LIGHT FIELD IMAGE CODING USING DUAL DISCRIMINATOR GENERATIVE ADVERSARIAL NETWORK AND VVC TEMPORAL SCALABILITY

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

LIGHT FIELD IMAGE CODING USING DUAL DISCRIMINATOR GENERATIVE ADVERSARIAL NETWORK AND VVC TEMPORAL SCALABILITY

Nader Bakir[*†], *Wassim Hamidouche*[*], *Sid Ahmed Fezza*[§], *Khouloud Samrouth*[†] *and Olivier Déforges*[*]

[*]Univ. Rennes, INSA Rennes, CNRS, IETR - UMR 6164, Rennes, France
[†]Lebanese University, Tripoli, Lebanon
[§]National Institute of Telecommunications and ICT, Oran, Algeria
whamidou@insa-rennes.fr

ABSTRACT

Light field technology represents a viable path for providing a high-quality VR content. However, such an imaging system generates a high amount of data leading to an urgent need for LF image compression solution. In this paper, we propose an efficient LF image coding scheme based on view synthesis. Instead of transmitting all the LF views, only some of them are coded and transmitted, while the remaining views are dropped. The transmitted views are coded using Versatile Video Coding (VVC) and used as reference views to synthesize the missing views at decoder side. The dropped views are generated using the efficient dual discriminator GAN model. The selection of reference/dropped views is performed using a rate distortion optimization based on the VVC temporal scalability. Experimental results show that the proposed method provides high coding performance and overcomes the state-of-the-art LF image compression solutions.

Index Terms— Light Field, Deep Learning, D2GAN, VVC, Coding Structure, RDO.

1. INTRODUCTION

Light Field (LF) image can be captured and sampled by a plenoptic camera composed of an array of microlens such as *Lytro* and *Raytrix* cameras. LF imaging has many advantages over traditional 2D imaging systems, as its allows the user to change various camera settings after capture, thus providing more flexibility. Specifically, LF image captures both spatial and angular information enabling several multimedia services: multi-focus, multi-perspective, viewpoint rendering and even 6-Degree of Freedom (6-DoF) viewing [1, 2]. A light field can be described by a 4 dimensional function with 2 parallel plans s, t and u, v denoted by $L(u, v, s, t)$. There are several ways to represent a LF image, including micro-image, epipolar image and sub-aperture image [3]. The latter, illustrated in **Fig. 1**, is the the widely used representation.

However, in sight of the huge amount of data involved by LF image, its processing, storage and transmission raise a real challenge that received increased research attention. In response, several solutions have been proposed to encode a LF image in sub-aperture representation. The straight forward coding approach organizes the LF views in a pseudo video sequence, which is then encoded with a classical 2D hybrid video encoder [4, 5, 6]. 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 [7, 8, 9], for instance, linear approximation has been investigated in [7] to estimate the views at the decoder from neighbour views, while a combination of linear approximation and convolutional neural network has been proposed in [8] to synthesize missing views at the decoder side. In the same way, Jia *et al.* [9] proposed to use the generative adversarial network 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. Jiang *et al.* [10] proposed a coding method called Homography-based Low Rank Approximation. This method jointly optimizes multiple homographies that align the LF views and low rank approximation matrices. Hou *et al.* [11] 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.

In this paper, we propose an efficient approach to encode the LF images, which consists in encoding a sparse set of views, and estimate the rest of views at the decoder side. In particular, the first set of selected reference views are coded with the next generation video coding standard called Versatile Video Coding (VVC). While the second set of views are either synthesized from the first decoded set of views using a Dual Discriminator Generative Adversarial Network (D2GAN) or decoded by a VVC decoder. The D2GAN have been trained with a large set of LF images coded at different distortions. The architecture offered by the D2GAN, composed by a generator and two discriminators, enables better training and thus synthesizes views with high visual quality. In addition, to increase the coding efficiency, a rate distortion optimization (RDO) is adopted to select which views

