Exploring Temporal Consistency in Image-Based Rendering for Immersive Video Transmission

Smitha Lingadahalli Ravi, Félix Henry, Luce Morin, Matthieu Gendrin

Exploring Temporal Consistency in Image-Based Rendering for Immersive Video Transmission

Smitha Lingadahalli Ravi
Orange Labs
35510 Cesson Sévigné, France
smitha.lingadahalliravi@orange.com

Félix Henry
Orange Labs
35510 Cesson Sévigné, France
felix.henry@orange.com

Luce Morin
INSA Rennes - IETR
35000 Rennes, France
luce.morin@insa-rennes.fr

Matthieu Gendrin
Orange Labs
35510 Cesson Sévigné, France
matthieu.gendrin@orange.com

***Abstract*—Image-based rendering methods synthesize novel views given input images captured from multiple viewpoints to display free viewpoint immersive video. Despite significant progress with the recent learning-based approaches, there are still some drawbacks. In particular, these approaches operate at the still image level and do not maintain consistency among consecutive time instants, leading to temporal noise. To address this, we propose an intra-only framework to identify regions of input images leading to temporally inconsistent synthesized views. Our method synthesizes better and more stable novel views, even in the most general use case of immersive video transmission. We conclude that the network seems to identify and correct spatial features at the still image level that produce artifacts in the temporal dimension.**

***Index Terms*—image-based rendering, temporal consistency, immersive video transmission**

I. Introduction

The moving picture experts group (MPEG) has been actively developing the MPEG Immersive Video (MIV) standard [1] to efficiently transmit 6 degrees of freedom for realistic and virtual environments. The MIV is a concrete step towards a complete chain for immersive video coding, delivery, and rendering. Figure 1 shows the workflow of immersive video transmission. The views in the server are transmitted through MIV reference software to the client side. The retrieved views on the client side are input in the synthesizer, and the generated novel views are visualized using a head-mounted display (HMD) or a monitor.

The MIV reference software has non-normative encoding, normative decoding, and non-normative rendering techniques. The non-normative rendering techniques use depth maps to perform novel view synthesis. The quality of the synthesized view highly depends on the quality of the depth maps in such depth image-based rendering (DIBR) techniques. The DIBR techniques can be broadly classified as conventional depth estimators [2, 3] and learning-based depth estimators [4, 5, 6]. As shown in [7, 8], the conventional and learning-based depth estimators have various limitations. The conventional depth estimation requires high computational complexity, high energy requirements, and slower run-time. The learning-based depth estimators underperform quality-wise as the ground-truth depth maps used for training the network (especially that of real-world scenes) are nowhere near perfect, and these imperfect depth maps may lead to disocclusions, ghosting artifacts in the novel synthesized view. Also, the learning-based methods only try to improve the quality of the depth, but they should be trained end-to-end with the synthesized view quality as a loss function. This is not feasible due to the synthesis being non-differentiable. On the other hand, image-based rendering techniques (IBR) [9] have been proposed for novel view synthesis. These techniques are compatible with a MIV transmission of the immersive video (using the Geometry Absent profile [7]).

In IBR methods, the source views are usually warped, resampled, and/or blended to obtain target viewpoints. Such methods enable high-resolution rendering, but they typically require dense input views or explicit proxy geometry. Without this, it is challenging to estimate high-quality views, resulting in artifacts in rendering. Earlier, researchers used dense sampling from the scene to create light fields [12, 13]. IBR techniques [14, 15] use proxy geometry of the scene to generate novel views. The next-generation techniques introduced better modeling of the scene structure [16, 17, 18]. Since the dawn of deep learning, learning-based methods [19, 20, 21, 22, 23] have generated promising results. Recently, techniques that combine novel representations [10, 24, 25, 26, 27, 28, 29, 30] with a differentiable rendering have produced high-quality novel views. Of those techniques, the prominent one is neural radiance fields (NeRF) [10], which encodes the 3D scene structure in a continuous 5D volumetric function. Although NeRF produces high-quality view synthesis results, it has to overfit every scene and requires tedious per-scene optimization. Besides, the network parameters must be transmitted for each time instant of the video. This is not practical for immersive video transmission.

