Perceptual Light Field Image Coding with CTU Level Bit Allocation

Panqi Jin[1], Gangyi Jiang[1(✉)], Yeyao Chen[1], Zhidi Jiang[2], and Mei Yu[1]

[1] Faculty of Information Science and Engineering, Ningbo University, Ningbo, China
jiangganyi@126.com
[2] College of Science and Technology, Ningbo University, Ningbo, China

Abstract. Light field imaging simultaneously records the position and direction information of light in scene, as one of the important techniques for digital media. The amount of light field image (LFI) data is huge, it needs to be effectively compressed. In this paper, a perceptual LFI coding method with coding tree unit (CTU) level bit allocation strategy is proposed. To remove angular redundancy, a hybrid coding framework with joint deep learning reconstruction networks is constructed. At the encoder side, only four corner sub-aperture images (SAIs) are compressed with new CTU level bit allocation, a complete SAIs array is reconstructed by a LFI angular super-resolution network at the decoder side. To remove perceptual redundancy, we design a CTU level bit allocation strategy with the assumption of perceptual consistency, considering the characteristics of the human visual system in the bit allocation process. Experimental results show that for the proposed method with the designed CTU level bit allocation strategy, an average BD-BR savings of 13.676% in Y-PPSNR metric and 2.045% in VSI metric can be achieved. Compared with the high efficiency video coding (HEVC) intra coding model, the proposed method can achieve an average BD-BR savings of over 90%.

Keywords: Light Field Image · Perceptual Coding · Light Field Reconstruction

1 Introduction

Light field imaging can simultaneously record the intensity and direction information of light in a scene [1]. Light field images (LFIs) have many applications, such as refocusing [2], 3D reconstruction [3], and multi view display [4]. But the rich scene information makes the data volume of LFIs much larger than 2D images of the same resolution. Therefore, efficient compression of LFI is crucial for its applications.

Generally, LFI compression methods can be mainly divided into the traditional encoder based approach and the view synthesis based approach. The former directly uses or improves existing encoders to compress LFIs, for example, treating LFI's sub-aperture images (SAIs) as pseudo video sequence (PVS), and compressing the PVS with video encoders [5]. LFI's spatial and angular redundancies are removed through intra and inter prediction of the video encoder. Monteiro et al. [6] improved high efficiency video coding (HEVC) and used a prediction method combining local linear embedding and

N. Tsapatsoulis et al. (Eds.): CAIP 2023, LNCS 14185, pp. 255–264, 2023.
https://doi.org/10.1007/978-3-031-44240-7_25

self-similarity for LFI compression. Ahmad et al. [7] proposed a coding method using the multi-view extension of HEVC to explore the correlation between SAIs. These methods can remove most of the data redundancy, but encoding all the data results in limited encoding efficiency.

For the latter (view synthesis based approach), only a subset of SAIs is selected for encoding, and the rest of SAIs will be synthesized at the decoder side. Bakir et al. [8] compressed sparsely sampled SAIs, and used the LF Dual Discriminator GAN at the decoder side to synthesize discarded SAIs. Hedayati et al. [9] used JPEG to compress the central SAI and designed a deep learning network that includes quality enhancement and depth estimation to reconstruct the SAIs array. Huang et al. [10] compressed the selected SAIs and the disparity maps corresponding to the unselected SAIs, and rendered the unselected SAIs at the decoder side. Liu et al. [11] compressed eight selected SAIs and constituted multi-disparity geometry to reflect abundant disparity characteristics; then, synthesizing remaining LFI's SAIs using the multi-stream view reconstruction network at the decoder side. These methods improve encoding efficiency through sparse sampling and view synthesis. However, the selected SAIs are generally coded with video coding techniques. The intra frame-based coding tree unit (CTU) level bit allocation algorithms for existing video coders do not fully consider the visual perception characteristics. This leads to perceptual redundancy in the compressed SAI subset. Due to the fact that the SAIs in the subset will be used as references at the decoder side, this perceptual redundancy will be further transmitted to the synthesized SAIs.

Therefore, in this paper, a perceptual LFI coding method with a new CTU level allocation strategy is proposed to improve LFI coding efficiency. The experimental results show that the effect of bit allocation is maintained in synthesized SAIs. At the same bit rate, the proposed method achieved better subjective quality and structural consistency in the salient regions.