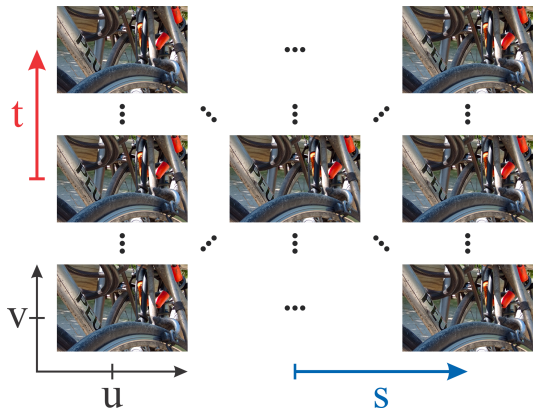

Fig. 1. Illustration of the Light Field image as an array of u, v slices arranged in s, t.

should be encoded and transmitted and which ones should be dropped and synthesized at the decoder side.

The remainder of this paper is organized as follows. Section 2 describes the concepts of D2GAN and VVC. Then, in Section 3, we describe the proposed LF image compression solution. Section 4 presents and discusses the experimental results. Finally, Section 5 concludes this paper.

2. BACKGROUND

As mentioned in Section 1, the proposed coding approach is based on Dual Discriminator Generative Adversarial Network (D2GAN) and Versatile Video Coding (VVC) standard. In this section, we briefly introduce these two concepts.

2.1. Dual Discriminator Generative Adversarial Nets

Generative Adversarial Networks (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 [12]. GAN has recently achieved great success in various fields, especially in fake video generation, super-resolution and objects detection [13, 14].

Dual discriminator generative adversarial network (D2GAN), is a novel framework based on GAN, which uses two discriminators D_1 and D_2, where D_1 tries to assign high scores for real data, and D_2 tries to assign high scores for the fake data. This technique uses the two discriminators to minimize the Kullback-Leibler (KL) divergence and reverse KL between the generated image and the target image [15].

Formally, D_1, D_2 and G now play the following three player minimax optimization game

$$\begin{aligned}\min_{G}\max_{D1,D2} \jmath\,(G, D1, D2) &= \alpha\,\mathbb{E}_{x\backsim P_{data}}[\log D_1(x)] \\ &+ \mathbb{E}_{z\backsim P_z}[-D_1(G(z))] + \mathbb{E}_{x\backsim P_{data}}[-D_2(x)] \\ &+ \beta\,\mathbb{E}_{z\backsim P_z}[\log D_2(G(z))],\end{aligned} \tag{1}$$

where z is a noise vector, $\mathbb{E}$ represents expected value, x is the real data, 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 KL and reverse KL divergences on the optimization problem [15].

More specifically, with a batch of M noise samples $z^{(1)}, z^{(2)}, ..., z^{(M)}$ given as inputs, the generator generates M artificial samples, and this process is defined as $G(z^{(i)})$. While, $x^{(1)}, x^{(2)}, ..., x^{(M)}$ represents a batch of M real data samples.

Three cost functions defined in (2), (3) and (4) 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. 2** (dash lines).

$$\nabla_{\theta_{D1}}\frac{1}{M}\sum_{m=1}^{M}[\alpha\,\log D_1(x^{(m)}) - D_1(G(z^{(m)}))], \tag{2}$$

$$\nabla_{\theta_{D2}}\frac{1}{M}\sum_{m=1}^{M}[\beta\,\log D_2(G(z^{(m)})) - D_2(x^{(m)})], \tag{3}$$

$$\nabla_{\theta_{G}}\frac{1}{M}\sum_{m=1}^{M}[\beta\,\log D_2(G(z^{(m)})) - D_1(G(z^{(m)}))]. \tag{4}$$

In this work, we use D2GAN to synthesize the dropped LF views, where the generator consists of two conventional neural networks (CNN) [16]. The first CNN estimates the disparity and the second one generates the color image.

