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

N. Bakir, Wassim Hamidouche, S.A. Fezza, K. Samrout, O. Deforges

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068563, IEEE Transactions on Multimedia

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 high coding performance and overcomes the state-of-the-art LF image compression methods by –36.22% in terms of BD-BR and 1.35 dB in BD-PSNR[1].**

***Index Terms*—Light Field, View Synthesis, Deep Learning, VVC, Coding Structure, RDO, Quality Enhancement.**

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

$$P(x, y, z, \theta, \phi, \lambda, t), \tag{1}$$

the rays in space are parameterized by 3D coordinates (x, y, z), two angles (θ, ϕ), wavelength λ 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.

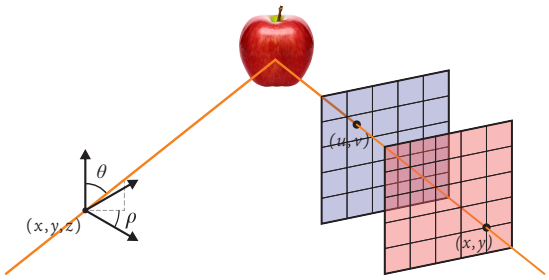


Fig. 1: 5D Plenoptic function on the left versus two-plane parametrization on the right (4D LF function).

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-image*s (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[2] 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

N. Bakir, W. Hamidouche and O. Déforges are with Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France (e-mail: whamidou@insa-rennes.fr).

SA. Fezza is with National Institute of Telecommunications and ICT, Oran, Algeria (e-mail: sfezza@inttic.dz).

K. Samrouth is with Lebanese University, Tripoli, Lebanon (e-mail: khouloud.samrout@gmail.com).

[1]The web page of this work is available at https://naderbakir79.github.io/LFD2GAN.html

[2]Plenoptic camera is also called narrow baseline plenoptic camera.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068563, IEEE Transactions on Multimedia

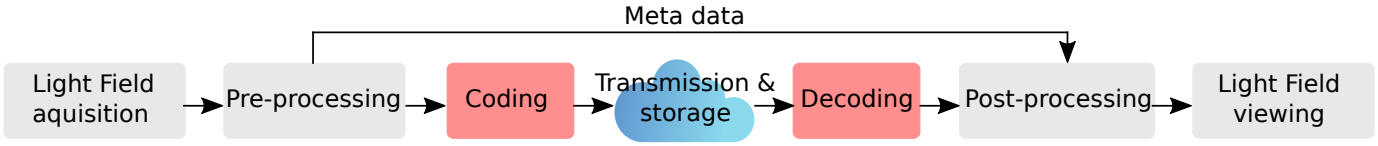


Fig. 2: Processing chain of light field technology from acquisition to end-user viewing.

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.

Fig. 3: Sparse representation of the 8×8 LF image in subaperture representation with 16 reference views including the center view highlighted in red.

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

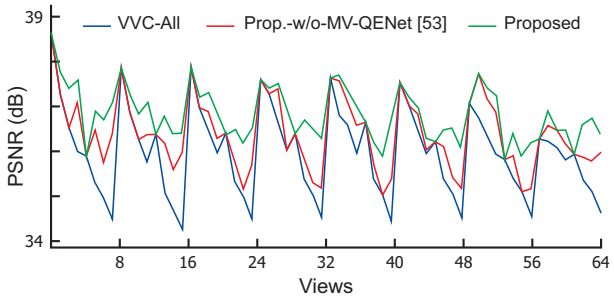


Fig. 4: Illustration of the quality fluctuation of the reconstructed LF views before and after performing the quality enhancement block (*Stone-Pillars-Outside* test LF image encoded at 0.0075 bpp).

(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.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068563, IEEE Transactions on Multimedia

TABLE I: Main features of the existing LLF image coding solutions.

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.* [10]	Lossy	SAIs	✖	✖	4D-DCT transform & Hexadeca-trees	JPEG Pleno
Astola *et al.* [11]	Lossy	SAIs	✔	✔	JPEG 2000 & Wraping and Sparse prediction	JPEG Pleno
R.A. Farrugia *et al.* [12]	Lossless	SAIs	✖	✖	4D wavelet transform	✖
Ahmad *et al.* [13] & [14]	Lossy	SAIs	✖	✖	Multi-View HEVC	MV-HEVC
Conti *et al.* [15]	Lossy	MI	✖	✖	HEVC & bi-prediction self-similarity (BI-SS)	✖
Liu *et al.* [16]	Lossy	MI	✖	✖	HEVC & Gaussian Process Regression (GPR)-based prediction	✖
Jiang *et al.* [17]	Lossy	SAIs	✔	✖	Homography-based Low Rank Approximation	✖
Dib *et al.* [18]	Lossy	SAIs	✔	✖	Super-Ray Based Low Rank Approximation	✖
Zhao *et al.* [19]	Lossy	SAIs	✖	✖	Pseudo-video sequence & JEM codec	✖
Liu *et al.* [20]	Lossy	SAIs	✖	✖	Pseudo-video sequence & JEM codec	✖
Hou *et al.* [21]	Lossy	SAIs	✖	✔	HEVC & CNN-based angular super-resolution	✖
Jia *et al.* [22]	Lossy	SAIs	✖	✔	HEVC & LF-GAN	✖
Zhao *et al.* [23]	Lossy	SAIs	✔	✔	HEVC & Linear Approximation	✖
Bakir *et al.* [24]	Lossy	SAIs	✔	✔	HEVC, Linear Approximation and CNN	✖
Wang *et al.* [25]	Lossy	SAIs	✖	✔	HEVC & Multibranch Spatial Transformer Networks	✖
Komatsu *et al.* [26]	Lossy	SAIs	✔	✔	Binary images representation of the LF	✖
Chen *et al.* [27]	Lossy	SAIs	✖	✔	HEVC & Global Multiplane Representation	HEVC
Conti *et al.* [28]	Lossy	SAIs	✖	✔	HEVC & field-of-view scalability	✖
Zhao *et al.* [29]	Lossy	SAIs	✖	✔	MV-HEVC & Super Resolution CNN	MV-HEVC
Proposed	Lossy	SAIs	✖	✔	VVC, LF-D2GAN and MV-QENet	VVC