Very recently, a new learning-based method called IBRNet [11] that leverages ideas from IBR and NeRF was introduced. Unlike NeRF, which embeds the signal in the network, IBRNet is a pure processor. It does not need per-scene optimization and operates as a general synthesizer which is able to work with any new content. The main idea of IBRNet is to obtain colors and densities by aggregating information present in the neighboring views. IBRNet can be divided into three steps: 1. Selecting the neighboring views of the target view and extracting dense features from each neighboring view 2. Predicting volume densities and colors at continuous 5D locations 3. Compositing the extracted colors and densities

978-1-6654-6623-3/22/$31.00

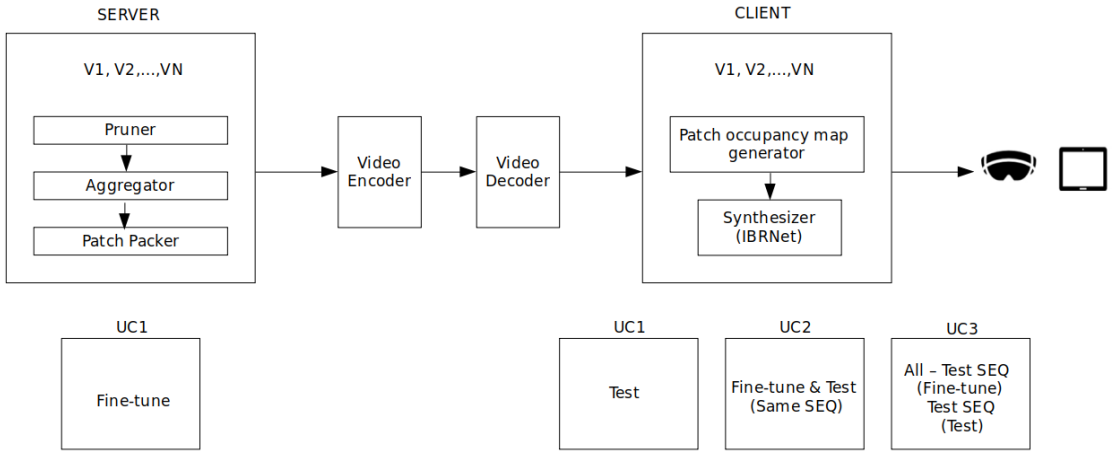

Fig. 1. The workflow of immersive video transmission along with the various use cases

along each camera ray to produce the target image. During training, the color is rendered along each camera ray, and the mean squared error is minimized between the ground-truth pixel color and the rendered pixel color. As IBRNet is a pixel-by-pixel rendering and produces state-of-the-art results, we are using it to synthesize novel views for immersive video transmission.

Even though IBRNet is a state-of-the-art rendering method, it has three main limitations. First, it is challenging to generate high-quality novel views based on a sparse set of input views. Second, it mainly considers static images with non-moving objects in the scene. Third, it suffers from intra-frame-based processing: views from a given temporal instant are synthesized only from views at the same instant, preventing any kind of temporal stability. As can be seen in videos [38], visually annoying artifacts appear temporally. Although they have a small magnitude and are almost invisible in a still image, they are quite visible when the video is displayed. To overcome this problem while maintaining the practical convenience of processing the views at intra level only, we propose an intra-framework to IBRNet, which synthesizes temporally consistent novel views. The remainder of the paper is structured as follows. Section 2 introduces the pipeline of our novel intra-framework. Section 3 shows the experimental details of our experiments. Section 4 reports the result, and section 5 concludes the paper.

