Subjective Evaluation of Light Field Image Compression Methods based on View Synthesis

Nader Bakir*†, Sid Ahmed Fezza§, Wassim Hamidouche*, 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
firstname.lastname@insa-rennes.fr

***Abstract*—Light field (LF) images provide rich visual information enabling amazing applications, from post-capture image processing to immersive applications. However, this rich information requires significant storage and bandwidth capabilities thus urgently raises the question of their compression. Many studies have investigated the compression of LF images using both spatial and angular redundancies existing in the LF images. Recently, interesting LF compression approaches based on view synthesis technique have been proposed. In these approaches, only sparse samples of LF views are encoded and transmitted, while the other views are synthesized at decoder side. Different techniques have been proposed to synthesize the dropped views. In this paper, we describe subjective quality evaluation of two recent compression methods based on view synthesis and comparing them to two pseudo-video sequence based coding approaches. Results show that view synthesis based approaches provide higher visual quality than the naive LF coding approaches. In addition, the database as well as subjective scores are publicly available to help designing new objective metrics or can be used as a benchmark for future development of LF coding methods.**[1]

***Index Terms*—Light field, Image compression, View synthesis, Subjective evaluation, CNN, Linear approximation**

I. Introduction

Recent years have witnessed the rapid development of immersive multimedia systems, and light field (LF) imaging is considered as an attractive representation that can enable these immersive applications, such as virtual reality, 3D gaming, view synthesis and depth estimation. LF images provide both spatial and angular information, allowing more functionalities, such as free viewpoint change and refocusing. However, this providing rich information involves a large amount of data, making its storage or transmission not plausible over existing bandwidth-limited infrastructure. Therefore, efficient compression methods are of paramount importance.

Several compression approaches have been proposed in the literature [1, 2], and a new standard called JPEG PLENO (Part 2) is currently under development by the JPEG standardization committee, which aims at standardizing light field coding technologies and associated metadata [3]. Globally, depending on the coding format and acquisition process, there are two kinds of LF image compression approaches [4]. The first coding approach acts directly after the LF acquisition step on the raw LF data, *i.e.*, lenslet image, while the second approach exploits the 4D representation of the LF image that can be obtained by applying some specific transformations to the raw LF data.

For the methods that directly compress the lenslet images, most of them apply intra coding relying on redundancies existing in the images [5–7] or use a pseudo-temporal sequence suitable for standard 2D video encoders [8, 9]. Whereas the compression approaches based on 4D LF representation, consist of rearranging LF elements (usually sub-aperture views) in a specific order to produce a pseudo-video sequence, which is then encoded with a classical 2D video encoder (intra and inter predictions) [10–14] or using Multi-View extension of High Efficiency Video Coding (MV-HEVC) [15]. In addition, different reordering ways have been explored to construct the pseudo-video sequence including zig-zag, spiral, raster, rotation and line.

Moreover, by exploiting the redundancy existing between neighboring views, instead of encoding all the LF views, other approaches proposed to compress only a small subset of views, which are subsequently used to reconstruct the other views at the decoder side. To reconstruct theses non-coded views, different approaches have been proposed in the literature. For instance, in [16, 17], depth image-based rendering techniques have been exploited. Zhao *et al.* [18] proposed linear approximation prior, where the non-encoded views are approximated with a weighted sum of coded views. Recently, convolutional neural networks (CNN)-based approaches have been adopted to synthesis the non-coded views [19–21], allowing to recover the whole LF views.

All of these view synthesis based coding approaches have shown their coding efficiency in terms of rate-distortion performance compared to encoding the whole LF views. However, most of these works considered only objective metrics such as PSNR and SSIM (Structural Similarity Index) [22] to evaluate the coding performance, without taking into account the subjective quality assessment. Given that these conventional image quality metrics cannot handle efficiently the distortions that can be induced from imperfect view synthesis process, which are quite different from those introduced by image/video compression, the subjective quality assessment remains the best and inevitable way to perform the visual quality evalu-

[1] Dataset and code are available upon request.

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

ation of these LF coding approaches. In addition, the recent subjective studies conducted for LF image quality assessment did not consider these kinds of compression methods [4, 23], especially the CNN-based view synthesis approaches, which motivates us to supplement these studies.

