Image Quality Assessment and Saliency Detection: human visual perception modeling and applications

Lu Zhang

HABILITATION
À DIRIGER DES RECHERCHES

Spécialité : Signal, Image, Vision
Présentée par : **Lu ZHANG**

Image Quality Assessment and Saliency Detection: human visual perception modeling and applications

soutenue le 07/02/2020 devant le jury composé de :

Ce ZHU, Professor, University of Electronic Science & Technology of China, China
Frederic DUFAUX, DR, L2S - CentraleSupelec - Université Paris-Sud, France
Daiqin YANG, Tenured associate professor, Wuhan University, China
Patrick LE CALLET, Professeur, Université de Nantes, France
Olivier LE MEUR, MCF avec HDR, Université de Rennes 1, France
Luce MORIN, Professeur, INSA Rennes, France
Olivier DEFORGES, Professeur, INSA Rennes, France

ACKNOWLEDGEMENT

Via this synthesis document of my research in preparation for my Habilitation degree (Habilitation à Diriger des Recherches), I would like to express my sincere gratitude towards all the doctors I co-supervised: Shishun TIAN, Qiong WANG, Meriem Outtas, Ines Saidi, Maxime Bichon, as well as the PhD students I currently co-supervise: Fang-Yi Chao, Théo Ladune. I am also very thankful for the directors of these theses: Olivier Déforges, Luce Morin, and Kidiyo Kpalma, who helped a lot all along the way. Without them, nothing presented in this manuscript would have been possible.

Thanks also to all the researchers that I met/collaborated during thesis meetings, project meetings and conferences for sharing their knowledges and insightful thoughts. My appreciation also goes to all my colleagues from the VAADER team of IETR, for their friendships, the good working environment and the pleasure time.

I would also like to acknowledge all the foreigner researchers with whom I have closely collaborated, from Southeast University, Shenzhen university, East China University of Technology, Cardiff University, for their friendships, fruitful discussions, and enjoyable collaborations.

Last but not least, I would like to thank my husband and my parents for their unceasing support, as well as my three children whose smiles and loves are worth it all.

LIST OF PUBLICATIONS

Peer-reviewed journal papers [IF=Impact Factor]:

[1] Y. Gu, H. Tang, T. Lv, Y. Chen, Z. Wang, **L. Zhang**, et al. “Discriminative feature representation for Noisy image quality assessment”. Multimedia Tools and Applications; January 2020. [IF=2.101]

[2] A. F. Perrin, V. Krassanakis, **L. Zhang**, V. Ricordel, M. P. Da Silva, O. Le Meur. “EyeTrackUAV2 : a Large-Scale Binocular Eye-Tracking Dataset for UAV Videos”. Drones; 2020.

[3] Q. Wang, **L. Zhang**, W. Zou, K. Kpalma. “Salient video object detection using a virtual border and guided filter”. Pattern Recognition; Volume 97, January 2020, 106998. [IF=5.898]

[4] H. Zhu, D. Tong, **L. Zhang**, et al. “Temporally Downsampled Cerebral CT Perfusion Image Restoration Using Deep Residual Learning”. International Journal of Computer Assisted Radiology and Surgery; October 2019; 1-9. [IF=2.155]

[5] Y. Gao, Y. Song, X. Yin, W. Wu, **L. Zhang**, Y. Chen, W. Shi, “Deep learning-based digital subtraction angiography image generation”. International Journal of Computer Assisted Radiology and Surgery; July 2019; 1-10. [IF=2.155]

[6] **L. Zhang**, I. Saidi, S. Tian, V. Barriac, O. Déforges, “Overview of full-reference video quality metrics and their performance evaluations for videoconferencing application”. J. Electron. Imaging; March 2019; 28(2), 023001. [IF=0.924]

[7] T. Wang, **L. Zhang**, H. Jia. “An effective general-purpose NR-IQA model using natural scene statistics (NSS) of the luminance relative order”. Signal Processing: Image Communication; February 2019; 71:100-109. [IF=2.814]

[8] W. Zou, Z. Zhuang, S. Jiao, **L. Zhang**, K. Kpalma. “Image Steganography Based on Digital Holography and Saliency Map”. Optical Engineering; January 2019; 58(1), 013102. [IF=1.209]

[9] H. Jia, **L. Zhang**, T. Wang. “Contrast and visual saliency similarity-induced index for assessing image quality”. IEEE Access; October 2018; 6:65885-65893. [IF=4.098]

[10] S. Tian, **L. Zhang**, L. Morin, O. Déforges. “A benchmark of DIBR Synthesized View Quality Assessment Metrics on a new database for Immersive Media Applications”. IEEE Transactions on Multimedia; October 2018; 21(5): 1235 - 1247. [IF=5.452]