ALGORITHM 1

Require: *T1* = Original view at *t*
T2 = Original view at *t*+1
M = Motion mask
Th1 = Predetermined threshold value
Let *(r1, g1, b1)*, *(r2, g2, b2)*, and *(rm, gm, bm)* be the RGB pixels of *T1*, *T2*, and *M* respectively at a given location.
For each pixel location:
*d = sqrt((r1 - r2)**2 + (g1 - g2)**2 + (b1 - b2)**2)*
if *d > Th1* **then**
(rm, gm, bm) = (255, 255, 255)
else:
(rm, gm, bm) = (0, 0, 0)
end if

ALGORITHM 2

Require: *O* = Masked original view at *t*
S = Masked synthesized view at *t*
T = Temporal guidance map at *t*
Th2 = Predetermined threshold value
Let *(ro, go, bo)*, *(rs, gs, bs)*, and *(rt, gt, bt)* be the RGB pixels of *O*, *S* and *T* respectively at a given location.
For each pixel location:
*d = sqrt((ro - rs)**2 + (go - gs)**2 + (bo - bs)**2)*
if *d > Th2* **then**
(rt, gt, bt) = (ro, go, bo)
else:
(rt, gt, bt) = (255, 255, 255)
end if

II. Intra-Framework Approach

To improve the temporal consistency among the consecutive frames, we constrain the fine-tuning phase of IBRNet to use original image pixels coming only from the temporal artifacts region of the image. To identify these pixels, we propose a temporal artifacts extraction method. As shown in Figure 2, the extraction can be carried out in four steps:

In the first step, we obtain the motion mask (Fig 2[b]) by taking a pixel-by-pixel difference between the *t*, *t+1* consecutive frames of an original view (Fig 2[a]). As shown in Algorithm 1, if the absolute difference value of the pixel is larger than an arbitrary threshold *Th1* then the value is changed to white pixels (255, 255, 255). Otherwise, it is set to black pixels (0, 0, 0). In the motion mask, black pixels are identified as having negligible motion: these are the pixels that should be synthesized with more temporal stability. In the second step, we synthesize the original view at *t* from neighboring views using IBRNet. This synthesized view (Fig 2[c]) has both static and temporal artifacts: static ones are due to the inability of IBRNet to fully reconstruct the signal, and other ones have

