLU | Fully Connected | Batch Normalization + ReLU |
|-------------------|------|-----------------|---------------------------|

Generator

Discriminator

- **Convolution Layer**
- **ReLU**
- **Fully Connected**
- **Batch Normalization + ReLU**
optimal and for which the Lagrangian optimization is giving the best performance.

It should be noted that the dropped views are not signaled in the bitstream and the decoder can detect the missing views based on the picture order count (POC) of the decoded frames in the GOP. The POC of a non-reference view not present in the bitstream is identified and its angular position $q$ in the GOP. The POC of a non-reference view not present in the bitstream and the decoder can detect the missing views based on the picture order count (POC) of the decoded frames in the bitstream and the decoder can detect the missing views based on the picture order count (POC) of the decoded frames in the bitstream and the decoder can detect the missing views based on the picture order count (POC) of the decoded frames in the bitstream.

### D. Multi-View Quality Enhancement Net

After analyzing the quality of each view at the output of the RDO block, we noticed that there is a significant fluctuation in the quality of the decoded views. Specifically, we have found that the non-Lagrange views have lower quality than the reference ones. This quality fluctuation is caused by the high QP values assigned to the views at high temporal layers in the VVC hierarchical coding structure on the one hand, and the unpredictable output quality of the LF-D2GAN block on the other hand. Thus, we propose to perform a post-processing on these non-reference views at the decoder side using MV-QENet block to further enhance their quality and reduce the quality fluctuation across LF views, as shown in Fig. 8. Here, the concept of quality enhancement consists in predicting the residual errors $R_q = V_q - \hat{V}_q$ of the non-reference views using a CNN. At the decoder side, the proposed CNN architecture uses 3 views as an input. These three views include the target decoded non-reference view $\hat{V}_t$, the decoded central view $\hat{V}_c$ (which is of the highest quality as it is encoded in Intra using low QP value) and one neighbor decoded reference view $\hat{V}_p$. $\hat{V}_p$ is selected among the reference views (except the central view already included in input) through a blind image quality assessment (IQA) metric called codebook representation for no-reference image assessment (CORNIA) [47]. This latter has the advantage of providing image quality scores without access to reference images and showed a high correlation with humane appreciations.

Thus, at the decoder, the reference view selection (RVS) block picks among 15 neighbors views the $\hat{V}_p$ view with the highest quality score computed by CORNIA metric. The 3 views ($\hat{V}_t, \hat{V}_c, \hat{V}_p$) are then fed to the MV-QENet which extracts the multiscale characteristics of the views and constructs a densely connected mapping in order to predict the residual errors and transfer the quality of neighbor views to the target view

$$\hat{V}_t = \hat{V}_t + h_\phi(\hat{V}_t, \hat{V}_c, \hat{V}_p),$$

(11)

with $h_\phi$ is the parametric function of quality enhancement neural network and $\phi$ its trainable parameters.

The architecture of this neural network is composed of two key components: the multiscale feature extraction (denoted by layers C1-9 in Fig. 8) and the densely connected mapping construct (denoted by layers C10-14 in Fig. 8). Multi-scale features extraction takes as input two reference views ($\hat{V}_c, \hat{V}_p$) and one target non-reference view $\hat{V}_t$. The spatial characteristics of these three views are extracted by multiscale convolutional filters. After feature extraction, all feature maps from the input three views are concatenated, then flow into the densely connected block component. After obtaining the feature maps of these three views, the densely connected architecture is applied to build the nonlinear mapping of feature maps in order to improve the residual part. In fact, there are 5 convolutional layers in the nonlinear mapping of the densely connected architecture. Each of them has 32 convolutional filters with size of $3 \times 3$. In addition, dense connection [48] is adopted to encourage feature reuse, strengthen feature propagation and mitigate the vanishing-gradient problem. Moreover, a batch normalization (BN) is applied to all 5 layers after PReLU activation to reduce internal covariate shift, thus accelerating the training process.

