MUSE: A Multi-view Synthesis Enhancer

Nour Hobloss
IRT b<>com
Rennes, FRANCE
nour.hobloss@conti-engineering.com

Joshua Maraval
IRT b<>com
Rennes, FRANCE
joshua.maraval@b-com.com

Jérôme Fournier
IRT b<>com
Rennes, FRANCE
jerome.fournier@orange.com

Nicolas Ramin
IRT b<>com
Rennes, FRANCE
nicolas.ramin@b-com.com

Lu Zhang
Univ Rennes, INSA Rennes
CNRS, IETR-UMR 6164
Rennes, FRANCE
lu.ge@insa-rennes.fr

***Abstract*—In this paper, we introduce the MUSE (MUlti-view Synthesis Enhancer) method, which is an evolution of our previously proposed HDSB method and is based on a hybrid algorithmic-learning-based scheme. MUSE generates novel views of a scene using an autoencoder specifically optimized for refining pre-synthesized views that have been derived from actual observations. Since the subjective test is the ultimate test of the visual rendering quality, we evaluate our proposed method by two subjective tests. Experimental results show that the MUSE brings a global gain compared to two tested state-of-the-art methods.**

***Index Terms*—View synthesis, subjective test, quality assessment**

I. Introduction

To enhance the user experience, advanced works in visual multimedia technologies and computer graphics have enabled the development of new immersive visual medias. In an immersive interactive experience, the "6 degrees of freedom (6DoF)" offers the user the possibility to be tracked not only by his head movement but also his body location as he physically moves left, right, forward, backward up, and down. The user can thus move around to freely change his location and point of view while watching a video.

The video transition is one possible intermediate step towards the 6DoF, where the user is not entirely free but has the choice to change his point of view using a transition that can seamlessly join two adjacent points of view of real cameras in the scene. A typical format for these applications is the Multi-view Video composed of a set of N video sequences representing the same scene, referred to as real views, acquired simultaneously by a system of N cameras positioned under different spatial configurations. An alternative representation is the Multi-View-Plus-Depth (MVD) format [1], where the depth and texture information are used for each viewpoint.

The MPEG (Moving Picture Experts Group) proved a considerable interest in the MVD formats for their capacity to support 3D video applications. They proposed the View Synthesis Reference Software (VSRS), Reference View Synthesis (RVS) and Versatile View Synthesizer (VVS) [2], consecutively. The VVS was released facing to the 6DoF challenges, which improves the precision of the backward texture warping and better preserves edges, compared to previous reference softwares.

Many other methods trying to improve the view synthesis quality have also been proposed in the litterature, such as rendering-algorithmic-based techniques [3], [4], disocclusion inpainting methods [5], [6], learning-based methods [7], [8] or methods based on radiance field estimation [9], [10]. Learning-based and radiance-field-based view synthesis methods achieving encouraging and outstanding results accelerated and condensed the work in this field. However, some of them do not attach importance to the number of input images and the computational/memory cost, and the improvement is mainly performed on small baselines from the target view-point and many input viewpoints.

The larger the distance between the cameras is, the wider the baseline is. Our works focus on developing solutions that deal with large baselines view synthesis while mixing the advantages in algorithmic-based warping method with the benefits of Convolutional Neural Networks (ConvNets) on improving the final rendered image quality. In our previous work [11], we proposed a Hybrid Dual Stream Blender for wide baseline view synthesis (HDSB), where reference views were algorithmically warped to the target position and then blended via a ConvNet, followed by a residual encoder-decoder for image blending with a Siamese encoder to keep the parameters low. However, the HDSB did not guarantee a network generalization due to the lack of variety in the limited available training database.

In this paper we propose an improved version of the HDSB, called MUSE: A Multi-view Synthesis Enhancer. It is still based on our previous idea of a hybrid algorithmic-learning scheme where reference views are preliminarily warped to the target position using an inpainting method built around a mean value to handle occlusions. But the MUSE differs from the HDSB in the following aspects:

- The warped and inpainted left and right reference views are preliminary merged by a sum weighted by a factor α, to increase the contribution of the closest reference view, reducing the impact of the projection of incorrect geometry estimated from distant points of view, and control the view-dependent appearance of non-Lambertian surfaces.