2.2. Versatile Video Coding

Based on High Efficiency Video Coding (HEVC), the Joint Video Exploration Team (JVET) is developing a new video coding standard called Versatile Video Coding (VVC) [17]. VVC already enables a bitrate saving of 35% to 40% with respect to HEVC for the same visual quality [18]. VVC introduces several new coding tools at different levels of the coding chain including frame partitioning, Intra/Inter predictions, transform, quantization, entropy coding and in-loop filters. For more details about the VVC coding tools the reader can refer to [19]. VVC supports by design the temporal scalability through the Random Access (RA) coding configuration. This latter, illustrated in **Fig. 3**, 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 removed without impacting the decoding of frames of lower temporal resolution t_j with $t_i > t_j$.

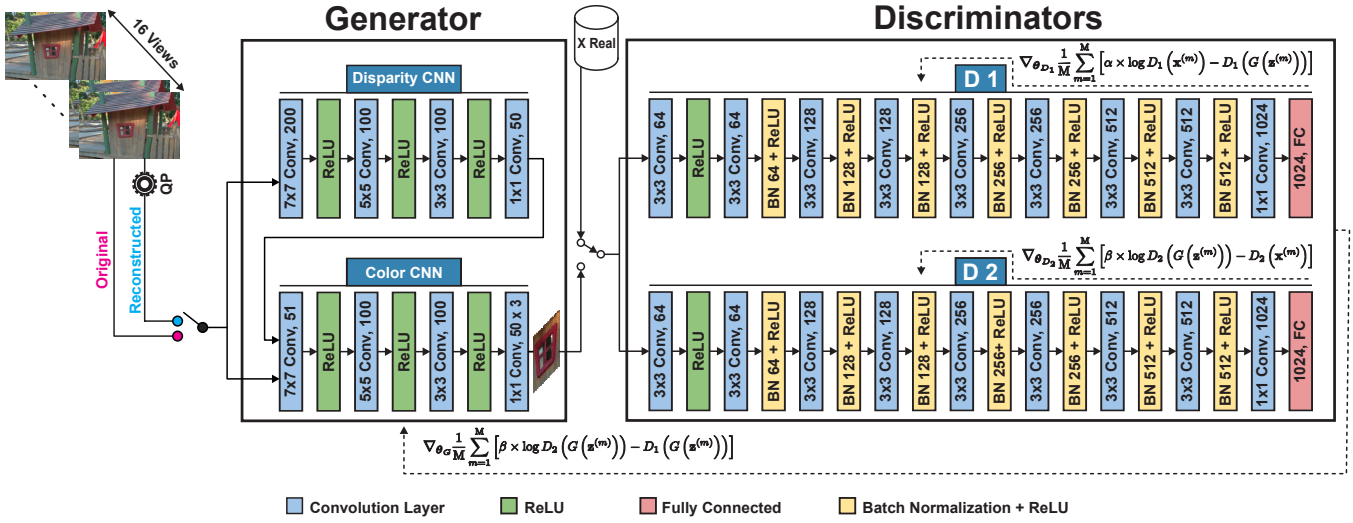

Fig. 2. Dual discriminator generative adversarial networks (D2GAN) architecture.

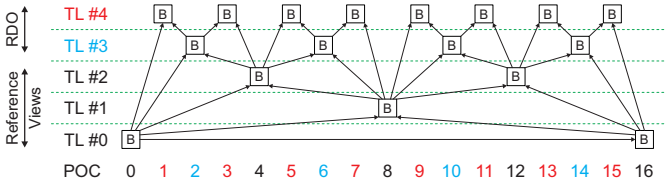

Fig. 3. Hierarchical prediction structure in VVC in Random Access (RA) coding configuration.

In the proposed coding approach, we exploit the concept of temporal resolution to drop views at the encoder without impacting the decoding process and thus performing the best rate distortion performance.