We denote the composite non-linear mapping as $H_t(\cdot)$, including Convolution (Conv), PReLU and BN. We further

### TABLE II: Convolutional layers of MV-QENet block.

| Layers | C1/4/7 | C2/5/8 | C3/6/9 | C10-14 | C15 |
|--------|--------|--------|--------|--------|-----|
| Filter size | $3 \times 3$ | $5 \times 5$ | $7 \times 7$ | $3 \times 3$ | $3 \times 3$ |
| Filter number | 32 | 32 | 32 | 32 | 1 |
| Stride | 1 | 1 | 1 | 1 | 1 |
| Function | BN+PReLU | BN+PReLU | BN+PReLU | BN+PReLU | BN+PReLU |

Fig. 8: Detailed architecture of the proposed multi-view quality enhancement convolutional neural Network (MV-QENet).
denote the output of the l-th layer as \(x_l\), such that each layer can be formulated as follows:

\[
\begin{align*}
x_{11} & = H_{11}(\{x_{10}\}), \\
x_{12} & = H_{12}(\{x_{10}, x_{11}\}), \\
x_{13} & = H_{13}(\{x_{10}, x_{11}, x_{12}\}), \\
x_{14} & = H_{14}(\{x_{10}, x_{11}, x_{12}, x_{13}\}),
\end{align*}
\]

where \(x_{10}, x_{11}, x_{12}, x_{13}\) refers to the concatenation of the feature maps produced in layers C10-C14. Finally, the enhanced target view \(\hat{V}_t\) is generated by the pixel-wise summation of learned enhancement residual \(R_t(\theta_qe)\) and input target view \(V_t\)

\[
\hat{V}_t = \hat{V}_t + R_t(\theta_qe),
\]

with \(\theta_qe\) is defined as the trainable parameters of the MV-QENet. The MV-QENet is trained with minimizing a mean squared error loss function

\[
\text{loss} = \|V_t - \hat{V}_t\|^2_2.
\]

It should be noted that the \(N_R\) reference views are not enhanced by the MV-QENet and thus \(\hat{V}_{p_i} = V_{p_i}, \forall i \in \{1, \ldots, N_R\}\).

IV. EXPERIMENTAL RESULTS

In this section, we first give the test material used to train the learning-based models and the testing conditions used to assess and compare the proposed solution with respect to state-of-the-art methods. The performance of the proposed solution are then assessed in terms of coding efficiency, visual quality and complexity at both encoder and decoder sides.

A. Experimental configurations

1) LF-D2GAN training: the proposed LF-D2GAN architecture described in the previous section was trained with 140 LLF images, where 70 LLF images are from EPFL dataset [49], 50 LLF images are from Stanford Lytro LF image dataset [50] and 20 LLF images are from HCI dataset [51]. A validation set was also considered with 14 LLF images from these three data sets (8, 4 and 2 images from EPFL, Stanford Lytro and HCI datasets, respectively). Each subaperture view was split into patches of size 60 \(\times\) 60, thus resulting in more than 150,000 patches that were used in the training phase. The training configuration of LF-D2GAN was set as follows: we trained the generator \(G\) and two discriminators \((D_1\) and \(D_2\)) with the Adam optimizer [52] by setting \(\beta_1 = 0.9, \beta_2 = 0.999\), learning rate = 0.0002, batch-size of 10 and kernel size of convolutional layers as depicted in Fig. 7. The regularization coefficients of \(D_1\) and \(D_2\) were set as \(\alpha = 0.2\) and \(\beta = 0.2\), respectively. For the generator, we used input patch of 60 \(\times\) 60, stride of 16, and output patch equal to 36 \(\times\) 36 (reduced size is due to the convolutions).

2) MV-QENet training: the same training data set used to train the LF-D2GAN was considered to train the MV-QENet. The training set includes both original and decoded views at different QPs. The views were segmented into patches of 64 \(\times\) 64 as the training samples. The batch size was set to 128 and Adam optimizer [52] was used with an initial learning rate of 0.0002. The configurations of the different layers are summarised in Table I.