- The masks of the disocclusion areas are preliminary merged by the boolean binary operator "*and*", to generate one mask that contains areas invisible by either the two reference views.
- A concatenation of the generated mask and the merged references form the only one input data considered by the network, and a single encoder can thus be used.
- The network does not blend the views but corrects the image artifacts resulting from the pre-merged image, and improves the quality of the final image.
- The learning process of the network is called "intra-content" and therefore, it is different from the inter-content learning process used for the HDSB. This ensures that the network learns the specific content features of each sequence by over-learning on the available views. Only the triplet images (left warped image, right warped image and the reference view) are considered for a given content.

Another contribution of the paper is that we conducted two subjective tests to evaluate the rendered quality of our method, different from many works where the methods are only evaluated by objective quality metrics which don't always correlate with human perception.

In the following, we describe the proposed method in section II, present the experiments for the performance evaluation in section III, then show and discuss the results in section IV, finally give the conclusions and perspectives in section V.

II. Proposed Method

We illustrate the MUSE architecture in Fig.1, which includes two main stages: a pre-synthesis process and a ConvNet-based view enhancer.

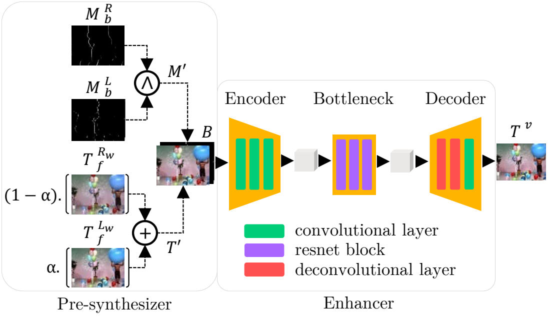


Fig. 1: MUSE architecture illustration, where α corresponds to the normalized distance between the virtual view and two left and right reference views that frame the target view and regulate the reference views' mixing within the pre-synthesized view.

A. Pre-synthesis process

The pre-synthsis process, in the same way as in the HDSB, uses the two reference views to generate the view to be synthesized. The steps leading to the generation of the warped and inpainted left and right reference texture images T_f^{Lw} T_f^{Rw}, and the generation of the binary masks (M^L, M^R) corresponding to the disocclusion area in (T^{Lw}, T^{Rw}), respectively, are the same as those described for the HDSB [11].

But here, the view T' is pre-synthesized by merging the warped ans inpainted texture images T_f^{Lw} T_f^{Rw} by a sum weighted by a factor α:

$$T' = \alpha \cdot (T_f^{Lw}) + (1-\alpha) \cdot T_f^{Rw} \tag{1}$$

We use a factor α that corresponds to the normalized distance between the virtual view and the two reference views and allows to increase the contribution of the reference view that is closest to the virtual one. For example, if the virtual view is very close to the left view, the factor α will be close to 1 and the contribution of the left view will be dominant. The fusion of the binary masks M^{Lb}, M^{Rb} is performed by the boolean operator *and*, and the produced common disocclusion map contains then only the areas that are not observed by either the left or the right view. The tensor B is the result of the concatenation of T' which is a 3-channels color image (in RGB or YUV format) and the mask M' of the disoccluded areas. B is the input to the neural network whose role is to improve the quality of the pre-synthesized view T^v.

B. ConvNet-based view enhancer

The architecture of the neural network enhancer is similar to the architecture used by the HDSB. However, there are some remarkable differences between HDSB architecture and the view-enhancing problem considered here. First, we only need to deal with a single input (the concatenation of the merged reference views with the mask B), rather than four inputs in HDSB, thus only one encoder is used here, instead of two. Second, the pre-synthesized image is characterized by particular types of artifacts due to algorithmic blending/inpainting imperfections (such as ghosting problems shown in Fig.2), instead of disocclusion problems.

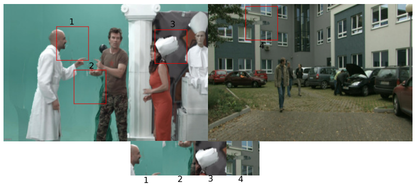


Fig. 2: Illustrations of algorithmic blending/inpainting imperfections in a pre-synthesized image.