2 The Proposed Method

In this paper, a perceptual LFI coding method with new CTU level bit allocation strategy is proposed, and its framework is shown in Fig. 1. At the encoder side, the original LFI $\boldsymbol{L}_{org}$ is sparsely sampled, and the SAIs at four corner positions are selected to form a subset of SAIs $\boldsymbol{S}_{sel}$, which are arranged into PVS for input into HEVC. At the same time, the depth map I_D and saliency map I_S of the central SAI I_C are extracted separately through deep learning networks. Subsequently, I_C, I_D and I_S are input into the proposed bit allocation model to calculate the bit weight for each CTU. The complete set of weights $\boldsymbol{W}$ is input into HEVC to guide the target bit rate allocation at the CTU level. At the decoder side, the decoded SAIs set $\boldsymbol{S}'_{sel}$ is input into the LFI angular super-resolution network to recover the dropped SAIs. Finally, the complete reconstructed LFI $\boldsymbol{L}_{rec}$ consist of a synthesized SAIs set $\boldsymbol{S}'_{unsel}$ and $\boldsymbol{S}'_{sel}$.

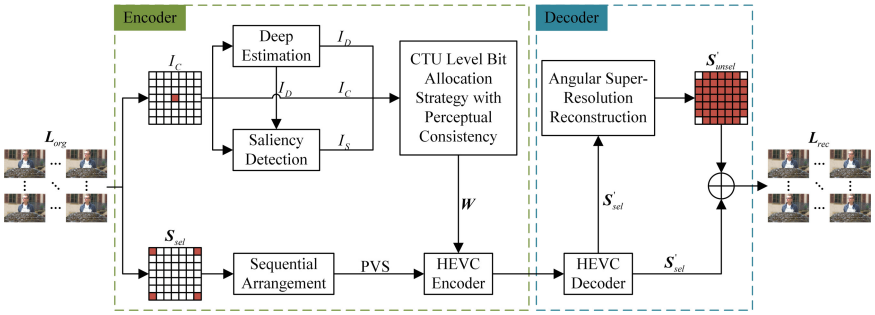

Fig. 1. The framework of the proposed LFI perceptual coding method.

2.1 Designed CTU Level Bit Allocation Strategy with Perceptual Consistency

The Assumption of Perceptual Consistency

There is a high content similarity between SAIs with different angular coordinates. Taking a 7×7 SAIs array as an example, Fig. 2 shows the residuals between I_C and other angular positional SAIs of the LFI *Fountain_&_Vincent_1*. Whether they are far away or adjacent, the residual between them is small. Therefore, the assumption of perceptual consistency for SAIs array is proposed, stating that the visual sensitive regions of SAIs with different angular coordinates are basically the same. Based on this assumption, each scene only needs to use I_C to calculate the weight of bit allocation once, rather than independently calculating for all selected SAIs. It is very meaningful for improving encoding speed.

(a) (b)

Fig. 2. The residual maps. (a) The residual between the SAIs located at (4,4) and (4,3). (b) The residual between the SAIs located at (4,4) and (1,1). (Here, pixel values are magnified by 4 times for visualization).

Calculation and Allocation of CTU Level Bit Weight

Initial Bit Weight Calculation

Compared to flat regions, complex texture regions have more complex prediction modes and deeper block depth. Generally, complex texture regions require more bit rate consumption to achieve the same quality as flat regions. In addition, studies have shown that humans pay more attention to complex texture regions than flat regions. Therefore, texture complexity is used as the initial bit weight for each CTU. CTUs with complex

textures are given larger initial weights, while flat CTUs are given smaller initial weights. The initial bit weight for each CTU is calculated as follows:

$$T_i = \sum_{x=1}^{M-1} \sum_{y=1}^{N-1} G_i(x, y) + c \tag{1}$$

where T_i is the texture complexity of the i-th CTU and also serves as the initial bit weight. M and N are the size of the CTU. c is a constant to avoid an initial bit weight of 0. $G_i(x, y)$ is the gradient value of the pixel at (x, y), and calculated as follows:

$$G_i(x, y) = |p_i(x, y) - p_i(x + 1, y)| + |p_i(x, y) - p_i(x, y + 1)| \tag{2}$$

where $P_i(x, y)$ is the pixel value at (x, y), $|\cdot|$ denotes an absolute value operation. The calculation is performed on the Y component of the image.