3) Testing conditions: for the testing phase, 9 LLF images different from the training and validation sets are selected, 6 LLF images are from EPFL dataset [49], 1 LLF image from Stanford Lytro LF dataset [50] and 2 LLF images from HCI dataset [51]. Each of these LLF images is composed of 8 \(\times\) 8 subaperture views \((N = 64)\).

The LF image views are arranged in a pseudo-video sequence using spiral order scan and encoded using VVC in random access (RA) coding configuration at 4 QP values of 18, 24, 28 and 32. The VVC test model (VTM) version 7.1 is used to encode the pseudo-video sequence in YCbCr 4:2:0 sampling color format. The \(N_R = 16\) reference views are selected as the four corner views of each quadrant as illustrated in Fig. 3. In this figure, the central reference view is highlighted in red color while the rest of disabled 48 views correspond to the non-reference views.

The proposed solution is compared with respect to six coding solutions including 1) VVC-All that encodes all views with VVC standard, 2) LF-D2GAN-16 that encodes the 16 reference views with VVC and the non-reference views are synthesized with the LF-D2GAN, 3) Liu et al. method [20], 4) Hou et al. method [21], 5) Jia et al. method [22] and 6) the proposed solution without the quality enhancement block (denoted as prop.-w/o-MV-QENet) [53]. The quality of the decoded views is assessed using both PSNR and structural similarity index measure (SSIM) [54] IQA metrics.

B. Coding and quality evaluation performance

Fig. 9 illustrates the average PSNR versus training iterations of the synthesis block on the validation set for three different architectures: CNN, GAN and the proposed LF-D2GAN. The CNN architecture is trained by minimizing the mean squared-error loss function between the synthesized and the original views, while the GAN and LF-D2GAN architectures are trained with one and two adversarial discriminators, respectively. It is clear from this figure that the proposed LF-D2GAN architecture relying on two discriminators provides higher PSNR quality performance on the validation set with smooth...
fluctuations of the quality and better convergence of the generator compared to both CNN and GAN architectures.

Fig. 10 gives the average PSNR performance versus the bitrate for the proposed and the six considered reference solutions on the 9 testing LF images. The first important observation is that the proposed solution performs better than the six reference solutions at all considered bitrates and for all test LF images. We can also notice that the three proposed components performing view synthesis (LF-D2GAN), rate-distortion (RD) optimization and quality enhancement (MV-QENet) bring significant quality improvements since our rate-distortion (RD) optimization and quality enhancement posed components performing view synthesis (LF-D2GAN), all test LF images. We can also notice that the three proposed components performing view synthesis (LF-D2GAN), rate-distortion (RD) optimization and quality enhancement (MV-QENet) bring significant quality improvements since our solution performs better than VVC-All, LF-D2GAN-16 and prop.-w/o-MV-QENet solutions.

Table 11 gives the performance of our solution and four reference solutions in terms of Bjøntegaard delta bit rate (BD-BR) and Bjøntegaard delta PSNR (BD-PSNR), both computed with respect to the anchor solution proposed by Liu et al. in [20]. We can notice that our solution achieves the highest BD-BR and BD-PSNR gains for the 9 test LF images. In average, our solution provides 36.22% bitrate reduction and increases the quality by 1.35 dB compared to the solution proposed in [20]. Compared to the second best performing solution, i.e., prop.-w/o-MV-QENet [53], our solution offers a relative bitrate gain of 8.12% and increases the quality by 0.52 dB. These scores highlight the significant gains brought by the different proposed blocks in terms of bitrate reduction and quality enhancement. These average gains are also substantial for the 9 individual test LF images.