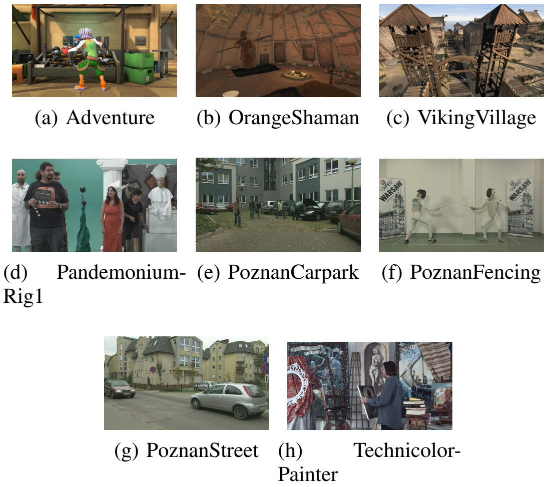
Our network should learn how to enhance the quality of the input blended image by reducing artifacts and ghosting effects and refining the image details using an adapted architecture. Thus, we use an encoder-decoder architecture that uses a single blended image as input and generates one output image and consists of three parts: *encoder*, *bottleneck* and *decoder*. The three parts are the same as in the HDSB architecture as they achieve the best tradeoff between the semantic depth and the spatial resolution of the output feature maps. As well as minimizing training time, we utilize the Mean Squared Error (MSE) loss function.

III. Experiments

For the performance evaluation, we did two subjective tests, one for the perceived quality evaluation of the view synthesis methods and the other for the quantification of the differences in the quality of the image synthesis methods. We compared our new method MUSE, the MPEG reference software VVS and our previously proposed HDSB with the subjective results.

A. Dataset

We used 8 multi-view video sequences here, of which 3 are synthetic (Adventure, OrangeShaman, and Viking Village) and the other 5 are from real video capturing (PandemoniumRig1, PoznanCarpark, PoznanFencing, PoznanStreet, and TechnicolorPainter), as illustrated in Fig. 3. The scenes Adventure and VikingVillage are generated by IRT b<>com from free Unity assets; OrangeShaman, PoznanCarpark, PoznanStreet and TechnicolorPainter are excerpts from the original scenes proposed to the MPEG Immersive Video working group; PandemoniumRig1 was captured by IRT b<>com with the same camera rig described in [12]. Tab. I summarizes the characteristics of the different sequences.


(a) Adventure (b) OrangeShaman (c) VikingVillage

(d) Pandemonium-Rig1 (e) PoznanCarpark (f) PoznanFencing

(g) PoznanStreet (h) Technicolor-Painter

Fig. 3: Illustration of the selected contents.

B. MUSE training settings

For each sequence, we extract three neighbor views from the first 750 frames: the left and right views are used as references (T^L, T^R), and the central view is used as the ground-truth. We firstly warp, inpaint, and blend T^L and T^R to obtain T'. Our network is then trained on the triplet images (T', the corresponding binary mask M' and the ground-truth image).

From each sequence, we extract 75k triplets of co-located patches of size 64×64, and get 52k training patches, 15k validation patches, and 7500 test patches. During the training, we used batches of 128 patches, a learning rate of 0.0001 (leading to the convergence of our learning algorithm after 100 epochs), and Adam optimization algorithm with weight decay $= 0$ and betas = (0.9, 0.999). All the experiments are performed on a server with an NVIDIA RTX2080GPU.

C. Subjective test-bed

A 24-inch LED PC monitor (ASUS VG248QE) is used to manage the stimulis presentation to be evaluated. The voting interface is displayed after.

In order to reproduce video sequences of different frame rates, a screen with V-SYNC/G-SYNC capabilities is used to guarantee a real-time display without visual artifacts. Therefore, the constituted test-bed optimizes the display of uncompressed video sequences (yuv/1080p/8bits format), recorded on SSD and played with “mpv.”s player (an open-source media player software based on ffmpeg [13]).

The test protocol follows the ITU-R BT.500-14 recommendation [14]. Twenty-seven non-expert observers between the age of 23 and 53 participated in the two tests.