Weight Adjustment of Visual Sensitive Regions

When human eye observes images, the visual sensitivity of different regions varies. Regions with higher visual sensitivity should be assigned more bits. In the proposed method, the foreground region and the salient object region are considered as high visual sensitivity regions, and the bit weights of the CTU in these regions are adjusted. Firstly, the depth map I_D and saliency map I_S of I_C are obtained using deep learning networks [12] and [13]. Secondly, I_D and I_S are binarized to obtain the foreground mask and salient object mask. Then, the masks are employed to calculate the foreground density ρ_D and salient density ρ_S of each CTU, respectively. The calculation is expressed as follows:

$$\rho = \sum CTU / (M \times N) \tag{3}$$

where $\sum CTU$ is the number of pixels with the value of 1 in the binary mask corresponding to the CTU, and $M \times N$ is the size of the CTU. If $\rho_D > 0.5$, the CTU belongs to the foreground region, and similarly, if $\rho_S > 0.5$, the CTU belongs to the salient region. Finally, the bit weights of visual sensitive regions are adjusted based on the judgment results, and the calculation is expressed as follows:

$$W_i = \begin{cases} T_i, \ if \ \rho_D < 0.5 \ \& \ \rho_S < 0.5 \\ T_i \times \alpha, \ if \ \rho_D > 0.5 \ \& \ \rho_S < 0.5 \\ T_i \times \beta, \ if \ \rho_D < 0.5 \ \& \ \rho_S > 0.5 \\ T_i \times \alpha \times \beta, \ if \ \rho_D > 0.5 \ \& \ \rho_S > 0.5 \end{cases} \tag{4}$$

where α and β are weight adjustment factors used for the foreground and salient regions, respectively, to increase the bit weights of CTUs belonging to these regions. Based on extensive experiments, α and β are taken as 1.1 and 1.5, respectively. W_i is the final bit weight of the *i-th* CTU, used for allocating the target bit.

CTU Level Target Bit Allocation

After calculating the bit weights of all CTUs, the target bit is allocated for each CTU, and the calculation is expressed as follows:

$$R_i = \frac{(R_p - R_h - R_c) \times W_i}{\sum_{k=i}^{N_c} W_K} \tag{5}$$

where R_i is the target bit of the *i-th* CTU, R_p is the total target bits of the current frame, R_h is the actual consumption bits of frame header information encoding, R_c is the actual consumption bits of the encoded CTU and N_c is the total number of CTUs in the current frame. After allocating all the bits, QP is calculated based on $R-\lambda$ and $QP-\lambda$ model [14].

2.2 Decoding and Reconstruction

In the proposed method, only the selected SAIs set $\boldsymbol{S}_{sel}$ is compressed and transmitted, while the remaining SAIs are synthesized at the decoder side. The network and pre trained model in [15] are selected for LFI reconstruction. Specifically, $\boldsymbol{S'}_{sel}$ is fed into the angular super-resolution network to reconstruct complete LFI, and represented as:

$$\boldsymbol{L'} = f(\boldsymbol{S'}_{sel}) \tag{6}$$

where $\boldsymbol{L'}$ is the LFI output by the network, and it is already a complete SAIs array, and $f(\cdot)$ denotes the angular super-resolution network.

Finally, the reconstructed LFI $\boldsymbol{L}_{rec}$ is obtained as follows:

$$\boldsymbol{L}_{rec} = \boldsymbol{S'}_{sel} + \boldsymbol{S'}_{unsel}, \boldsymbol{S'}_{unsel} \in \boldsymbol{L'} \tag{7}$$

where $\boldsymbol{S'}_{unsel}$ is the set of SAIs from $\boldsymbol{L'}$ except for the four corner positions. The SAI at the four corner positions still uses $\boldsymbol{S'}_{sel}$ to minimize the reconstruction distortion caused by the angular super-resolution network.