Consequently, in this paper, we propose to conduct subjective experiments of LF compression methods based on view synthesis technique. Specifically, four compression approaches have been considered in this study, two methods are view synthesis basis, while the remaining are naive LF coding methods. All these methods have been subjectively and objectively evaluated. The dataset, including non-compressed and compressed LF images, along with subjective scores are provided publicly to facilitate future research works, such as developing new reliable objective quality metrics for LF images based view synthesis methods.

The rest of this paper is organized as follows. Section II presents the LF coding methods considered in this study. Section III describes the performed subjective experiment, including the preparation of the test material, environmental setup and the test methodology. Next, the results and analysis of subjective evaluation are provided in Section IV. Finally, Section V concludes the paper.

II. LIGHT FIELD CODING STRATEGIES

The LF contents evaluated in the subjective experiments were compressed using four coding strategies. Given that the widely explored coding approach for LF contents is the pseudo-video sequence coding method, we have therefore considered two methods from this category. For both coding methods, all the sub-aperture images are rearranged into a pseudo-sequence using spiral order scan starting from the center view, which is then encoded with a classical video encoder. Two video encoders have been selected for this purpose, the High Efficiency Video Coding (HEVC) standard and the Joint Exploration Test Model (JEM) that led to the starting point of future video coding standard named Versatile Video Coding (VVC). For HEVC, the HM reference software (version 16.9) was used, while for the second method the JEM software (version 7.0) was exploited, both in random access coding configuration. For both methods, all views are encoded and we refer to them as HM-All and JEM-ALL for the rest of this paper. In addition, in order to avoid the darkness and distorted remote views, only the middle 8×8 views were encoded.

Furthermore, two light field compression methods based view synthesis have been included in this study. Instead of coding all views, in these approaches, only sparse samples of LF views are encoded and transmitted, while the other views are synthesized at the decoder side. One of the selected methods is described in [18], where at the encoder side the views are equally divided into two sets, the selected reference views set and the dropped views set, that is 32 views each. The selected reference views are then rearranged into a pseudo-sequence using horizontal zigzag scan order and compressed with a 2D video encoder standard (JEM in our implementation). The decoded versions of theses latter views are used to linearly approximate the dropped views and only the approximation coefficients are transmitted to the decoder. At the decoder side, the selected reference views are decoded and the dropped views are approximated by the weighted sum of the decoded selected views. For the rest of this paper, we refer to this method as LA-32.

Finally, the fourth and last method that we included is the CNN-based view synthesis approach proposed in [19]. In this method, the authors proposed a learning-based approach to synthesize new views from a sparse set of input views. The proposed architecture includes two phases: a disparity estimator and color predictor, which are performed by two sequential CNNs. Based on the features extracted from the sparse input views (four views at the corners), four layers CNN firstly estimates the disparity of the dropped views. The second CNN uses all the warped disparity views, derived from the first CNN, along with few other features to predict the color and synthesize the dropped views. For training the CNN, we used 100 LF images, 28 from Stanford Lytro LF dataset [24] and 72 from California Lytro LF dataset [19]. We split each sub-aperture view into patches of size 60×60, which results in more than 100,000 patches exploited for training. For this method, which will be referred to as DL-16, 16 sparse views are encoded with the JEM, while the remaining dropped views are synthesized by the trained CNN block at the decoder side.

III. SUBJECTIVE EVALUATION

A. Dataset Preparation

A total of ten LF images have been carefully selected for subjective experiments, six from EPFL *Light-Field Image Dataset* (*Bikes*, *Fountain_&_Vincent_2*, *Friends-1*, *Overexposed-Sky*, *Rusty-Fence*, and *University*) [25], two from INRIA *Light-Field Image Dataset* (*Bee1* and *Cactus*) [26] and two that we acquired by a Lytro Illum camera, namely *Flowers* and *KidsHouse*. These LF images represent different content, including indoor and outdoor scenes and a wide range of colors, textures and depth properties [27]. In order to cover a wide range of features, the spatial complexity, color features and the amount of occluded pixels of each LF image have been analyzed using Spatial Information (SI) [28], ColorFulness (CF) [29] and occlusion model proposed in [30], respectively. The Figure 1 shows the values of SI, CF and occlusions for all the selected images.

Each image was extracted from LF raw file format using Light Field Matlab Toolbox v0.4 [31], thus providing a 4D LF of dimensions $15 \times 15 \times 434 \times 625 \times 4$, where 434×625 represents the resolution of each view, 4 corresponds to the RGB channels including additional weighting image component, while 15×15 represents the number of views [25]. As mentioned previously, we only encoded the central 8×8 sub-aperture views after being converted to YUV format and downsampled to 4:2:0 with 10-bit depth.