Fig. 11 gives the SSIM performance of the VVC-All, prop.-w/o-MV-QENet and proposed methods for the 9 test LF images at two QP values 18 and 32. We can notice from this figure that our solution gives the highest SSIM scores for all LF images at both considered QPs. Table IV shows the SSIM-based BD-BR and Bjøntegaard delta SSIM (BD-SSIM) of our solution and two other methods with respect to the anchor method proposed in [20]. Our solution achieves the highest bitrate saving in average with around 42.95% compared to [20]. These scores are even higher compared to the PSNR-based bitrate savings reported in Table III. Compared to the prop.-w/o-MV-QENet solution [53], we can notice a relative bitrate saving of 12.94% in average, which highlights the contribution of the proposed MV-QENet post-processing.

Fig. 12 illustrates the visual quality of the decoded non-reference views resulting from Liu et al. [20], VVC-All, prop.-w/o-MV-QENet and proposed methods for 3 test LF images: Danger de Mort, Herbs and Stone-Pillars-Outside. We can see that our method provides a higher visual quality of the reconstructed views, especially after applying the MV-QENet post-processing. This latter enhances the visual quality of views by providing more details (high frequencies), especially
TABLE III: BD-BR and BD-PSNR performance calculated with respect to the anchor method proposed in [20].

| Image             | VVC-All | Jia et al. [22] | Hou et al. [21] | Prop.-w/o-MV-QENet [53] | Proposed |
|-------------------|----------|-----------------|-----------------|--------------------------|----------|
|                   | BD-BR | BD-PSNR | BD-BR | BD-PSNR | BD-BR | BD-PSNR | BD-BR | BD-PSNR |
| Bikes             | –11.7% | 0.72    | –6.3% | 0.48     | –6.9% | 0.49     | –22.4% | 0.96     | –31.56% | 1.19    |
| Danger de Mort    | –7.8%  | 0.22    | –10.8% | 0.28    | –8.7% | 0.26     | –16.5% | 0.40     | –25.69% | 0.78    |
| Flowers           | –12.3% | 0.56    | –11.9% | 0.54    | –16.2% | 0.72     | –16.6% | 0.74     | –23.66% | 1.03    |
| Ankylosaurus Dip1 | –13.2% | 0.44    | –14.9% | –0.72   | –12.3% | 0.39     | –18.0% | 0.57     | –31.17% | 1.15    |
| Aloe              | –26.4% | 0.85    | –9.1%  | 0.31     | –2.46% | 0.12     | –42.3% | 1.23     | –56.59% | 1.84    |
| Stone-pillars-outside | –18.3% | 0.61    | –15.1% | 0.52     | –11.9% | 0.28     | –35.6% | 0.98     | –49.76% | 1.42    |
| Bedroom           | –5.3%  | 0.46    | –4.0%  | 0.32     | –2.3%  | 0.18     | –9.5%  | 0.85     | –24.78% | 2.11    |
| Desktop           | –19.6% | 0.32    | –7.5%  | 0.11     | 44.1%  | –0.61    | –26.3% | 0.45     | –40.58% | 0.79    |
| Herbs             | –26.0% | 1.14    | –4.4%  | –0.11    | 6.9%   | –0.20    | –29.8% | 1.32     | –42.25% | 1.85    |
| Average           | –13.6% | 0.59    | –8.3%  | 0.35     | –0.54% | 0.15     | –24.1% | 0.83     | –36.22% | 1.35    |

Fig. 11: SSIM performance of three LF image coding methods for the 9 considered test LF images at two QP values.

TABLE IV: SSIM-based BD-BR and BD-SSIM performance calculated with respect to the anchor [20]. 1) Bikes, 2) Danger de Mort, 3) Flowers, 4) Ankylosaurus Dip1, 5) Aloe, 6) Stone-pillars-outside, 7) Bedroom, 8) Desktop, 9) Herbs.