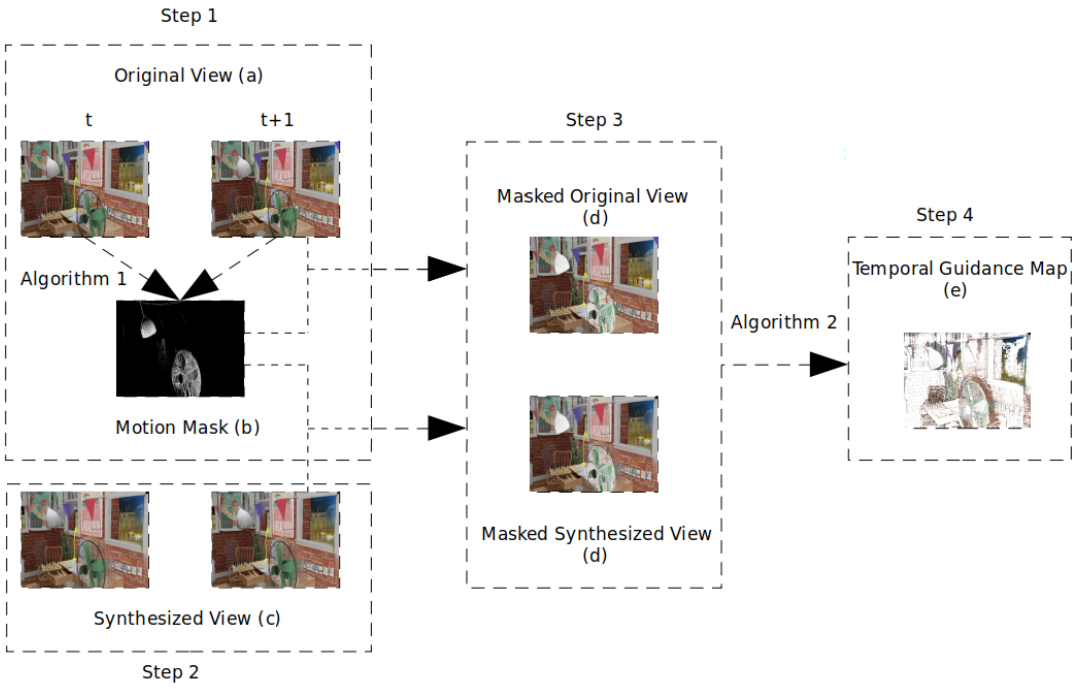

Fig. 2. The production of a temporal guidance map.

a temporal incoherent behaviour which are detectable to the human eye. In the third step, we add the mask obtained in step 1 to both the original and synthesized view (Fig 2[d]), which allows us to remove pixels affected by the original motion. Finally, in the fourth step, we extract only the region of temporal artifacts (Fig 2[e]). As shown in Algorithm 2, if the distance between the original and synthesized value of a pixel is larger than a second threshold *Th2*, we keep the pixels as they are. Otherwise, the pixels are substituted by white pixels. Table 1 shows the values of threshold used for each sequence. The thresholds are chosen experimentally such that the amount of active motion pixels is roughly similar to the number of pixels which are affected by temporal noise (typically, between 15 to 30 percent of the entire image, but this may vary between sequences).

Thus, we are able to isolate pixels affected by temporal incoherence, and the final output is a temporal guidance map with only relevant pixels in temporally unstable regions. There is one temporal guidance map associated with each original image. During fine-tuning, the input to the IBRNet network will be original views along with their temporal guidance maps and camera parameters: each pixel of the temporal guidance map is read, and if a white pixel is found, its corresponding pixel in the original view is simply skipped.

III. Experiments

A. Experimental Settings

To study the performance of the proposed method in the context of immersive video transmission, we consider three use cases, as shown in Figure 1.

Sequence	Type	Resolution	Frames	Camera Setup	Th1	Th2
Fan	Real-World	1920x1080	97	5x3	25	10
Mirror	Synthetic	1920x1080	97	5x3	25	10
Frog	Real-World	1920x1080	300	13x1	35	20
Shaman	Synthetic	1920x1080	300	5x5	25	10
Carpark	Real-World	1920x1088	150	9x1	25	10
Street	Real-World	1920x1088	150	9x1	25	10

Table 1. The MPEG-I sequences used for evaluation along with their type, resolution, number of frames, and the threshold utilized for tests.

a) Use Case 1: In this case, the fine-tuning is carried out on the server side. We fine-tune and test on the same set of frames of a sequence, and assume that the fine-tuning parameters can be transmitted losslessly to the client.

b) Use Case 2: This is a per-scene fine-tuning on the client side where the fine-tuning and testing are done on a different set of frames of the same sequence. In this use case, it is not necessary to transmit any parameter as the fine-tuning can be done in the client (at the expense of substantial complexity).

c) Use Case 3: This is a universal solution where we use five different sequences to fine-tune the network and then test it with a new sequence. This is the most realistic scenario, as the resulting synthesizer is not data dependent: the parameters are retrained once and for all (using the temporal guidance maps). After, it can be used as a classical IBRNet.