The ten LF images have been encoded using the previously described four compression methods at four compression

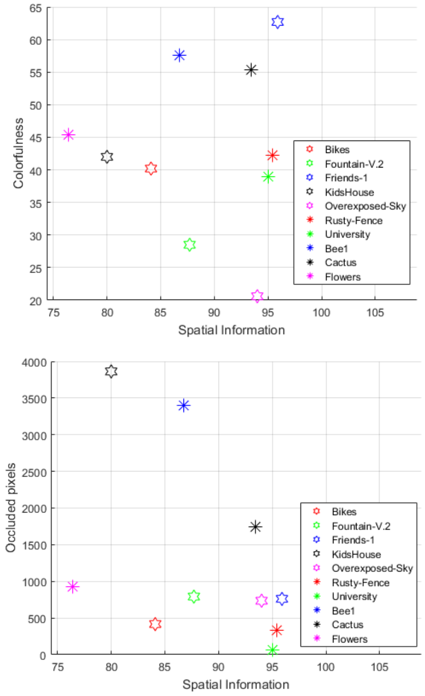

Fig. 1. Distributions of the three properties of the selected LF contents.

bitrates, namely *R1* = 0.0074 bpp, *R2* = 0.0171 bpp, *R3* = 0.0384 bpp and *R4* = 0.1112 bpp.

B. Environment Setup and Test Methodology

The subjective evaluations were conducted in a laboratory psychovisual test room, calibrated according to ITU-R BT.500-13 Recommendations [32], equipped with a controlled lighting system and the color of all background walls and curtains is mid-gray. A full HD 27-inch Dell UltraSharp U2717D was used to display the test stimuli. The distance of the subjects from the monitor was approximately equal to 7 times the image height, as recommended in [33].

The subjective experiments have been performed using the recently introduced methodology, named *passive test methodology* [4], without refocusing effect, which is out of the scope of this paper. The methodology is based on Double Stimulus Impairment Scale (DSIS) [32], where both the non-compressed reference and stimulus were displayed in a side-by-side arrangement on the same monitor. The non-compressed reference and stimulus were always displayed on the left and right side, respectively, and the subjects were aware of these positions. In addition, the LF contents were presented as a video sequence navigating between the viewpoints. The pseudo-video was created using horizontal scan, starting from the view in the left upper corner down, and proceeding from left to right and right to left in alternate order, which mimics the parallax effect. In [34], it has been noticed that this visualization technique is preferred among six possible different visualization strategies, because it reduces the shift among consecutive frames. Moreover, the created videos were displayed with a frame rate of 9 frames per second offering a smooth switching between views.

At the end of the presentation of each pair of videos, a dedicated user interface was displayed on the screen for about five seconds during which the subject gives its judgment. The participants were asked to rate the level of impairment of the stimulus with respect to the non-compressed reference, using a five-grade discrete impairment scale (1: very annoying, 2: annoying, 3: slightly annoying, 4: perceptible, but not annoying, 5: imperceptible).

Given the large number of stimuli, a session would exceed 30 minutes, making it hard to show all of them in a single session. Consequently, in order to avoid visual fatigue effects, the subjective experiment was divided into two sessions whose duration does not exceed 20 minutes each. Subjects took a break between each two sessions. Moreover, each test session involved only one subject assessing the stimuli. In order to avoid possible contextual and memory effects, the display order of these stimuli was randomized in a way that the same content was never shown consecutively.

Before the experiment starts, instructions explaining the task were provided to subjects. In addition, training session was held with additional LF contents, allowing the subjects to practice and become familiarized with the test procedure. The quality of these training samples was chosen so that it covers the full rating scale.

A total of 18 naive subjects (10 females and 8 males) took part in the subjective experiments. The age of subjects was ranging from 20 to 58, with an average of 29.4. All subjects were screened for color blindness and visual acuity using Ishihara and Snellen charts, respectively.

C. Data Processing

First, the subjective scores were screened to detect and exclude possible outliers. Outliers detection was performed as specified in [32], and no outlier subjects were found in this study.

Second, the Mean Opinion Score (MOS) was computed as the mean across scores provided by different subjects as follows:

$$MOS_j = \frac{1}{N}\sum_{i=1}^{N} s_{ij} \qquad (1)$$

where N is the number of subjects and s_{ij} is the score given by subject i for the stimulus j.