| Image            | VVC-All | Jia et al. [22] | Hou et al. [21] | Prop.-w/o-MV-QENet [53] | Proposed |
|------------------|----------|-----------------|-----------------|--------------------------|----------|
|                   | BD-BR | BD-SSIM | BD-BR | BD-SSIM | BD-BR | BD-SSIM |
| Bikes             | –10.3% | 0.011   | –24.91% | 0.022    | –40.1% | 0.032   |
| Danger de Mort    | –12.4% | 0.013   | –22.10% | 0.021    | –33.7% | 0.027   |
| Flowers           | –13.0% | 0.020   | –23.15% | 0.028    | –32.8% | 0.042   |
| Ankylosaurus Dip1 | –16.0% | 0.011   | –22.8%  | 0.024    | –40.0% | 0.035   |
| Aloe              | –30.0% | 0.028   | –50.30% | 0.046    | –66.6% | 0.049   |
| Stone-pillars-outside | –19.80% | 0.004 | –40.49% | 0.021    | –54.0% | 0.031   |
| Bedroom           | –9.4%  | 0.010   | –17.22% | 0.015    | –30.6% | 0.024   |
| Desktop           | –18.0% | 0.011   | –37.09% | 0.029    | –42.5% | 0.036   |
| Herbs             | –19.8% | 0.020   | –32.10% | 0.041    | –46.1% | 0.059   |
| Average           | –16.54%| 0.014   | –30.01% | 0.027    | –42.95%| 0.037   |

C. Complexity analysis

The complexity of the proposed coding approach is evaluated and compared to the other methods on both CPU and GPU platforms. The performance has been carried-out on a PC equipped with an Intel core i9-7900X CPU running at 3.3 GHz with 64 GB memory and a TITAN Xp NVIDIA GPU. The complexity of our solution is assessed on CPU, where all modules run on the CPU, and on GPU when both LF-D2GAN and MV-QENet modules run on the GPU. Table V gives the processing time of four LF image coding methods. As we can see in Fig. 5, the decoder can be optimized by processing several decoding blocks in parallel. In addition, the MV-QENet block is optional and may or may not be applied depending on the computational resources available.
Fig. 12: Visual illustration of 3 test LF images encoded at around 0.014 bpp for Danger de Mort, 0.02 bpp for Herbs and 0.01 bpp for Stone-Pillars-Outside. The objective quality scores are provided for each illustrated view in this format PSNR(SSIM). Danger de Mort [view 49 (TL #4), RDO: Synthesized by LF-D2GAN], Liu et al. 37.0(0.84), VVC-All 37.2(0.87), prop.-w/o-MV-QENet 37.4(0.88), proposed 37.8(0.91); Herbs [view 47 (TL #4), RDO: Synthesized by LF-D2GAN], Liu et al. 30.8(0.80), VVC-All 31.7(0.85), prop.-w/o-MV-QENet 32.1(0.87), Proposed 32.6(0.90) and Stone-Pillars-Outside, [view 33 (TL #4), RDO: decoded by VVC], Liu et al. 35.7(0.81), VVC-All 36.2(0.84), prop.-w/o-MV-QENet 36.2(0.84), prop. 36.7(0.87).

at the decoder to the detriment of lower quality.

V. Conclusion

In this paper, we have proposed an efficient lossy coding scheme for LLF imaging in subaperture representation. The coding scheme is composed of four elementary blocks, including 2D video coding, view synthesis, rate-distortion optimization and view quality enhancement. The LF views are first arranged in a pseudo-video sequence which is encoded with the VVC standard in hierarchical temporal scalability configuration. The reference views are encoded at low temporal layers, while the rest of views are encoded at higher temporal layers. This coding structure enables to drop thanks to RDO block the non-reference views without impacting the decoding of reference views. The training of the proposed LF-D2GAN synthesis block is guided by two adversarial discriminators enabling better convergence of the generator and providing higher PSNR quality performance of the synthesized views. A novel quality enhancement block MV-QENet is applied at the decoder side on the non-reference views to further enhance their quality and ensure quality consistency between views.

The proposed coding solution has been assessed in terms of bitrate saving and visual quality using both PSNR and SSIM objective quality metrics. A significant bitrate saving has been achieved by the proposed method without affecting the visual quality. The obtained results clearly demonstrated the superiority of our solution with respect to the state-of-the-art methods.