B. Dataset

For our experiments, we have used the following MPEG-I test sequences: Fan, Mirror, Frog, Shaman, Carpark, and Street. All the sequences are multi-view captured by a sparse camera setup. In these sequences, Fan and Shaman are syn-

MPEG-I Sequences	MSE ↓				PSNR ↑			
	Anchor	*UC1*	*UC2*	*UC3*	*Anchor*	*UC1*	*UC2*	*UC3*
Fan	507.61	**343.54**	361.82	486.19	22.67	**24.05**	23.64	22.93
Mirror	722.53	**601.39**	645.24	702.41	23.15	**24.68**	23.91	23.46
Frog	4517.12	**3341.92**	3402.54	3845.92	13.55	**17.21**	16.38	14.92
Shaman	321.89	**186.33**	214.82	299.43	23.41	**25.97**	24.37	23.82
Carpark	1419.26	**1128.47**	1206.31	1397.21	16.22	**17.85**	17.02	16.68
Street	1628.44	**1541.85**	1595.26	1633.75	17.48	**18.52**	17.98	17.65

Table 2. The comparison of average quality of synthesized views using MSE (lower means better), and PSNR (higher means better) metrics with respect to various use cases.

MPEG-I Sequences	VMAF ↑				MS-SSIM ↑				PSNR ↑			
	Anchor	*UC1*	*UC2*	*UC3*	*Anchor*	*UC1*	*UC2*	*UC3*	*Anchor*	*UC1*	*UC2*	*UC3*
Fan	52.62	53.05	**55.28**	53.86	0.9177	0.9203	**0.9365**	0.9248	27.18	27.72	**27.96**	27.53
Mirror	64.26	65.28	**66.85**	65.92	0.9311	0.9365	**0.9483**	0.9426	28.17	**29.36**	29.16	28.48
Frog	49.38	61.52	**63.74**	56.48	0.8928	0.9223	**0.9509**	0.9342	22.97	25.11	**25.35**	23.65
Shaman	57.72	**64.85**	64.32	58.54	0.9816	**0.9894**	0.9852	0.9821	32.84	**33.12**	32.91	32.86
Carpark	65.35	68.70	**70.18**	67.11	0.9426	0.9458	**0.9528**	0.9491	28.34	29.25	**29.97**	29.12
Street	73.63	75.28	**76.81**	74.92	0.9501	0.9546	**0.9612**	0.9558	29.01	29.67	**29.82**	29.38

Table 3. The comparison of average quality of synthesized views using VMAF (higher means better), MS-SSIM (higher means better), and PSNR (higher means better) metrics with respect to various use cases.

Fig. 3. A specialized MSE calculation only in the region of temporal inconsistency of the image. The top row contains original and synthesized views and the bottom row contains the temporal guidance map.

thetic or computer-generated sequences and the rest are real-world sequences. Table 1 shows the characteristics of the selected MPEG-I sequences.

C. Quality Metrics

For our analysis, we have used the following metrics to determine the performance of our method. These metrics are reliable and have been evaluated as relevant for estimating the subjective quality of temporal coherence [37].

a) Video Multimethod Assessment Fusion (VMAF) [31]: It predicts the subjective video quality based on a reference and synthesized video sequence by combining various quality metrics such as Visual Information Fidelity (VIF) [32], Detail Loss Metric (DLM) [33], Mean Co-Located Pixel Difference (MCPD) [34]. The MCPD metric measures the temporal difference between the frames.

b) Multi-scale Structural Similarity (MS-SSIM) [35]: The structural similarity (SSIM) [36] index quantifies the SSIM between an image and a reference image to determine perceived quality. The MS-SSIM index is calculated by merging the SSIM index of various versions of the image at various scales with the multissim function.

c) Mean Squared Error (MSE): It measures the average squared difference between the estimated pixels and original pixels. The MSE is calculated only in the region of temporal inconsistency. As shown in Figure 3, we keep the temporal guidance map as a reference and select only the region of temporal artifacts by skipping the white pixels in the synthesized and original images to measure the MSE.

d) Peak Signal-to-Noise Ratio (PSNR): It is a quality measurement between an original and synthesized view. In our experiments, the PSNR is calculated for both the full image and only the region of temporal artifacts.