In order to evaluate the reliability of the obtained results from statistical point of view, 95% confidence intervals (CI), assuming a Students *t*-distribution of the scores, were computed together with MOS values.

IV. RESULTS AND DISCUSSION

R-D curves based on weighted PSNR (wPSNR) of the four evaluated methods are provided in Figure 2. In these plots, the horizontal axis reports the bitrate required to encode the LF image and the vertical axis represents the average wPSNR across all sub-aperture images calculated for YUV

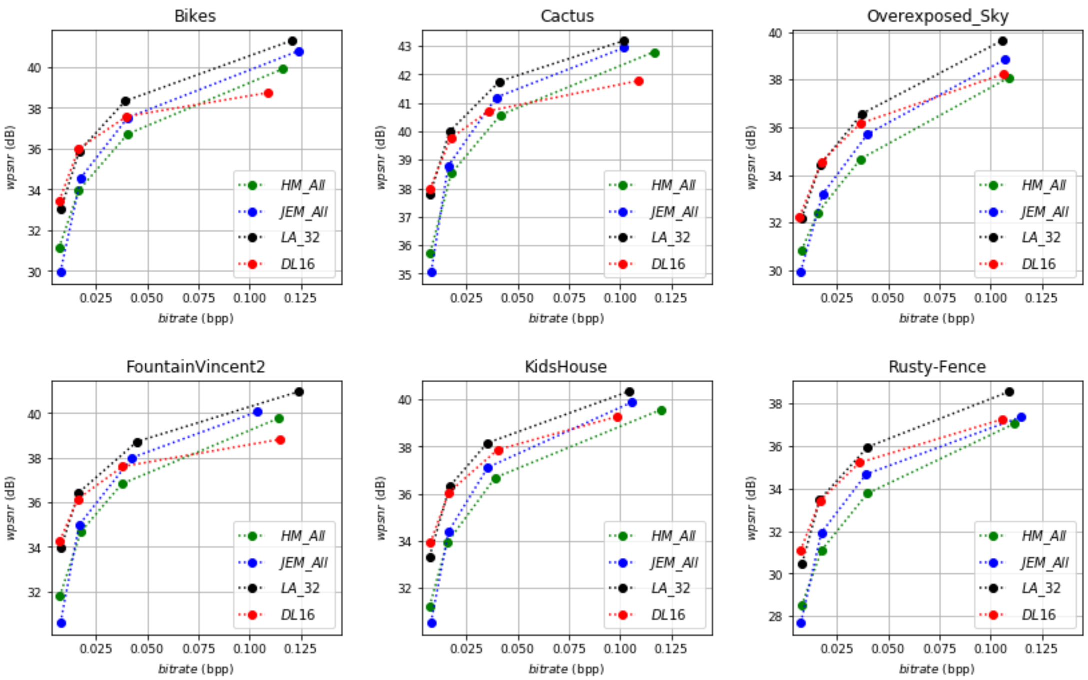

Fig. 2. R-D curves based on wPSNR of the four considered solutions for six different LF images.