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 uni-predictive 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 hexadeca-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 *1* is encoded otherwise *0* 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 hexadeca-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 farther 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 maps[3] 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-

[3]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 encore 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 [26] 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 straight forward 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.* [20] 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 site 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068563, IEEE Transactions on Multimedia

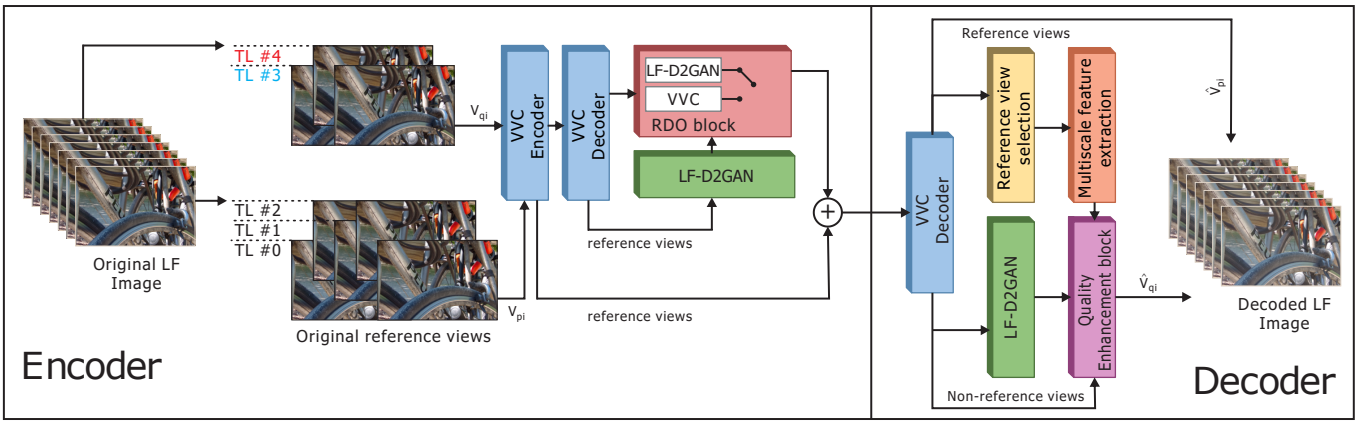


Fig. 5: Overall scheme of the proposed coding solution.

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068563, IEEE Transactions on Multimedia

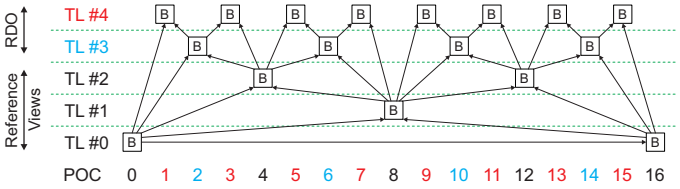


Fig. 6: Hierarchical prediction structure in VVC in RA coding configuration.

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 scalabilty 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 $\tilde{V}_{p_1}, \tilde{V}_{p_2}, \ldots, \tilde{V}_{p_{N_R}}$

$$\tilde{V}_q = f\left(\tilde{V}_{p_1}, \tilde{V}_{p_2}, \ldots, \tilde{V}_{p_{N_R}}, q\right), \tag{2}$$

where $p_1, p_2 \cdots, p_{N_R}$ 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

$$D_q = g_d\left(K\right) \tag{3}$$

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

$$\bar{V}_{p_i}(s) = \tilde{V}_{p_i}\left[s + (p_i - q)D_q(s)\right], \tag{4}$$

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

The N_R wrapped reference views $\bar{V}_{p_1}, \bar{V}_{p_2}, \ldots, \bar{V}_{p_{N_R}}$ 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 $\bar{V}_{p_1}, \bar{V}_{p_2}, \ldots, \bar{V}_{p_{N_R}}$, its disparity map D_q estimated by the disparity CNN and its position q.

$$\tilde{V}_q = g_c\left(\bar{V}_{p_1}, \ldots, \bar{V}_{p_{N_R}}, D_q, q\right). \tag{5}$$

As mentioned in Section I, the proposed coding approach is based on the LF-D2GAN block. GANs are deep neural net 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 ($\mathbb{P}_{train}$) and returns low score for data generated by the generator ($\mathbb{P}_G$). 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

$$\begin{aligned}\min_{\theta_G} \max_{\theta_{D1}, \theta_{D2}} \mathcal{L}\left(\theta_G, \theta_{D1}, \theta_{D2}\right) &= \alpha\, \mathbb{E}_{x \backsim \mathbb{P}_{train}}[\log D_1(x)] \\ &+ \mathbb{E}_{\tilde{x} \backsim \mathbb{P}_G}[-D_1(G(\tilde{x}))] + \mathbb{E}_{x \backsim \mathbb{P}_{train}}[-D_2(x)] \\ &+ \beta\, \mathbb{E}_{\tilde{x} \backsim \mathbb{P}_G}[\log D_2(G(\tilde{x}))],\end{aligned} \tag{6}$$