As future work, we plan to consider more advanced LF image features such as the visual attention and viewing conditions.

REFERENCES

[1] G. Arun, “The light field,” Journal of Mathematics and Physics, vol. 18, no. 1-4, p. 51–151, 1939.
[2] M. Levoy and P. Hanrahan. “Light field rendering,” in Proc. 23rd Annu. Conf. Comput. Graph. Interactive Tech., May 1996, p. 31–42.
[3] F.-C. Huang, K. Chen, and G. Wetzstein, “The light field stereocope: Immersive computer graphics via factored near-eye light field displays with focus cues,” ACM Trans. Graph., vol. 34, no. 4, Jul. 2015.
Nader Bakir received the master’s degree in Applied mathematics from the Faculty of Science, Lebanese University, Beirut, Lebanon, in 2003, and the Ph.D. degree in signal and image processing from National Institute of Applied Sciences of Rennes, Institute of Electronics and Telecommunications of Rennes, Rennes, France, in 2020. Currently, he is an Lecturer at the Faculty of Business Administration, and researcher in the Center of Research and Studies in Legal Informatics at Lebanese University. His research is in the image processing field.

Wassim Hamidouche received Master’s and Ph.D. degrees both in Image Processing from the University of Poitiers (France) in 2007 and 2010, respectively. From 2011 to 2013, he was a junior scientist in the video coding team of Canon Research Center in Rennes (France). He was a post-doctoral researcher from Apr. 2013 to Aug. 2015 with VAADER team of IETR where he worked under collaborative project on HEVC video standardisation. Since Sept. 2015 he is an Associate Professor at INSA Rennes and a member of the VAADER team of IETR Lab. He has joined the Advanced Media Content Lab of b<>com IRT Research Institute as an academic member in Sept. 2017. His research interests focus on video coding and multimedia security. He is the author/coauthor of more than one hundred and thirty (+130) papers at top journals and conferences in Image Processing, two MPEG standards, two patents, several MPEG contributions, public datasets and open source software projects.

Sid Ahmed Fezza received his engineer degree from University of Dr. Tahar MOULAY, Saida, Algeria, in 2007 and the Ph.D. degree from Djillali Liabes University of Sidi-Bel-abbes, Algeria, in 2015, both in computer science. He is currently Associate Professor at National Institute of Telecommunications and ICT (INTTIC), Oran, Algeria. Sid Ahmed Fezza was a recipient of two top 10% best paper awards in ICIP 2014, the 2015 Algerian Paper of the Year Awards from the Algerian Network for Academics, Scientists and Researchers and has authored several publications in top journals and conferences on image and video processing. His research interests are in the fields of image/video processing, image/video coding, visual quality assessment, immersive multimedia communication and multimedia security.

Khouloud Samrout received the master’s degree in informatics and telecommunication engineering from the Department of Electronics and Telecommunication, Faculty of Engineering, Lebanese University, Tripoli, Lebanon, in 2011, and the Ph.D. degree in signal and image processing from Lebanese University and National Institute of Applied Sciences of Rennes, Institute of Electronics and Telecommunications of Rennes, Rennes, France, in 2014. Currently, she is an assistant professor and a researcher at the Faculty of Engineering, Electronic and Telecommunication Department at Lebanese University. Her research is in the image processing field.

Olivier Déforges received the Ph.D. degree in image processing in 1995. He is a Professor with the National Institute of Applied Sciences (INSA) of Rennes. In 1996, he joined the Department of Electronic Engineering, INSA of Rennes, Scientific and Technical University. He is a member of the Institute of Electronics and Telecommunications of Rennes (IETR), UMR CNRS 6164 and leads the IMAGE Team, IETR Laboratory including 40 researchers. He has authored over 130 technical papers. His principal research interests are image and video lossy and lossless compression, image understanding, fast prototyping, and parallel architectures. He has also been involved in the ISO/MPEG standardization group since 2007.