D. Two Subjective Tests

Our subjective tests aim to evaluate the visual quality of our method in the case of an actual view synthesis application, which is a *video transition*. A video transition is a technique used in post-production video-editing in order to join two shots together. In our tests, we consider the transition from one point of view to another in the same scene. Therefore, we use our method to create a virtual transition between two cameras in the same scene to change the user’s point of view. To achieve this, we create several intermediate views that are played one after another, such as a ”path” between the starting and ending points. Tab. I shows the scene properties for each sequence to create the transitions, i.e., the number of real cameras in the scene, the baseline... It also shows the number of synthesized images in every scene, necessary to transition between two real cameras, with its duration.

1) Test 1: Evaluation of the perceived quality of the view synthesis methods: Five-scale Absolute Category Rating (ACR) method (cf. P.910 [15]) is adopted in this test, because of the absence of the natural contents.

With 8 different sequences and 3 synthesis methods (VVS, HDSB, and MUSE), we have 24 test transitions in total. For each observer, the order of presentation of the sequences is different, and each sequence is repeated once. Before starting the test, the protocol and the objectives of the test are presented to each participant in a test instruction sheet.

2) Test 2: Quantification of the differences in the quality of the image synthesis methods: This test is more accurate than Test 1 in estimating the differences in rendering between the synthesis methods from a subjective point of view. Therefore, for each scene, it is possible to determine if MUSE offers a better visual rendering than the two other methods and quantify this difference on an appropriate perceptual scale. However, this method does not determine the level of perceived visual quality for each measurement point (a view synthesis method associated with a scene), as was the first subjective test case. The two tests are thus complementary.

Sequence	PandemoniumRig1	PoznanStreet	PoznanFencing	PoznanCarpark	OrangeShaman	TechnicolorPainter	Adventure	VikingVillage
Baseline [m]	0.79	0.14	0.22	0.14	0.2	0.072	0.8	0.8
Horizontal field of view [deg]	36	58	58	58	90	46	66	92
depth n5 [m]	3.7	4.8	0.63	4.6	1.3	2.4	3.1	7.7
Disparity max [pix]	630	49	600	52	150	73	390	97
Number of images	600	250	250	250	300	300	510	510
Length [s]	10	10	10	10	10	10	10	10
Frame rate [image/s]	60	25	25	25	30	30	50	50
Type	natural	natural	natural	natural	synthetic	hybrid	synthetic	synthetic
Number of cameras	10	9	10	9	5	4	22	22
Resolution [pix]	1920 x 1080	1920 x 1088	1920 x 1080	1920 x 1088	1920 x 1080	2048 x 1088	1920 x 1080	1920 x 1080

TABLE I: Summary of the characteristics of each scene and their associated capturing system

Test 2 is similar to the Double Stimulus Continuous Quality Scale (DSCQS) (cf. BT.500 [14]). It requires prior scheduling of the sequences to be tested because the view synthesis methods will be compared in pairs (video A and video B). For a given scene, three pairs are compared: (MUSE vs VVS), (MUSE, HDSB), and (VVS, HDSB). After that each pair is displayed twice, participants are asked to report their 7-scale score on the perceived quality difference (whether the quality of video B is "Much worse," "Less good," "Slightly worse," "Equivalent," "Slightly better," "Better," or "Much better" than video A).

IV. Results and Discussions

A. *From Subjective Test 1*

Tab. II lists the Mean Opinion Score (MOS) obtained by computing the average of the inter-observer and inter-scene scores for each tested method. The ANOVA test is also done and the p-values are alway below the level of 5% of significance. The results demonstrate thus that the performance of MUSE is significantly better than those of VVS and HDSB.

Method	MOS
HDSB	2.4
MUSE	**3.1**
VVS	2.7

TABLE II: MOS for each view synthesis method and the bold font indicates the best result.

If we look further into each scene (cf. Fig. 4), we can find that MUSE is better than VVS and HDSB on all the tested scences, except that VVS performs favorably for the VikingVillage sequence. For this scene, we could observe that prominent, annoying, and time-varying artifacts are predominant in the areas where occlusion alternate with disocclusion zones. We can reasonably hypothesize that the temporal smoothing provided by VVS would bring again compared to the pure spatial processing performed by MUSE. Thus, the MUSE method brings a global gain compared to the VVS method. The taking into account of the temporal neighborhood represents an improvement perspective of the MUSE method.