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068563, IEEE Transactions on Multimedia

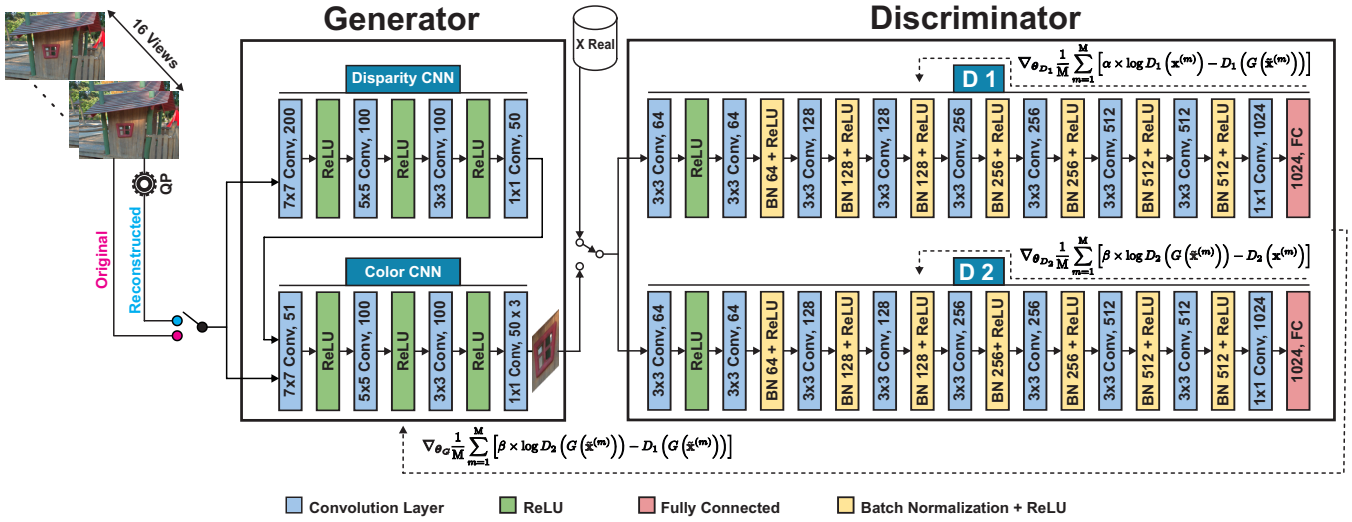


Fig. 7: Architecture of the proposed LF dual discriminator generative adversarial network (LF-D2GAN).

the weights of the discriminators D_1 and D_2, respectively, by ascending the obtained stochastic gradient.

$$\nabla_{\theta_{D1}} \frac{1}{M} \sum_{m=1}^{M} \alpha \log D_1(x^{(m)}) - D_1(G(\tilde{x}^{(m)})), \quad (7)$$

$$\nabla_{\theta_{D2}} \frac{1}{M} \sum_{m=1}^{M} \beta \log D_2(G(\tilde{x}^{(m)})) - D_2(x^{(m)}), \quad (8)$$

while (9) represents the cost function to gain the error that should be given to the generator G for its weight updating.

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

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 of temporal levels 0, 1 and 2 are encoded using the VVC codec, which constitute the reference views used in the synthesis process. Then, 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 a non-reference view should be encoded or synthesized at the decoder side. The proposed RDO process is described in the Algorithm 1 and explained in the following.

As illustrated in Fig. 6, we apply RDO process on the 3 consecutive frames, i.e., frame i at level 4, frame $i+1$ at level 3 and frame $i+2$ at level 4. It should be noted that if one of the views at temporal level 4 (frame i or $i+2$) is encoded using VVC, then the frame $i+1$ at level 3 must also be encoded using VVC, because this layer will be used as a reference for the frames at temporal level 4. The main reasons behind only considering the 2 upper levels exclusively to the RDO block

Algorithm 1 RDO block based Lagrangian optimization

Require: $\mathcal{J} \leftarrow \{\forall\ v \in TL\#[3\ or\ 4],\ \forall\ m \in \{VVC, LF-D2GAN\}, \mathcal{J} = D + \lambda R\}$
for all $v \in TL\#4$ **do**
  if $\mathcal{J}(VVC) < \mathcal{J}(LF\text{–}D2GAN)$ **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 $\mathcal{J}(VVC) > \mathcal{J}(LF\text{–}D2GAN)$ **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.

$$\mathcal{J} = D + \lambda R, \quad (10)$$

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068563, IEEE Transactions on Multimedia