3. PROPOSED LF IMAGE COMPRESSION METHOD

The idea behind the proposed coding method is, instead of transmitting all the LF views, to drop a sub-set of views at the encoder side and synthesize them at decoder side, thus considerably reducing the required bitrate for LF images. To efficiently achieve that, we exploit the temporal scalability of VVC and use the D2GAN model, all within a rate distortion optimization process (RDO).

At the encoder side, first, LF sub-aperture views are organized into groups of 16 views that form Groups of Pictures (GOPs), as illustrated in **Fig. 3**. 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 later in the synthesis process at the decoder side. Then, the images at the remaining levels 3 and 4 are either coded using the VVC codec or dropped. In contrast to fix the number of dropped views, in our approach this is done adaptively on the basis of the proposed RDO process described in the Algorithm 1 and

Algorithm 1 RDO block based Lagrangian optimization

Require: $\mathcal{J} \leftarrow \{\forall\, m,\ \forall\, v \in TL\#[3\ or\ 4],\ \mathcal{J} = D + \lambda R\}$
m: metod {VVC, D2GAN}
for all $v \in TL\#4$ **do**
  if $\mathcal{J}(VVC) < \mathcal{J}(D2GAN)$ **then**
    Encode v by VVC
    flag(v) $\leftarrow$ False
  else
    generate v by D2GAN
    flag(v) $\leftarrow$ True
  end if
end for
for all $v \in TL\#3$ **do**
  if $\mathcal{J}(VVC) < \mathcal{J}(D2GAN)$ **then**
    Encode v by VVC
    flag(v) $\leftarrow$ False
  else {flag(previous(v)) and flag(next(v))}
    generate v by D2GAN
    flag(v) $\leftarrow$ True
  end if
end for

explaining in the following.

As illustrated in **Fig. 3**, 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 $i+2$) must be encoded using VVC, then the frame $i+1$ at level 3 is also encoded using VVC, because this layer will be used as a reference for the frames at temporal level 4.

Main reasons behind only considering the 2 upper levels exclusively to the RDO block are, firstly, after an extensive study, we found that these levels occupy together around 28%

Table 1. The average coding gains in terms of BD-BR of D2GAN, trained with reconstructed views, in comparison with the anchor D2GAN training with original views.

	wPSNR-based		SSIM-based	
	BD-BR	BD-PSNR	BD-BR	BD-SSIM
vs. D2GAN Reconstructed	-11.0%	0.25	-20.3%	0.013
vs. D2GAN Recons. separately	**-16.6%**	**0.39**	**-25.5%**	**0.022**

of the total bitrate. Second, the views at the upper levels are not used as references in the VVC coding scheme.

Thus, we proposed a RDO block deciding which views from the upper level can be encoded using VVC or dropped and synthesized using D2GAN. To reach this goal, the encoder computes the rate distortion (RD) cost function J given by (5) for both the VVC decoded view and the one synthesized by the D2GAN.

$$\mathcal{J} = D + \lambda R \tag{5}$$

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 optimal and for which the Lagrangian optimization is giving the best performance.

At the decoder side, the dropped views are synthesized using D2GAN block. As a reminder, the D2GAN is composed of a generator G and two discriminators D_1 and D_2. G consists of two CNNs [16], the first CNN estimates the disparity and the second one generates the color image. A set of features (mean and standard deviation) of a sparse set of views (16 views) are fed to the disparity CNN that estimates the disparity at an intermediate view, and then used it to warp (backward) all the input views to the intermediate view. The second color CNN uses all the warped images, derived from the first CNN, to predict the color and synthesizes the dropped views.

Given that the generator G and discriminators ($D1$ and $D2$) are CNN-based blocks, a training phase is required to fix respectively their parameters θ_G, θ_{D1} and θ_{D2}. Unlike GAN, in D2GAN, the scores returned by G are values in $\mathbb{R}^+$ rather than probabilities in $[0, 1]$. The discriminators and generator are alternatively updated using stochastic gradient ascent and descent, respectively. The backward propagation of errors (*i.e.*, cost functions) is applied to update the discriminators and generator with mini-batch size equal to M, as shown in **Fig. 2**.