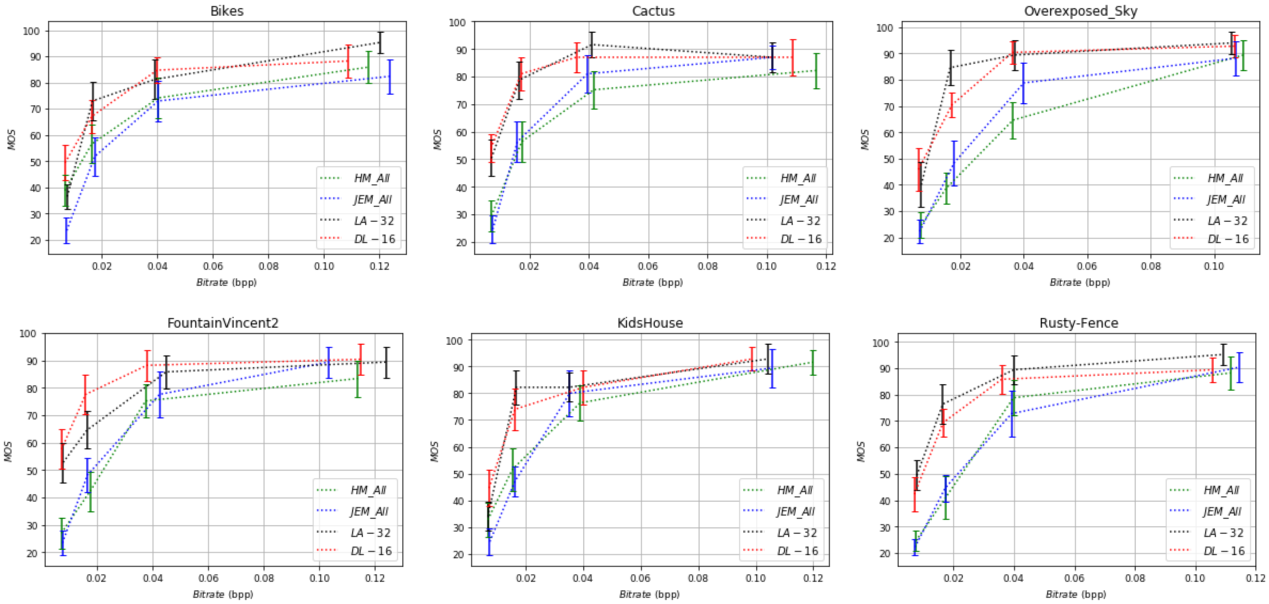

Fig. 3. MOS vs bitrate with associated confidence intervals for six different LF images.

channels, where the factor 6 is assigned to the luminance channel and the factor 1 for each chrominance channel [35]. One can observe that for all LF images and for all bitrates the LA-32 method provides the best result and outperforms the other compression solutions. The CNN-based view synthesis approach (DL-16) performs well at low and medium bitrates compared to HM-ALL and JEM-ALL methods, whereas provides low performance for the high bitrates. As expected, JEM-ALL outperforms HM-ALL for all tested LF images and for all bitrates, because it includes different improvements compared to HM, thus leading to an improvement of R-D performances. However, these results are reported according to wPSNR objective metric, which is not the best way for assessing the visual quality of LF images.

Thus, in the Figure 3, the fitted R-D curves based on the MOS are illustrated. The same conclusion may be drawn from this figure regarding the LA-32 method. However, for

DL-16 method, the results are quite different from objective evaluation, since this method achieves clearly better visual quality than HM-ALL and JEM-ALL methods, especially at low and medium bitrates. Globally, the LF coding methods based on view synthesis (LA-32 and DL-16) provide the highest visual quality at all bitrates. For instance, for most LF images their visual quality provided at medium bitrate is roughly the same as the one achieved by the naive coding approaches (HM-ALL and JEM-ALL) at high bitrate. Thus, the coding methods based on view synthesis can achieve high coding performance and demonstrate their effectiveness by providing the best visual quality compared to the two other methods.

V. Conclusion

In this paper, two recent LF compression methods based on view synthesis have been compared subjectively and objectively to two pseudo-video sequence based coding approaches. Experimental results show that the methods based on view synthesis achieve significant better coding performance without affecting the visual quality. Specifically, the subjective quality assessment showed that the view synthesis based methods provide substantial supperior visual quality, especially at low and medium bitrates.

Future works will include more LF compression methods based on view synthesis, as well as other more recent compression methods for LF images.

References

[1] C. Conti et al., “Light Field Image Compression,” in: Assunção P., Gotchev A. (eds) 3D Visual Content Creation, Coding and Delivery. Springer, Cham, 2019.

[2] I. Viola, M. Rerabek, T. Bruylants, P. Schelkens, F. Pereira, and T. Ebrahimi, “Objective and subjective evaluation of light field image compression algorithms,” in *IEEE Picture Coding Symposium (PCS)*, 2016.

[3] T. Ebrahimi, S. Foessel, F. Pereira, and P. Schelkens, “JPEG Pleno: Toward an efficient representation of visual reality,” *IEEE Multimedia*, vol. 23, no. 4, pp. 14–20, 2016.

[4] I. Viola, M. Rerabek and T. Ebrahimi, “Comparison and evaluation of light field image coding approaches,” *IEEE Journal of selected topics in signal processing*, vol. 11, no. 7, pp. 1092–1106, 2017.

[5] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, “Efficient intra prediction scheme for light field image compression,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 539–543.

[6] Y. Li, R. Olsson, and Sjöström, “Compression of unfocused plenoptic images using a displacement intra prediction,” in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, Jul. 2016.