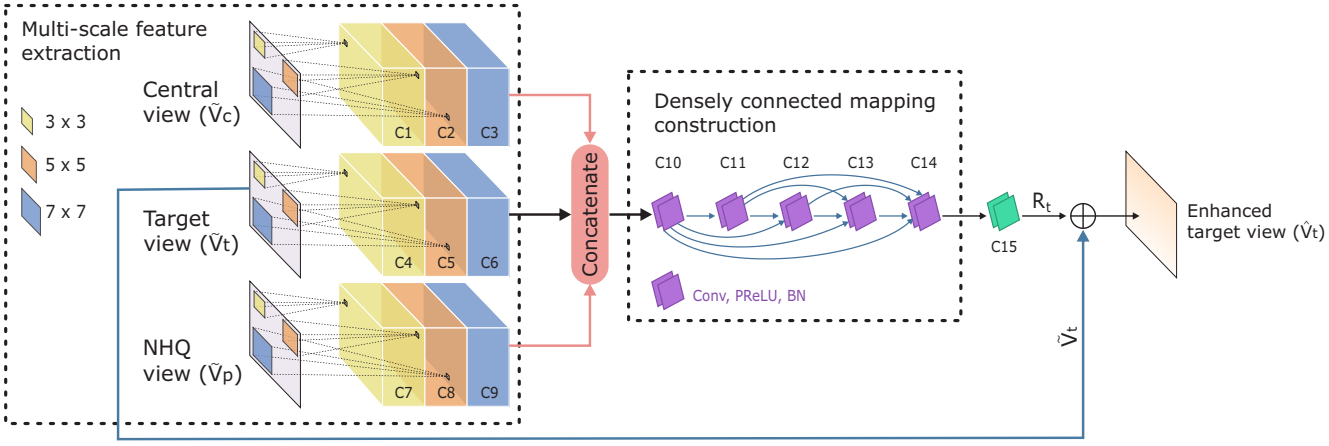


Fig. 8: Detailed architecture of the proposed multi-view quality enhancement convolutional neural Network (MV-QENet).

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 is sent to the LF-D2GAN block. This also has the advantage to make the bitstream compliant with the VVC standard.

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-reference 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. 5. Here, the concept of quality enhancement consists in predicting the residual errors $R_q = V_q - \tilde{V}_q$ 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 $\tilde{V}_t$, the decoded central view $\tilde{V}_c$ (which is of the highest quality as it is encoded in Intra using low QP value) and one neighbor decoded reference view $\tilde{V}_p$. $\tilde{V}_p$ 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 $\tilde{V}_p$ view with the highest quality score computed by CORNIA metric. The 3 views ($\tilde{V}_t$, $\tilde{V}_c$, $\tilde{V}_p$) 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

$$\hat{V}_t = \tilde{V}_t + h_\phi(\tilde{V}_t, \tilde{V}_c, \tilde{V}_p), \tag{11}$$

with h_ϕ is the parametric function of quality enhancement neural network and ϕ its trainable parameters.

TABLE II: Convolutional layers of MV-QENet block.

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

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 ($\tilde{V}_c$, $\tilde{V}_p$) and one target non-reference view $\tilde{V}_t$. 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 3×3. 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 $H_l(.)$, including Convolution (Conv), PReLU and BN. We further

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068563, IEEE Transactions on Multimedia

denote the output of the l-th layer as x_l, such that each layer can be formulated as follows

$$\begin{aligned} x_{11} &= H_{11}([x_{10}]), \\ x_{12} &= H_{12}([x_{10}, x_{11}]), \\ x_{13} &= H_{13}([x_{10}, x_{11}, x_{12}]), \\ x_{14} &= H_{14}([x_{10}, x_{11}, x_{12}, x_{13}]), \end{aligned} \tag{12}$$

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 $\tilde{V}_t$

$$\hat{V}_t = \tilde{V}_t + R_t\left(\theta_{qe}\right), \tag{13}$$

with θ_{qe} is defined as the trainable parameters of the MV-QENet. The MV-QENet is trained with minimizing a mean squared error loss function

$$loss = \|V_t - \hat{V}_t\|_2^2. \tag{14}$$

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

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 60×60, 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 60×60, stride of 16, and output patch equal to 36×36 (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 64×64 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 II.

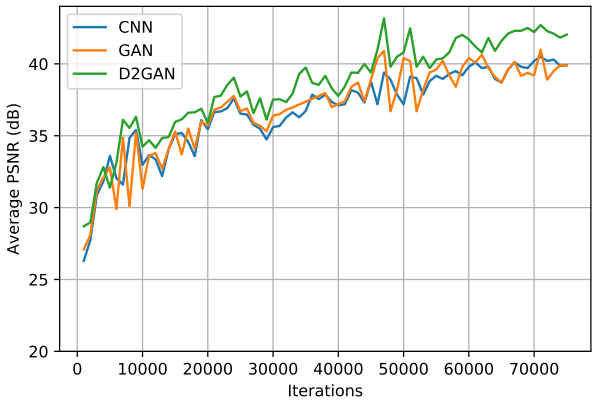


Fig. 9: Average PSNR performance during the training iterations on the validation set for CNN, GAN and the proposed LF non-reference views synthesis (LF-D2GAN) architectures.