B. *From Subjective Test 2*

Tab. III lists the inter-observer and inter-scene MOS for each pair of the compared methods, from which we observe that: the quality results of HDSB are relatively equivalent to those of VVS; the quality results of MUSE are better than those of HDSB; the quality results of MUSE are better than those of VVS. Fig. 5 illustrates then the inter-observer MOS for each scene and pair of compared methods.

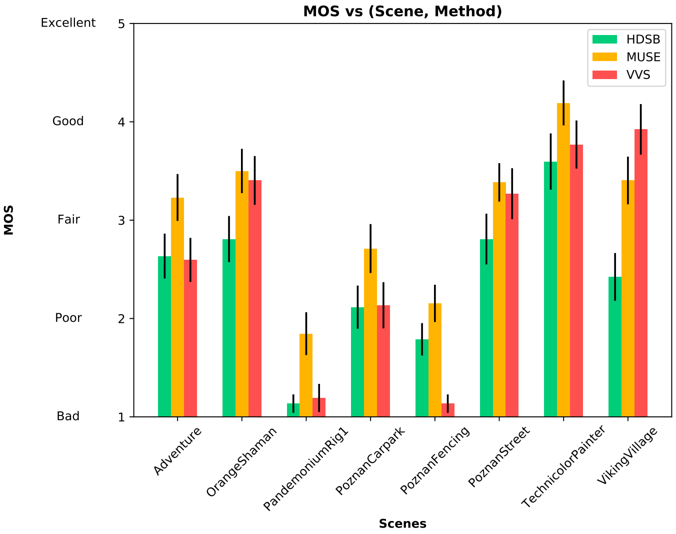


Fig. 4: MOS and 95% confidence intervals obtained for each scene and each view synthesis method.

Compared methods	MOS
HDSB vs VVS	-0.335
MUSE vs HDSB	1.157
MUSE vs VVS	0.780

TABLE III: MOS for each pair of compared methods

Test 2 allows us to quantify the difference in visual quality between the different view synthesis algorithms (HDSB, VVS and MUSE). The analysis of the results corroborates the conclusions of the first test and shows that:

- MUSE performs favorably compared to VVS for all the scenes except VikingVillage. The contributions are most significant for the real scenes PandemoniumRig1 and PoznanFencing. The results for the OrangeShaman and PoznanStreet sequences are relatively comparable
- MUSE is a real improvement over HDSB. The results for PoznanStreet and TechnicolorPainter are slightly better, while the results for the other contents are much better.

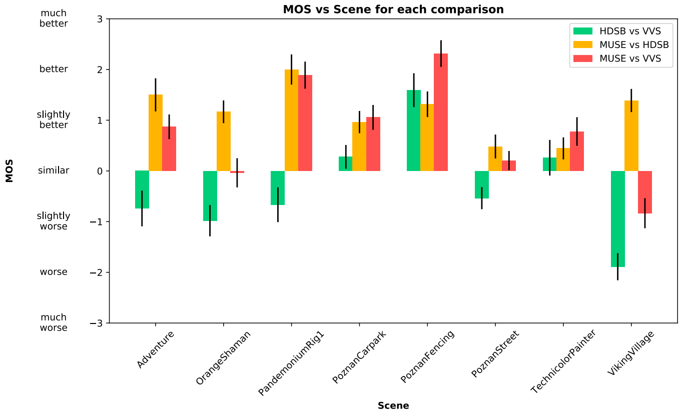


Fig. 5: MOS and 95% inter-user confidence intervals for each scene and pair of compared methods

- HDSB produces similar or worse results compared to VVS, except for the PoznanFencing scene, for which the visual quality of the HDSB method is judged better.

The analysis of the two tests' results also shows that the quality of the view synthesis is strongly dependent on the content. The visible artifacts are related to errors in the depth maps, which are more critical when the disparity amplitude is more significant. As well as the camera focal lengths, the spacing, and the object relative distances in the scene, affect the quality of the synthesized views. The synthetic contents for which ideal depth maps have been used escape this rule. A more detailed study of the link between the configuration of the rig and the quality of the synthesized views would allow us to confirm these hypotheses.