3 Experimental Results and Analyses

3.1 Experimental Setup

The proposed method is tested on the commonly used EPFL light field database [16], which provides multiple scenes captured by a Lytro Illum camera. Here, the MATLAB light field toolbox [17] is adopted to decode the RAW light field data into a SAIs array, with angular and spatial resolutions are 15×15 and 434×625, respectively. Figure 3 shows the SAI thumbnails corresponding to the scenes used in this paper. In specific experiments, the central 7×7 SAIs array is selected, and the spatial resolution of each SAI is cropped to 432×624 to meet the requirements of the encoder for encoding block size. In addition, the SAIs in $\boldsymbol{S}_{sel}$ are arranged into PVS and converted into the format of 4:2:0 YUV. Due to only comparing intra encoding mode, the arrangement order of PVS will not affect the final performance.

The proposed bit allocation method is implemented using the HEVC reference software (HM16.20). Specifically, the PVS is encoded with All Intra coding structure. The size of the CTU is set to 64×64, and the maximum division depth is set to 4. Rate Control and LCU Level Rate Control are set to 1. Besides, the target bitrate of each sequence is collected under the platform of HM16.20 with fixed QPs (*i.e.*, QP = 22, 27, 32, 37, respectively).

Fig. 3. SAI thumbnails: *Caution_Bees*, *Danger_de_Mort*, *Fountain_&_Vincent_1*, *Stone_Pillars_Outside*, *Sophie_&_Vincent_on_a_Bench*, *Sophie_Krios_&_Vincent*.

Perceptual Peak Signal to Noise Ratio (PPSNR) [18] and Visual Saliency induced Index (VSI) [19] are adopted as the perceptual quality metrics. Between them, PPSNR is a quality metric that only targets salient region, and calculated as follows:

$$PPSNR = 10\log_{10} \times \frac{255^2}{\frac{1}{L\times H}\sum_{x=1}^{L}\sum_{y=1}^{H}(I(x,y) - I'(x,y))^2 \times \delta(x,y)} \tag{8}$$

where $\delta(x, y) = 1$ indicates the salient region, and $\delta(x, y) = 0$ indicates the non-salient region

Note that it is meaningful to calculate PSNR only for salient regions, as these regions are more susceptible to attention and have a greater impact on perceived quality. However, when the total bitrate is fixed, the increase of the bitrate in the salient region will inevitably be accompanied by the decrease of the bitrate in the non-salient region. Therefore, this paper also adopts VSI to evaluate the global quality of images. VSI considers the visual saliency and has been validated to be in line with human perception [19].

This paper measures encoding performance by calculating Bjontegaard Delta bitrate (BD-BR) [20]. A negative BD-BR value indicates that under the same quality, the proposed method can save more bitrate compared to the benchmark method, while conversely, it means consuming more bitrate. The bitrate is measured in bit per pixel (bpp) and calculated as follows:

$$bpp = \frac{R_{LF}}{x \times y \times u \times v} \tag{9}$$

where R_{LF} denotes the size of the bitstream, $x \times y$ and $u \times v$ are the spatial and angular resolutions of the LFI, respectively. In addition, the quality of each LFI is represented by the average quality of all SAIs.

Here, two compression methods are used for comparison to evaluate the effectiveness of the proposed method. The abbreviations for these methods are as follows:

- *HM*: Encode all SAIs on HM16.20. Except for the target bitrate collected when encoding 49 SAIs. The other configurations are consistent with the ones described earlier.
- *HM&ASR*: It can be seen as a version of the proposed method using HM's rate allocation strategy. Specifically, only four corner SAIs are encoded, and the remaining SAIs are synthesized by an angular super-resolution network.

3.2 Rate-Distortion Performance

Table 1 gives the Bjontegaard metrics [20] of the proposed method with *HM* and *HM&ASR* as the baselines, respectively. Y-PPSNR indicates to the average PPSNR metric for all SAIs calculated on the Y component. Compared with the *HM* method, the proposed method achieves average BD-BR savings of 90.305% in the Y-PPSNR metric and 90.825% in the VSI metric, respectively. This is mainly due to the sparse encoding, which saves a lot of bitrates. Compared to the *HM&ASR* method, the proposed method achieves an average BD-BR savings of 13.676% in the Y-PPSNR metric. This indicates that the proposed bit allocation method significantly improves the visual quality of salient regions. In addition, the proposed bit allocation method effectively balances the bitrates of non-salient regions, thereby improving the global quality of the image. This can be reflected in the average BD-BR savings of 2.045% in the VSI metric.

Table 1. The BD-BR comparison of the proposed method with *HM* and *HM&ASR* methods as baselines, respectively.