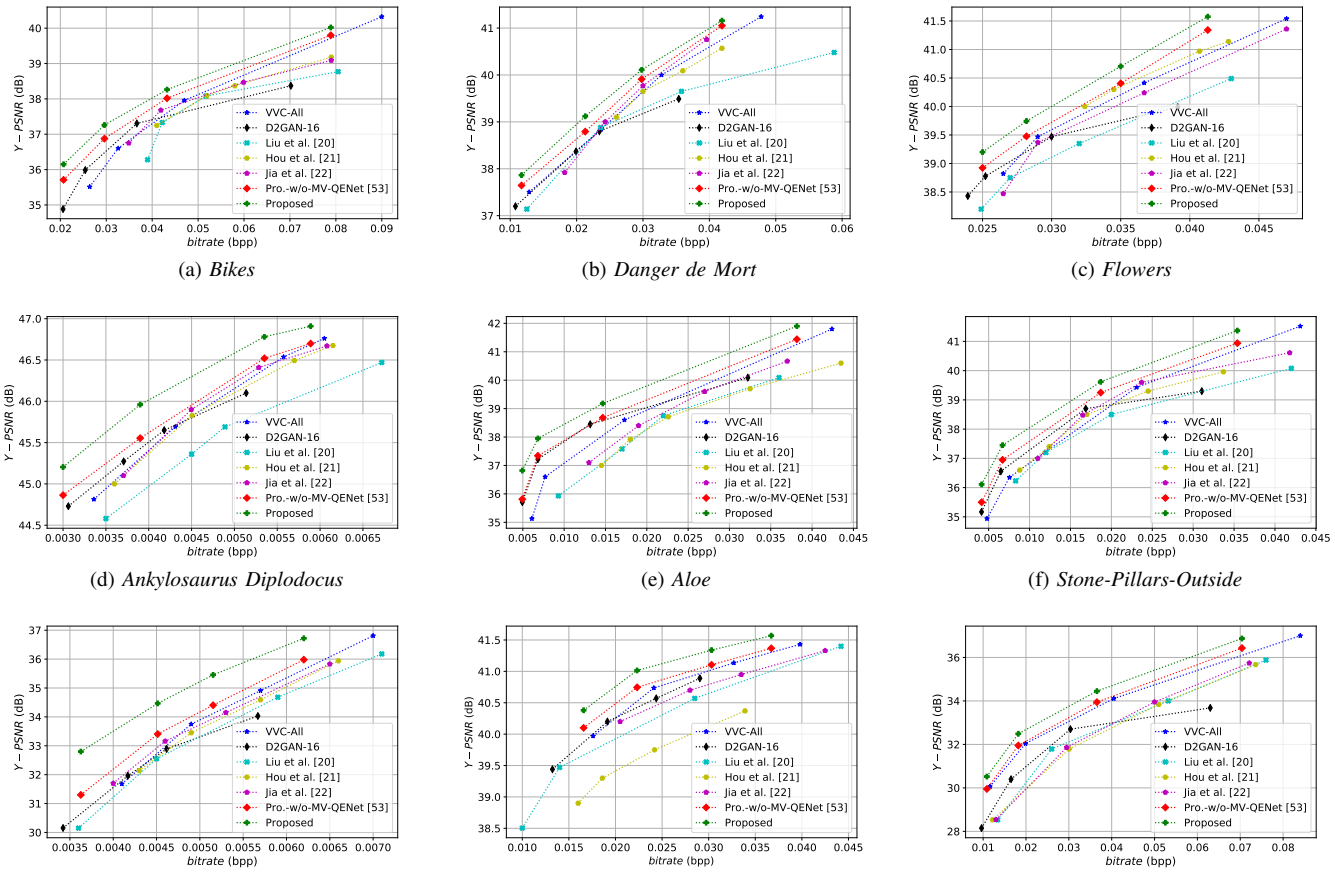
3) Testing conditions: for the testing phase, 9 LLF images different from the training and validation sets are selected, 6 LLF images are from EPFL dataset [49], 1 LLF image from Stanford Lytro LF dataset [50] and 2 LLF images from HCI dataset [51]. Each of these LLF images is composed of 8×8 subaperture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



(a) *Bikes* (b) *Danger de Mort* (c) *Flowers*

(d) *Ankylosaurus Diplodocus* (e) *Aloe* (f) *Stone-Pillars-Outside*

(g) *Bedroom* (h) *Desktop* (i) *Herbs*

Fig. 10: RD curves of the five considered solutions for the 9 LF images using four QP values.

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 solution performs better than VVC-All, LF-D2GAN-16 and prop.-w/o-MV-QENet solutions.

Table III 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.

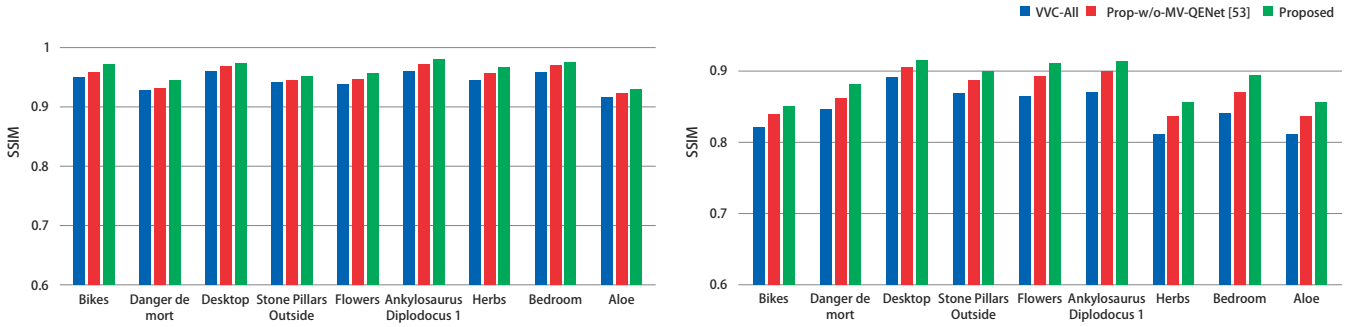
Fig. 11 gives the SSIM performance of the VVC-All, prop.-w/o-MV-QENet and proposed methods for the 9 test LF images at two QP values 18 and 32. We can notice from this figure that our solution gives the highest SSIM scores for all LF images at both considered QPs. Table IV shows the SSIM-based BD-BR 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 birate 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068563, IEEE Transactions on Multimedia

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	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	–15.6%	0.59	–8.3%	0.35	–0.54%	0.15	–24.1%	0.83	**–36.22%**	**1.35**



(a) QP 18 (b) QP 32

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*.