V. Conclusions and Perspectives

In this paper, we propose a new view synthesis method MUSE. We qualitatively evaluate our method by conducting two subjective tests that compare MUSE's performances with two other state-of-the-art view synthesis methods (HDSB and VVS). These tests were performed on an actual view synthesis use-case which is a video transition between two points of view in the same scene.

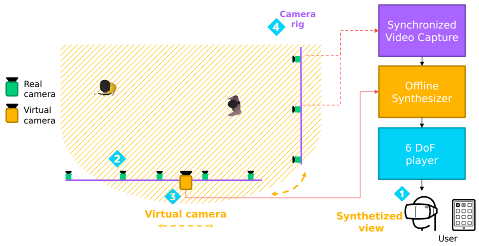


Fig. 6: From multi-view capture to 6DoF rendering

Fig. 6 illustrates the final targeted user experience towards the 6DoF video navigation, that can benefit from our new method. The user (1) visualizes, with the help of a 6-DoF player installed on a device such as a PC, a tablet, or a virtual reality headset, the video sequence associated with a real camera (2). He/she can navigate from one real point of view to another real point of view. Between two real points of view, the images associated with a virtual camera (3) are generated by an off-line synthesizer such as MUSE and played back by the player (1).

Our future research regarding the view synthesis technique aims to exploit the temporal information to improve the hole filling procedure and impose temporal consistency among neighboring frames as an additional constraint at training and inference stages. Since the subjective tests are costly and time-consuming, we only compared the MUSE with two other methods this time. But we will compare our method with more state-of-the-art view synthesis methods in the future.

References

[1] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *2007 IEEE International Conference on Image Processing*, vol. 1, 2007, pp. I – 201–I – 204.

[2] J. Jung and P. Boissonade, "VVS: Versatile View Synthesizer for 6-DoF Immersive Video," Apr. 2020, working paper or preprint. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02541110

[3] J. Lin, W. Wang, J. Yao, T. Guo, E. Chen, and Q. F. Yan, "Fast multi-view image rendering method based on reverse search for matching," *Optik*, vol. 180, pp. 953 – 961, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0030402618319272

[4] S. Prakash, T. Leimkühler, S. Rodriguez, and G. Drettakis, "Hybrid image-based rendering for free-view synthesis," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 4, no. 1, May 2021. [Online]. Available: http://www-sop.inria.fr/reves/Basilic/2021/PLRD21

[5] J. Thatte and B. Girod, "A statistical model for disocclusions in depth-based novel view synthesis," in *2019 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2019, pp. 1–4.

[6] S. Satapathy and R. R. Sahay, "Robust depth map inpainting using superpixels and non-local gauss–markov random field prior," *Signal Processing: Image Communication*, vol. 98, p. 116378, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923596521001752

[7] I. Choi, O. Gallo, A. Troccoli, M. H. Kim, and J. Kautz, "Extreme view synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7781–7790.

[8] T. Volker, G. Boisson, and B. Chupeau, "Learning light field synthesis with multi-plane images: Scene encoding as a recurrent segmentation task," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 633–637.

[9] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision*. Springer, 2020, pp. 405–421.

[10] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics (ToG)*, vol. 41, no. 4, pp. 1–15, 2022.

[11] N. Hobloss, L. Zhang, S. Lathuiliere, M. Cagnazzo, and A. Fiandrotti, "Hybrid dual stream blender for wide baseline view synthesis," *Signal Processing: Image Communication*, vol. 97, p. 116366, 2021.

[12] N. Hobloss, L. Zhang, and M. Cagnazzo, "A multi-view stereoscopic video database with green screen (mtf) for video transition quality-of-experience assessment," in *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2021, pp. 201–206.

[13] MPV, "a free, open source, and cross-platform media player," 2021. [Online]. Available: https://mpv.io/

[14] ITU-R, "Recommendation itu-r bt.500-14: Methodologies for the subjective assessment of the quality of television images," *International Telecommunications Union: Geneva, Switzerland*, 2019.

[15] I. Rec, "P. 910: Subjective video quality assessment methods for multimedia applications," *International Telecommunication Union, Geneva*, vol. 2, 2008.