Scenes	*Proposed* vs *HM*		*Proposed* vs *HM&ASR*	
	BD-BR (Y-PPSNR)	BD-BR (VSI)	BD-BR (Y-PPSNR)	BD-BR (VSI)
101	–91.494%	–90.551%	–17.490%	–2.477%
102	–91.283%	–90.538%	–10.982%	–0.635%
103	–88.590%	–90.152%	–14.895%	–0.998%
104	–91.728%	–91.482%	–23.650%	–4.545%
105	–88.552%	–90.843%	–13.245%	–3.162%
106	–90.182%	–91.381%	–1.795%	–0.451%
Avg	**–90.305%**	**–90.825%**	**–13.676%**	**–2.045%**

Figure 4 show the visual comparison results of the decoded central SAI, where the red box is taken from the salient region of the image and the blue box is taken from non-salient region, and the PSNR values of these regions are given. It can be found that the proposed method maintains better details in salient regions, such as the eye details. Correspondingly, the quality of the proposed method has decreased in non-salient regions. However, this has a small impact on the overall perceived quality, as these regions have a low level of attention. Moreover, in the proposed method, the central SAI is not encoded, but synthesized by the decoder side. The experimental results indicate that the designed bit allocation strategy not only affects $\boldsymbol{S}_{sel}$, but also affects $\boldsymbol{S'}_{unsel}$, thereby generating results with better visual quality.

Fig. 4. Comparison of the decoded central SAI of *Sophie_&_Vincent_on_a_Bench* (105). The number in the sub-figure indicates the PSNR of the region. Here, 0.324bpp for *HM* methods, 0.026bpp for *HM&ASR* method and the proposed method.

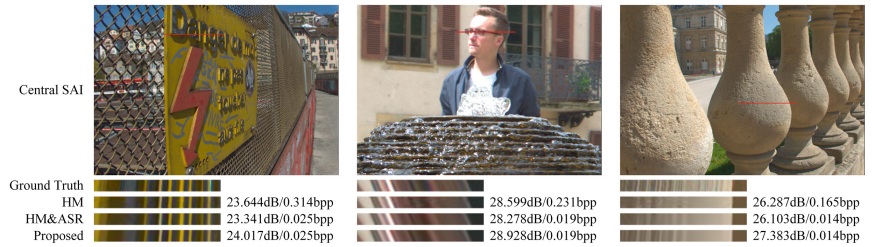

Fig. 5. Comparison of the EPI consistency of salient regions of *Danger_de_Mort* (102), *Fountain_&_Vincent_1* (103), *Stone_Pillars_Outside* (104). Unit: PSNR/bpp.

3.3 Structural Consistency of Reconstruction

The structural consistency of LFIs is considered key to techniques such as refocusing and depth inference. Epipolar Plane Images (EPIs) contain parallax changes and object occlusion information, and the continuity of their polar lines can well reflect the structural consistency of the LFI. Hence, Fig. 5 shows the EPIs in the salient regions extracted from the decoded results. It can be observed that the proposed method achieves higher PSNR and visual quality at lower bitrates. The HM method independently encodes each SAI, consuming large bitrates while damaging structural consistency, resulting in significant distortion on the EPI. The HM&ASR method saves bits through sparse encoding, but also introduces reconstruction distortion generated by deep learning networks. In contrast, the proposed method increases the bitrates of the salient regions by reallocating the CTU level bit, thereby improving the quality of the regions, and at the same time enhancing the structural consistency of the salient regions.

4 Conclusions

This paper presents a perceptual light field image (LFI) coding method with coding tree unit (CTU) level bit allocation strategy. At the encoder side, the four corner sub-aperture images (SAIs) are compressed. In order to remove the perceptual redundancy, a CTU level bit allocation strategy with perceptual consistency is proposed. Firstly, the texture features of each CTU of central SAI are extracted as the initial bit weight. Then, the bit weight of CTU belonging to foreground and salient regions are adjusted to obtain the final bit weight. Finally, the calculated weights are employed to allocate the target bit of each CTU. At the decoder side, the complete SAIs array is reconstructed by the LFI angular super-resolution network. The experimental results show that the proposed method can effectively improve the quality of the salient regions and the overall image at the same bitrate, while maintaining better structural consistency of the salient regions.