Im.	VVC-All		Prop.-w/o-MV-QENet [53]		Proposed	
	BD-BR	BD-SSIM	BD-BR	BD-SSIM	BD-BR	BD-SSIM
1)	–10.3%	0.011	–24.91%	0.022	**–40.1%**	**0.032**
2)	–12.4%	0.013	–22.10%	0.021	**–33.7%**	**0.027**
3)	–13.0%	0.020	–23.15%	0.028	**–32.8%**	**0.042**
4)	–16.01%	0.011	–22.8%	0.024	**–40.0%**	**0.035**
5)	–30.0%	0.028	–50.30%	0.046	**–66.6%**	**0.049**
6)	–19.80%	0.004	–40.49%	0.021	**–54.0%**	**0.031**
7)	–9.4%	0.010	–17.22%	0.015	**–30.6%**	**0.024**
8)	–18.0%	0.011	–37.09%	0.029	**–42.5%**	**0.036**
9)	–19.8%	0.020	–32.10%	0.041	**–46.1%**	**0.059**
Av.	–16.54%	0.014	–30.01%	0.027	**–42.95%**	**0.037**

TABLE V: Processing time of four LF image coding methods.

QP	Encoder time in seconds				
	VVC-All	Jia *et al.* [22]	Liu *et al.* [20]	Our	
	CPU	GPU	CPU	CPU	GPU
18	**259**	450	3535	559	514
22	**152**	350	3030	452	402
28	**101**	220	2478	401	349
34	**66**	142	1710	366	315
Average	**144**	291	2688	445	395
	Decoder time in seconds				
Average	**4**	53	5	333	285

at the edges.

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 NVDIA 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 encoding and decoding times in second for our solution and three other methods including VVC-All, Jia *et al.* [22] and Liu *et al.* [20] methods. The complexity of the proposed encoder is in the same range as the complexity of the solution proposed in [22] that also relies on a GAN to synthesize the non-reference views at the encoder. We can also notice that the GPU enables to speedup the LF-D2GAN and MV-QENet blocks at both encoder and decoder. The complexity of the proposed encoder is in average 6× faster than the encoder proposed in [20]. This latter relies on the JEM codec which is more complex than the VTM codec. However, the proposed decoder is more complex than the other decoders. On average, the VVC decoding takes 3 seconds, view synthesis using LF-D2GAN 92 seconds and finally MV-QENet block 190 seconds on GPU, which corresponds to 1.05%, 32.28% and 66,66% of the total decoder time, respectively. This clearly shows that the increase in the complexity of decoder is mainly due to the synthesis block and in particular to the quality enhancement block.

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3068563, IEEE Transactions on Multimedia