For the training phase of D2GAN, 3 configurations were considered : 1) training with the original views , 2) training with reconstructed views at multiple distortion levels including the original views and 3) one training for each distortion level separately. We compared the three configurations, and the obtained results are given in **Table 3**. Based on these results, the third configuration, *i.e.*, D2GAN reconstructed separately, outperforms the other configurations and hence we used it for the D2GAN training.

4. RESULTS AND DISCUSSIONS

4.1. Experimental setup

The proposed deep learning-based architecture described in the previous section was trained with 140 LF images, where 70 LF images are from EPFL dataset [20], 50 LF images are from Stanford Lytro LF image dataset [21] and 20 LF images are from HCI dataset [22]. Each sub-aperture view was split into patches of size 60×60, thus resulting in more than 150,000 patches that were used in the training phase. For the testing phase, 9 LF images are selected, 6 LF images are from EPFL dataset [20], 1 LF image from Stanford Lytro LF dataset [21] and 2 LF images from HCI dataset [22]. Each of these LF images is composed of 8×8 sub-aperture views. These views are rearranged in a pseudo sequence using spiral order scan and coded using VVC in random access (RA) coding configuration at 4 QP values of 18, 24, 28 and 32.

The training configuration of D2GAN was set as follows: we trained the generator G and two discriminators (D_1 and D_2) with the ADAM optimizer 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 Figure 2. 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).

4.2. Evaluations

We compared the proposed scheme with four state-of-the-art methods: 1) VVC-All that encodes all views with the VVC in RA coding configuration, 2) LF-GAN method proposed in [9], where a sub-set of views are coded with HEVC, while the remaining views are generated by GAN and the residual error of views are transmitted to the decoder, 3) the method proposed in [4] encodes the views as a pseudo-video sequence using specific order scan, 4) the method of Hou *el al.* [11] that exploits the Inter- and Intra-views correlation to encode the views using HEVC. The latter method is considered as the anchor method.

4.3. Results

The BD-BR [23] is a Peak Signal to Noise Ratio (PSNR) based metric. It is used in this paper to assess the gain of the proposed approach compared to the anchor solution. A negative BD-BR value refers to a bitrate reduction compared to the anchor method, while a positive value expresses a bitrate overhead.