This work was supported in part by the Natural Science Foundation of China under Grant Nos. 62271276, 62071266 and 61931022, in part by the Natural Science Foundation of Ningbo under Grant No. 202003N4088, and in part by Science and Technology Innovation 2025 Major Project of Ningbo under Grant No. 2022Z076.

References

1. Xiang, J., Jiang, G., Yu, M., Jiang, Z., Ho, Y.-S.: No-reference light field image quality assessment using four-dimensional sparse transform. IEEE Trans. Multimedia **25**, 457–472 (2023)
2. Yang, N., et al.: Detection method of rice blast based on 4D light field refocusing depth information fusion. Comput. Electron. Agric. **205**, 107614 (2023)
3. Yuan, L., Gao, J., Wang, X. and Cui, H.: Research on 3D reconstruction technology based on the fusion of polarization imaging and light field depth information. In: 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), pp. 1792–1797. IEEE, Xi'an, China (2022)
4. Shen, S., Xing, S., Sang, X., Yan, B., Chen, Y.: Virtual stereo content rendering technology review for light-field display. Displays **76**, 102320 (2022)
5. Dai, F., Zhang, J., Ma, Y. and Zhang, Y.: Lenselet image compression scheme based on subaperture images streaming. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 4733–4737. IEEE, Quebec City, QC, Canada (2015)
6. Monteiro, R., Lucas, L., Conti, C., et al.: Light field HEVC-based image coding using locally linear embedding and self-similarity compensated prediction. In: 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1–4. IEEE, Seattle, WA, USA (2016)
7. Ahmad, W., Olsson, R., Sjöström, M.: Interpreting plenoptic images as multi-view sequences for improved compression. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 4557–4561. IEEE, Beijing, China (2017)
8. Bakir, N., Hamidouche, W., Fezza, S.A., Samrouth, K., Déforges, O.: Light field image coding using VVC standard and view synthesis based on dual discriminator GAN. IEEE Trans. Multimedia **23**, 2972–2985 (2021)
9. Hedayati, E., Havens, T.C., Bos, J.P.: Light field compression by residual CNN-assisted JPEG. In: 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–9. IEEE, Shenzhen, China (2021)

10. Huang, X., An, P., Chen, Y., Liu, D., Shen, L.: Low bitrate light field compression with geometry and content consistency. IEEE Trans. Multimedia **24**, 152–165 (2022)
11. Liu, D., Huang, Y., Fang, Y., Zuo, Y., An, P.: Multi-Stream Dense View Reconstruction Network for Light Field Image Compression. IEEE Transactions on Multimedia, early access (2022)
12. Miangoleh, S.M.H., Dille, S., Mai, L., Paris, S., Aksoy, Y.: Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9680–9689. IEEE, Nashville, TN, USA (2021)
13. Wang, F., Pan, J., Xu, S., Tang, J.: Learning discriminative cross-modality features for RGB-D saliency detection. IEEE Trans. Image Process. **31**, 1285–1297 (2022)
14. Li, B., Li, H., Li, L., Zhang, J.: λ domain rate control algorithm for high efficiency video coding. IEEE Trans. Image Process. **23**(9), 3841–3854 (2014)
15. Wang, Y., et al.: Disentangling light fields for super-resolution and disparity estimation. IEEE Trans. Pattern Anal. Mach. Intell. **45**(1), 425–443 (2023)
16. EPFL dataset. https://www.epfl.ch/labs/mmspg/downloads/epfl-light-field-image-dataset/. Accessed 28 April 2023
17. Dansereau, D.G., Pizarro, O., Williams, S.B.: Decoding, calibration and rectification for lenselet-based plenoptic cameras. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1027–1034. IEEE, Portland, OR, USA (2013)
18. Majid, M., Owais, M., Anwar, S.M.: Visual saliency based redundancy allocation in HEVC compatible multiple description video coding. Multimed. Tools Appl. **77**, 20955–20977 (2018)
19. Zhang, L., Shen, Y., Li, H.: VSI: a visual saliency-induced index for perceptual image quality assessment. IEEE Trans. Image Process. **23**(10), 4270–4281 (2014)
20. Bjontegaard, G.: Calculation of average PSNR differences between RD-curves. ITU SG16 Doc. VCEG-M33 (2001)