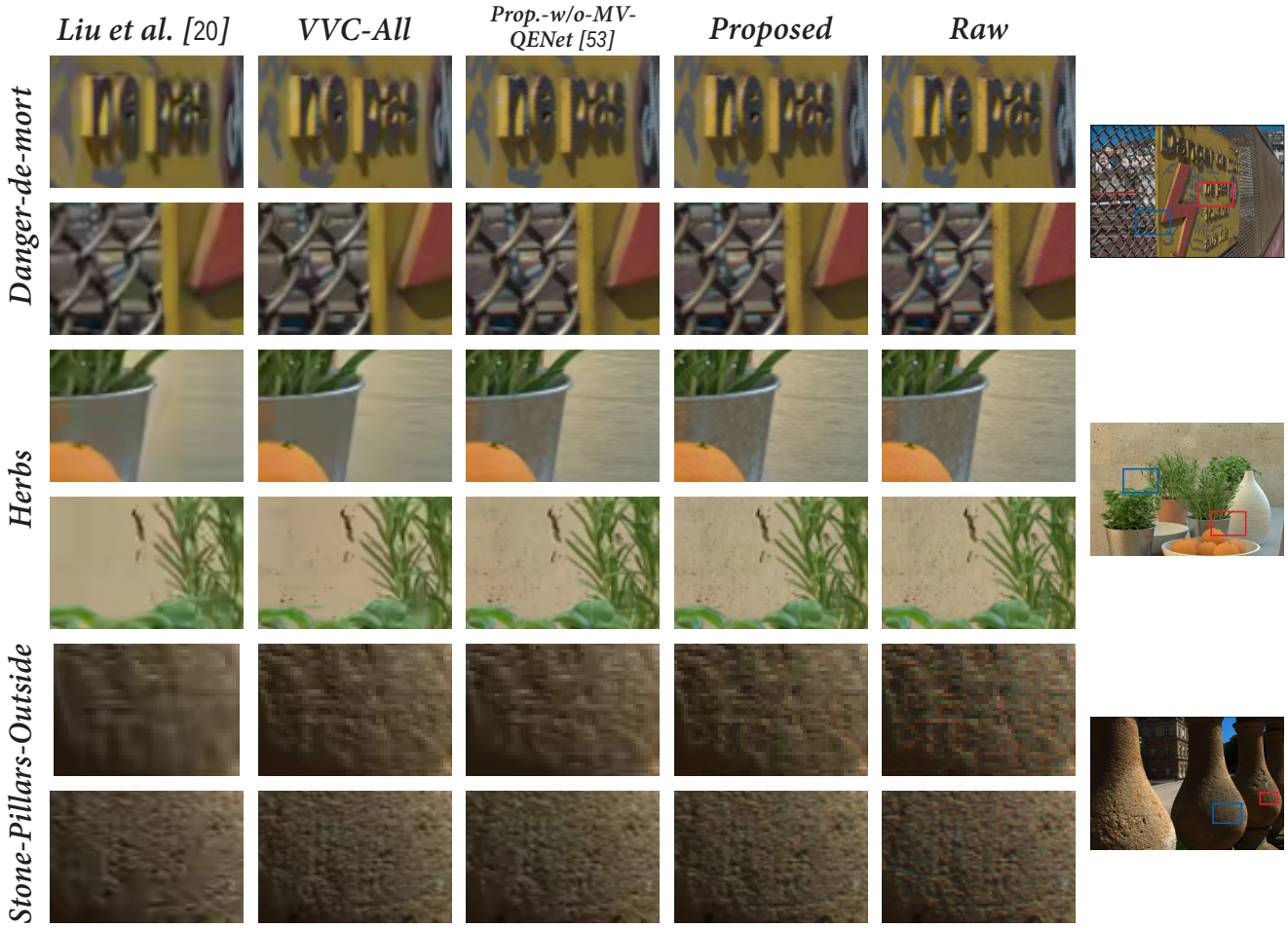


Fig. 12: Visual illustration of 3 test LF images encoded at around 0.014 bpp for *Danger de Mort*, 0.02 bpp for *Herbs* and 0.01 bpp for *Stone-Pillars-Outside*. The objective quality scores are provided for each illustrated view in this format PSNR(SSIM). *Danger de Mort* [view 49 (TL #4), RDO: Synthesized 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 stereoscope: Immersive computer graphics via factored near-eye light field displays with focus cues," *ACM Trans. Graph.*, vol. 34, no. 4, Jul. 2015.

[4] J. Chen, J. Hou, and L. Chau, "Accurate light field depth estimation with superpixel regularization over partially occluded regions," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 4889–4900, 2018.

[5] H. Jeon *et al.*, "Accurate depth map estimation from a lenslet light field camera," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1547–1555.

[6] C. Galea and C. Guillemot, "Denoising of 3d point clouds constructed from light fields," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 1882–1886.

[7] C. Kim *et al.*, "Scene reconstruction from high spatio-angular resolution light fields," *ACM Trans. Graph.*, vol. 32, no. 4, Jul. 2013.

[8] J. Fiss, B. Curless, and R. Szeliski, "Refocusing plenoptic images using depth-adaptive splatting," in *IEEE International Conference on Computational Photography (ICCP)*, 2014, pp. 1–9.

[9] V. Irene, M. Hermina Petric, F. Pascal, and E. Touradj, "A graph learning approach for light field image compression," in *SPIE 10752, Applications of Digital Image Processing XLI*, September 2018.

[10] M. B. de Carvalho *et al.*, "A 4d dct-based lenslet light field codec," in *IEEE International Conference on Image Processing (ICIP)*, 2018.

[11] P. Astola and I. Tabus, "Wasp: Hierarchical warping, merging, and sparse prediction for light field image compression," in *2018 7th European Workshop on Visual Information Processing (EUVIP)*, 2018, pp. 1–6.

[12] R. A. Farrugia and J. A. Briffa, "Lossless light field compression using 4d wavelet transforms," in *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 121–125.

[13] W. Ahmad, R. Olsson, and M. Sjostrom, "Interpreting plenoptic images as multi-view sequences for improved compression," in *IEEE International Conference on Image Processing (ICIP)*, Sep. 2017.

[14] L. Li, Z. Li, B. Li, D. Liu, and H. Li, "Pseudo-sequence-based 2-d hierarchical coding structure for light-field image compression," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1107–1119, 2017.

[15] C. Conti, P. Nunes, and L. Ducla Soares, "Light field image coding with jointly estimated self-similarity bi-prediction," *Signal Processing: Image Communication*, vol. 60, pp. 144 – 159, 2018.

[16] D. Liu, P. An, R. Ma, W. Zhan, X. Huang, and A. A. Yahya, "Content-based light field image compression method with gaussian process regression," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 846–859, 2020.

[17] X. Jiang, M. L. Pendu, R. A. Farrugia, and C. Guillemot, "Light field compression with homography-based low-rank approximation," *IEEE J. Sel. Topics Signal Process*, vol. 11, no. 7, pp. 1132–1145, Oct 2017.

[18] E. Dib, M. Le Pendu, X. Jiang, and C. Guillemot, "Super-ray based low rank approximation for light field compression," in *2019 Data Compression Conference (DCC)*, 2019, pp. 369–378.

[19] S. Zhao, Z. Chen, K. Yang, and H. Huang, "Light field image coding with hybrid scan order," in *Visual Communications and Image Processing (VCIP)*, 2016, pp. 1–4.

[20] D. Liu, L. Wang, L. Li, Z. X., F. W., and W. Z., "Pseudo-sequence-based light field image compression," in *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, July 2016, pp. 1–4.

[21] J. Hou, J. Chen, and L. Chau, "Light field image compression based on bi-level view compensation with rate-distortion optimization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 517–530, 2019.