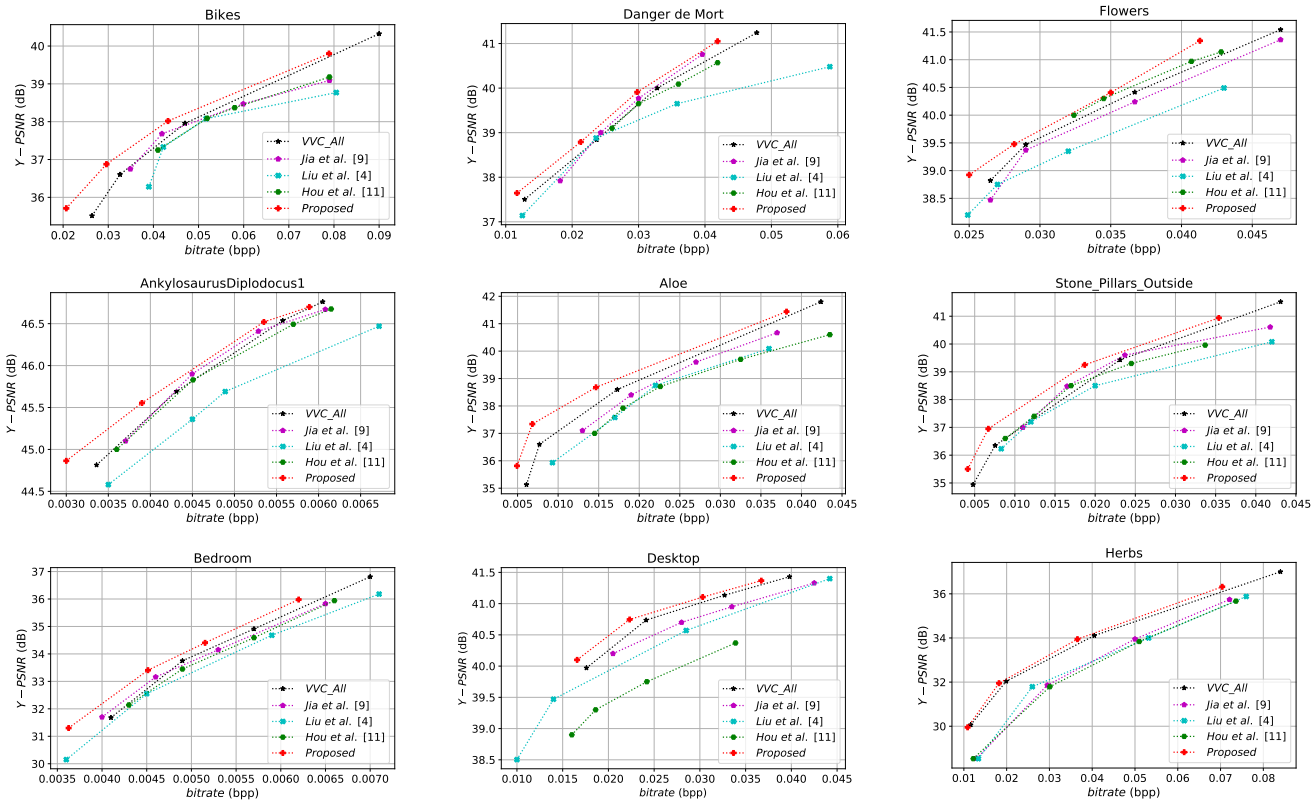

Fig. 4. RD curves of the five considered solutions for the 9 LF images using four QP values.

Table 2. BD-BR and BD-PSNR gains calculated against anchor method described in [4].

Image	VVC-All		Jia *et al.* [9]		Hou *et al.* [11]		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**
DangerDeMort	–7.8%	0.22	–10.8%	0.28	–8.7%	0.26	**–16.5%**	**0.40**
Flowers	–12.3%	0.56	–11.9%	0.54	–16.2%	0.72	**–16.6%**	**0.74**
Ankylosaurus_Dip1	–13.2%	0.44	–14.9%	–0.72	–12.3%	0.39	**–18.0%**	**0.57**
Aloe	–26.4%	0.85	–9.1%	0.31	–2.46%	–0.12	**–42.3%**	**1.23**
Stone_pillars_outside	–18.3%	0.61	–15.1%	0.52	–11.9%	0.28	**–35.6%**	**0.98**
Bedroom	–5.3%	0.46	–4.0%	0.32	–2.3%	0.18	**–9.5%**	**0.85**
Desktop	–19.6%	0.32	–7.5%	0.11	44.1%	–0.61	**–26.3%**	**0.45**
Herbs	–26.0%	1.14	–4.4%	–0.11	6.9%	–0.20	**–29.8%**	**1.32**
Average	–15.6%	0.59	–8.3%	0.35	–0.54%	0.15	**–24.1%**	**0.83**

R-D curves based on PSNR for the 9 LF images are provided in **Fig. 4**. We can notice that for all considered images, the proposed coding method provides the highest performance for all bitrates. The previous conclusion is confirmed by **Table 2**, providing the Bjøntegaard results of the four coding solutions compared to the anchor one [4]. The proposed method achieved an average BD-BR gain of -24.1% and BD-PSNR of 0.83 dB compared to the anchor method [4].