IV. Results and Discussion

The quality of synthesized views for various use cases is presented in Table 2 and Table 3. Quality measures are produced by synthesizing each existing view (without using the target view) and computing the measure over this view. The final measure is averaged over all views. The anchor column corresponds to the views synthesized using a model fine-tuned on other MPEG-I test sequences to avoid domain adaptation issues. All the use cases are compared to the anchor views. As shown in Figure 4, our novel views consistently appear sharper and are less noisy than the anchor views [38]. The novel views in all the use cases tend to synthesize more details from the original views than the anchor, for instance, the background wallpaper of the frog sequence.

In Table 2, the MSE is computed only on the active pixels of the temporal guidance map. This is to evaluate whether we did improve the performance locally as intended. As shown in Table 2, UC1 has better quality views than all use cases in every sequence: this is expected as the fine-tuning is done on the same frames as the inference. The UC2 still produces

Fig. 4. Qualitative comparison of temporal artifacts in synthesized views of Fan, Frog and Shaman (top to bottom)

a substantial improvement over the anchor, even when the fine-tuning is done on previous frames. The most interesting result is UC3, where the fine-tuning is done on all sequences except the evaluated sequence: in this scenario, a gain in MSE is still observed. It is likely that because of the temporal guidance maps, the network is able to identify the specific spatial features of the signal at a given time instant that are producing temporal artifacts in relation to subsequent frames. Since the fine-tuned IBRNet is not data-dependent, it can be deployed once (there is not parameter transmission) and still produce an improvement.

Since we have shown that the temporal guidance maps based fine-tuning can help to improve the synthesis over the specific areas subject to temporal incoherence, the next question is whether it impacts the rest of the synthesized view. To answer, we show in Table 3 that VMAF, MS-SSIM and PSNR are calculated on full images. In UC1 and UC2, an increase in quality are obtained on both measures. UC2 performs better, which could be explained by the fact that UC1 is a rather extreme form of local fine-tuning which could be done at the expense of the rest of the view. Here again, the most important result is in UC3: there is still an improvement in the synthesis quality compared to the anchor in all sequences. This seems to confirm that the fine-tuning using temporal guidance maps has allowed the network to improve the synthesis, even in an intra-frame synthesis approach and in the most general case of offline fine-tuning. This allows the fine-tuned IBRNet to be deployed in the same way as the anchor version.

V. Conclusion

In this paper, we propose a novel intra-framework approach to improve temporal consistency in IBRNet for immersive video transmission. Our technique is easy to implement, requires no architectural changes to the network, and shows that the proposed method significantly improves temporal stability in all use cases. Despite the intra-frame nature of IBRNet, the experiments show that temporal synthesis performance can be improved even in the most general case of offline fine-tuning, which leads us to suggest that the network was able to pick up better spatial features that inherently produce temporal artifacts. In future work, we would like to continue improving the temporal stability by incorporating the motion information as an input to the synthesis network.

References

[1] Test Model 4 for Immersive Video, ISO/IEC JTC1/SC29/WG11 MPEG/N18795, October 2019, Geneva, Switzerland.

[2] Eduardo Juarez et. al., Manual of Depth Estimation Reference Software, (DERS 8.0), ISO/IEC JTC 1/SC 29/WG 11 N18450, Geneva, Switzerland, March 2019.

[3] Dawid Mieloch, Adrian Dziembowski, Jakub Stankowski, Olgierd Stankiewicz, Marek Domanski, Gwangsoon Lee, and Yun Young Jeong, "Immersive video depth estimation," ISO/IEC JTC 1/SC29/WG 11 m53407, Apr. 2020.

[4] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "GA-Net: Guided aggregation net for end-to-end stereo matching," in CVPR, 2019.