[22] C. Jia, X. Zhang, S. Wang, S. Wang, and S. Ma, "Light field image compression using generative adversarial network-based view synthesis," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 177–189, 2018.

[23] S. Zhao and Z. Chen, "Light field image coding via linear approximation prior," in *IEEE International Conference on Image Processing (ICIP)*, Sep. 2017, pp. 4562–4566.

[24] N. Bakir, W. Hamidouche, O. Déforges, K. Samrouth, S. A. Fezza, and M. Khalil, "Rdo-based light field image coding using convolutional neural networks and linear approximation," in *Data Compression Conference (DCC)*, 2019, pp. 554–554.

[25] J. Wang, Q. Wang, R. Xiong, Q. Zhu, and B. Yin, "Light field image compression using multi-branch spatial transformer networks based view synthesis," in *Data Compression Conference (DCC)*, 2020, pp. 397–397.

[26] K. Komatsu, K. Takahashi, and T. Fujii, "Scalable light field coding using weighted binary images," in *IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 903–907.

[27] Y. Chen, P. An, X. Huang, C. Yang, D. Liu, and Q. Wu, "Light field compression using global multiplane representation and two-step prediction," *IEEE Signal Process. Lett.*, vol. 27, pp. 1135–1139, 2020.

[28] C. Conti, L. D. Soares, and P. Nunes, "Light field coding with field-of-view scalability and exemplar-based interlayer prediction," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 2905–2920, 2018.

[29] J. Zhao, P. An, X. Huang, C. Yang, and L. Shen, "Light field image compression via cnn-based epi super-resolution and decoder-side quality enhancement," *IEEE Access*, vol. 7, pp. 135 982–135 998, 2019.

[30] C. Brites, J. Ascenso, and F. Pereira, "Lenslet light field image coding: Classifying, reviewing and evaluating," *IEEE Trans. Circuits Syst. Video Technol.*, 2020.

[31] C. Conti, L. D. Soares, and P. Nunes, "Dense light field coding: A survey," *IEEE Access*, vol. 8, pp. 49 244–49 284, 2020.

[32] D. Taubman and M. Marcellin, *JPEG2000 Image Compression Fundamentals, Standards and Practice*. Springer Publishing Company, Incorporated, 2013.

[33] M. M. Hannuksela, Y. Yan, X. Huang, and H. Li, "Overview of the multiview high efficiency video coding (mv-hevc) standard," in *IEEE International Conference on Image Processing (ICIP)*, 2015.

[34] W. Ahmad, M. Ghafoor, S. A. Tariq, A. Hassan, M. Sjöström, and R. Olsson, "Computationally efficient light field image compression using a multiview hevc framework," *IEEE Access*, vol. 7, 2019.

[35] M. Wien, V. Baroncini, J. Boyce, C. Segall, and T. Suzuki, "Preliminary Joint Call for Evidence on Video Compression with Capability beyond HEVC," Geneva, Switzerland 2017.

[36] N. Sidaty, W. Hamidouche, O. Déforges, P. Philippe, and J. Fournier, "Compression Performance of the Versatile Video Coding: HD and UHD Visual Quality Monitoring," in *Picture Coding Symposium (PCS)*, 2019.

[37] J. C. Y. Ye and S.-H. Kim, "Algorithm description for Versatile Video Coding and Test Model 5 (VTM 5)," in *document JVET-N1002, 14th JVET meeting*, Gothenburg, SE, Mar. 2019.

[38] X. Zhao *et al.*, "Joint Separable and Non-Separable Transforms for Next-Generation Video Coding," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2514–2525, May 2018.

[39] X. Zhao *et al.*, "NSST: Non-separable secondary transforms for next generation video coding," in *Picture Coding Symposium (PCS)*, 2016.

[40] G. Chaurasia, O. Sorkine, and G. Drettakis, "Silhouette-aware warping for image-based rendering," in *Proceedings of the Twenty-Second Eurographics Conference on Rendering*, ser. EGSR '11. Goslar, DEU: Eurographics Association, 2011, p. 1223–1232.

[41] G. Chaurasia, S. Duchene, O. Sorkine-Hornung, and G. Drettakis, "Depth synthesis and local warps for plausible image-based navigation," *ACM Trans. Graph.*, vol. 32, no. 3, Jul. 2013.

[42] N. Kalantari, T. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 193:1–193:10, Nov. 2016.

[43] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[44] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[45] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Sod-mtgan: Small object detection via multi-task generative adversarial network," in *The European Conference on Computer Vision (ECCV)*, 2018.

[46] T. Nguyen, T. Le, H. Vu, and D. Phung, "Dual discriminator generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2017, pp. 2670–2680.

[47] P. Ye, J. Kumar, L. Kang, and D. Doermann, "K-means sparse-coding random samples raw-image-patch hard-assignment soft-assignment," 2012.

[48] G. Huang *et al.*, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[49] M. Rerabek and T. Ebrahimi, "New light field image dataset," in *https://mmspg.epfl.ch/EPFL-light-field-image-dataset*, 2016.

[50] S. Raj, L. Michael, and A. Sunder, "Stanford lytro light field archive," in *http://lightfields.stanford.edu/*, 2016.

[51] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldlücke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Asian Conference on Computer Vision*. Springer, 2016.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[53] N. Bakir, W. Hamidouche, S. A. Fezza, K. Samrouth, and O. Déforges, "Light field image coding using dual discriminator generative adversarial network and vvc temporal scalability," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.

[54] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.