The complexity of the proposed coding approach is also evaluated and compared to the other methods on both CPU and GPU platforms. The performance has been carried-out on an Intel core i9-7900X CPU running at 3.3GHz PC with 64 GB memory and a TITAN Xp NVDIA GPU. It is important to note that the GPU is only used when the D2GAN block is involved in the coding scheme.

Table 3 gives the encoding and decoding run times in seconds. We can notice that the proposed solution requires almost the same complexity in the encoding for all QP compared to [9] and [11] methods. The GPU enables to speedup the encoding part related to the D2GAN block. However, the decoder of the proposed solution is more complex than the other solutions due the D2GAN block.

5. CONCLUSION

In this paper, we have proposed a view synthesis based LF image compression approach. In the proposed coding scheme,

Table 3. Processing time in seconds of the four LF image coding methods.

QP	Encoder				
	VVC-All	Jia *et al.* [9]	Hou *et al.* [11]	Our	
	CPU	GPU	CPU	CPU	GPU
18	**259**	450	6028	559	449
22	**152**	350	6028	452	342
28	**101**	220	6028	401	291
34	**66**	142	6028	366	256
Average	**66**	291	6028	445	335
	Decoder				
Average	**4**	53	583	124	94

a set of views are encoded using VCC, while the remaining views are dropped. The dropped views are synthesized using enhanced GAN-based approach known as D2GAN. The transmitted and dropped views are selected using RDO process. In addition, in order to avoid impacting the decoder with the dropped views, the latter are determined according to the temporal scalability of VCC. All these features allow reducing bitrate required by LF image, while providing views with high visual quality.

The experimental results showed the efficiency of our scheme, which achieved bitrate reduction of -24.1% in terms of BD-BR and increased the visual quality by 0.83 dB in BD-PSNR with respect to the state-of-the-art solution.

6. REFERENCES

[1] C. Guillemot and R. A. Farrugia, "Light Field Image Processing : Overview and Research Issues," *IEEE COMSOC MMTC Communications - Frontiers*, vol. 14, no. 4, pp. 37–43, 2017.

[2] G. Wu, B. Masia, A. Jarabo, Y. Zhang, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field image processing: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 926–954, Oct 2017.

[3] D. Dansereau, "Plenoptic signal processing for robust vision in field robotics," *Ph.D. dissertation, University of Sydney Graduate School of Engineering and IT School of Aerospace, Mechanical and Mechatronic Engineering*, 2014.

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

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

[6] 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, pp. 4557–4561.

[7] 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.

[8] 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 *2019 Data Compression Conference (DCC)*, March 2019, pp. 554–554.

[9] C. Jia, X. Zhang, S. Wang, S. Wang, S. Pu, 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*, pp. 1–1, 2018.

[10] X. Jiang, M. Le Pendu, R. A. Farrugia, and C. Guillemot, "Light field compression with homography-based low-rank approximation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1132–1145, Oct 2017.

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

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

[13] C. Ledig, L. Theis, F. Huszr, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 105–114.

[14] 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)*, September 2018.

[15] 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.

[16] 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.

[17] M. Wien, V. Baroncini, J. Boyce, A. Segall, and T. Suzuki, "Preliminary joint call for evidence on video compression with capability beyond hevc," Janvier 2017.

[18] 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)*, November 2019.

[19] K. Reuze, W. Hamidouche, P. Philippe, and O. Deforges, "Dynamic lists for efficient coding of intra prediction modes in the future video coding standard," in *2019 Data Compression Conference (DCC)*, March 2019, pp. 601–601.

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

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

[22] K. Honauer, O. Johannsen, D. Kondermann, and B. Goldlücke, "A dataset and evaluation methodology for depth estimation on 4d light fields," in *Computer Vision - ACCV 2016 : 13th Asian Conference on Computer Vision*, Cham, 2016, Springer.

[23] G. Bjøntegaard, "Improvements of the bd-psnr model," *ITU-T SG16 Q*, vol. 6, pp. 35, 2008.