[5] Yao, Yao, et al. "MVSNet: Depth inference for unstructured multi-view stereo." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

[6] Wei, Zizhuang, et al. "AA-RMVSNet: Adaptive aggregation recurrent multi-view stereo network." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[7] Mieloch, D., Garus, P., Milovanović, M., Jung, J., Jeong, J. Y., Ravi, S. L., Salahieh, B. (2022). Overview and Efficiency of Decoder-Side Depth Estimation in MPEG Immersive Video. IEEE Transactions on Circuits and Systems for Video Technology.

[8] Smitha Lingadahalli Ravi, Marta Milovanović, Luce Morin, Félix Henry, "A Study of Conventional and Learning-Based Depth Estimators for Immersive Video Transmission", unpublished.

[9] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. SIGGRAPH Asia, 2018.

[10] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. ECCV, 2020.

[11] Wang, Qianqian, et al. "IBRNet: Learning multi-view image-based rendering." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.
[12] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, pages 43–54, 1996.
[13] Marc Levoy and Pat Hanrahan. Light field rendering. In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, pages 31–42, 1996.
[14] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, pages 425–432, 2001.
[15] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, pages 11–20, 1996.
[16] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In Proceedings of the 25th annual conference on Computer graphics and interactive techniques, pages 231–242, 1998.
[17] Eric Penner and Li Zhang. Soft 3d reconstruction for view synthesis. ACM Transactions on Graphics (TOG), 36(6):1–11, 2017
[18] Abe Davis, Marc Levoy, and Fredo Durand. Unstructured light fields. In Computer Graphics Forum, volume 31, pages 305–314. Wiley Online Library, 2012.
[19] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5515–5524, 2016.
[20] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. ACM Transactions on Graphics (TOG), 37(6):1–15, 2018.
[21] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. ACM Transactions on Graphics (TOG), 35(6):1–10, 2016.
[22] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. ACM Transactions on Graphics (TOG), 38(6):1–15, 2019.
[23] Zexiang Xu, Sai Bi, Kalyan Sunkavalli, Sunil Hadap, Hao Su, and Ravi Ramamoorthi. Deep view synthesis from sparse photometric images. ACM Trans. Graph., 38(4), July 2019.
[24] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2367–2376, 2019
[25] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. Local implicit grid representations for 3d scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6001–6010, 2020.
[26] Kai-En Lin, Zexiang Xu, Ben Mildenhall, Pratul P Srinivasan, Yannick Hold-Geoffroy, Stephen DiVerdi, Qi Sun, Kalyan Sunkavalli, and Ravi Ramamoorthi. Deep multi depth panoramas for view synthesis. In ECCV, 2020.
[27] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. arXiv preprint arXiv:1906.07751, 2019.
[28] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG), 2019.
[29] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2437–2446, 2019.
[30] Pratul P. Srinivasan, Richard Tucker, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. CVPR, 2019.
[31] VMAF: Perceptual video quality assessment based on multi-method fusion, Netflix, Inc., 2017-07-14, retrieved 2017-07-15
[32] Sheikh, Hamid R., and Alan C. Bovik. "A visual information fidelity approach to video quality assessment." The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics. Vol. 7. No. 2. sn, 2005.
[33] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," IEEE Transactions on Multimedia, vol. 13, pp. 935–949, Oct 2011.
[34] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," The Netflix Tech Blog, vol. 6, no. 2, 2016.
[35] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multiscale structural similarity for image quality assessment," The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003, 2003, pp. 1398-1402 Vol.2, doi: 10.1109/ACSSC.2003.1292216.
[36] Wang, Zhou; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. (2004-04-01). "Image quality assessment: from error visibility to structural similarity". IEEE Transactions on Image Processing. 13 (4): 600–612.
[37] J. Jung, J.G. Lopez, X. Li, S. Liu (Tencent). "Evaluation of objective quality metrics on HEVC, VVC and AV1 contents" m59384, All AG 05 MPEG Visual quality assessment, April 2022, Virtual.
[38] https://tinyurl.com/temporalresults