[7] C. Conti, P. Nunes, and L. D. Soares, “HEVC-based light field image coding with bi-predicted self-similarity compensation,” in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2016.

[8] D. Liu, L. Wang, L. Li, Z. Xiong, F. Wu, and W. Zeng, “Pseudosequence-based light field image compression,” in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, July 2016.

[9] C. Perra and P. Assuncao, “High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement,” in *IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2016.

[10] J. Hou, J. Chen, and L.-P. Chau, “Light field image compression based on bi-level view compensation with rate-distortion optimization,” I*EEE Transactions on Circuits and Systems for Video Technology*, 2018.

[11] S. Zhao, Z. Chen, K. Yang, and H. Huang, “Light field image coding with hybrid scan order,” in *IEEE Visual Communications and Image Processing (VCIP)*, Nov. 2016.

[12] R. Olsson, M. Sjöström, and Y. Xu, “A combined pre-processing and h.264-compression scheme for 3d integral images,” in *IEEE International Conference on Image Processing*, 2006, pp. 513–516.

[13] P. Helin, P. Astola, B. Rao, and I. Tabus, “Sparse modelling and predictive coding of subaperture images for lossless plenoptic image compression,” in *IEEE 3DTV-Conference (3DTV-CON)*, Jul. 2016.

[14] F. Dai, J. Zhang, Y. Ma, and Y. Zhang, “Lenselet image compression scheme based on subaperture images streaming,” in *IEEE International Conference on Image Processing (ICIP)*, sep. 2015, pp. 4733–4737.

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

[16] X. Jiang, M. Le Pendu, and C. Guillemot, “Light field compression using depth image based view synthesis,” in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, July 2017, pp. 19–24.

[17] X. Huang, P. An, L. Shen, and R. Ma, “Efficient Light Field Images Compression Method Based on Depth Estimation and Optimization,” *IEEE Access*, vol. 6, pp. 48984–48993, 2018.

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

[19] N-K. Kalantari, T-C 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.

[20] Z. Zhao, S. Wang, C. Jia, X. Zhang, S. Ma, and J. Yang, “Light field image compression based on deep learning,” in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2018.

[21] N. Bakir, W. Hamidouche, O. Déforges, K. Samrouth, and M. Khalil, “Light field image compression based on convolutional neural networks and linear approximation,” in *Proc. IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 1128–1132.

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[23] P. Paudyal, F. Battisti, M. Sjöström, R. Olsson, and M. Carli, “Towards the perceptual quality evaluation of compressed light field images,” *IEEE Transactions on Broadcasting*, vol. 63, no. 3, pp. 507–522, 2017.

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

[25] M. Rerabek and T. Ebrahimi, “New light field image dataset,” in *Proc. IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, 2016.

[26] X. jiang, M. Le Pendu, R. Farrugia, and C. Guillemot, “Light Field Compression with Homography-based Low Rank Approximation,” *IEEE J. on Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1132–1145, Oct. 2017.

[27] P. Paudyal, J. Gutiérrez, P. Le Callet, M. Carli, and F. Battisti, “Characterization and selection of light field content for perceptual assessment,” in *Proc. IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, 2017.

[28] ITU-T P.910, “Subjective video quality assessment methods for multimedia applications,” *International Telecommunication Union*, Apr. 2008.

[29] D. Hasler and S. E. Suesstrunk, “Measuring colorfulness in natural images,” *Human vision and electronic imaging VIII*, Jun. 2003.

[30] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, “Depth estimation with occlusion modeling using light-field cameras,” *IEEE Trans. on pattern analysis and machine intelligence*, vol. 38, no. 11, pp. 2170–2181, 2016.

[31] D. G. Dansereau, O. Pizarro, and S. B. Williams, “Decoding, calibration and rectification for lenselet-based plenoptic cameras,”in *Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1027–1034.

[32] ITU-R BT.500-13, “Methodology for the subjective assessment of the quality of television pictures,” *International Telecommunication Union*, Jan. 2012.

[33] ITU-R BT.2022, “General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays,” *International Telecommunication Union*, Aug. 2012.

[34] F. Battisti, M. Carli, and P. Le Callet, “A Study on the Impact of Visualization Techniques on Light Field Perception,” in *Proc. IEEE European Signal Processing Conference (EUSIPCO)*, 2018.

[35] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, “Comparison of the coding efficiency of video coding standards–including high efficiency video coding (HEVC),” IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1669–1684, 2012.