[11] C. Xiang, **L. Zhang**, Y. Tang, W. Zou, C. Xu. “MS-CapsNet: A Novel Multi-Scale Capsule Network”. IEEE Signal Processing Letters; October 2018; 25(12):1850-1854. [IF=3.268]

[12] M. Outtas, **L. Zhang**, O. Déforges, A. Serir, W. Hammidouche, Y. Chen. “Subjective and Objective Evaluations of feature selected Multi Output Filter for Speckle reduction on Ultrasound Images”. Physics in Medicine and Biology; August 2018; 63(18):185014. [IF=3.03]

[13] S. Tian, **L. Zhang**, L. Morin, O. Déforges. “NIQSV+: A No Reference Synthesized View Quality Assessment Metric”. IEEE Transactions on Image Processing; December 2017; 27(4):1652-1664. [IF=6.79]

[14] Y. Chen, **L. Zhang**, W. Yuan, et al. “Extended PCJO for the detection-localization of hypersignals and hyposignals in CT images”. IEEE Access; June 2017; 5(99):24239-24248. [IF=4.098]

[15] T. Wang, **L. Zhang**, H. Jia, et al. “Multiscale contrast similarity deviation: an effective and efficient index for perceptual image quality assessment”. Signal Processing: Image Communication; April 2016; 45:1-9. [IF=2.814]

[16] T. Wang, **L. Zhang**, H. Jia, Y. Kong, B. Li, H. Shu. “Image quality assessment based on perceptual grouping”. Journal of Southeast University (English Edition); March 2016; 32(1):29-34. [IF=0.12]

[17] **L. Zhang**, C. Cavaro-Ménard, P. Le Callet. “An overview of model observers”. Innovation and Research in BioMedical engineering (IRBM); June 2014; 35(4):214-224. [IF=0.934]

[18] D. GE, **L. Zhang**, C. Cavaro-Ménard, P. Le Callet. “Numerical Stability issues on Channelized Hotelling Observer under different background assumptions”. Journal of the Optical Society of America A (JOSA A); April 2014; 31(5):1112-1117. [IF=1.861]

[19] **L. Zhang**, B. Goossens, C. Cavaro-Ménard, P. Le Callet, D. GE. “Channelized model observer for the detection and estimation of signals with unknown amplitude, orientation, and size”. Journal of the Optical Society of America A (JOSA A); November 2013; 30(11):2422-32. [IF=1.861]

[20] **L. Zhang**, C. Cavaro-Ménard, P. Le Callet, J. Tanguy. “A Perceptually relevant Channelized Joint Observer (PCJO) for the detection-localization of parametric signals”. IEEE Transactions on Medical Imaging; October 2012; 31(10):1875-1888. [IF=7.816]

International conference papers:

[1] J. Chen, **L. Zhang**, C. Bai, K. Kpalma. “Review of Recent Deep Learning Methods for Image-Text Retrieval”. IEEE 3rd International Conference on Multimedia Information Processing and Retrieval (MIPR), April 2020, Shenzhen, China. (Invited Paper)

[2] Y. Zhang, **L. Zhang**, W. Hammidouche, O. Deforges. “Key Issues for the Construction of Salient Object Datasets with Large-Scale Annotation”. IEEE 3rd International Conference on Multimedia Information Processing and Retrieval (MIPR), April 2020, Shenzhen, China.

[3] A. F. Perrin, **L. Zhang**, O. Le Meur. “How well current saliency prediction models perform on UAVs videos?”. Computer Analysis of Images and Patterns (CAIP), September 2019, Salerno, Italy.

[4] I. Saidi, **L. Zhang**, O. Déforges, V. Barriac. “Laboratory and crowdsourcing studies of lip sync effect on the audio-video quality assessment for videoconferencing application”. ICIP, September 2019, Taipei, Taiwan.

[5] S. Tian, **L. Zhang**, L. Morin, O. Déforges. “SC-IQA: Shift compensation based image quality assessment for DIBR-synthesized views”. Visual Communications and Image Processing (VCIP), December 2018, Taichung, Taiwan.

[6] F. Chao, **L. Zhang**, W. Hammidouche, O. Déforges. “SalGAN360: Visual Saliency Prediction on 360 Degree Images with Generative Adversarial Networks”. ICME2018, July 2018, San Diego, California, USA.

[7] S. Tian, **L. Zhang**, L. Morin, O. Déforges. “Performance comparison of objective metrics on free-viewpoint videos with different depth coding algorithms”. SPIE Optical Engineering + Applications, August 2018, San Diego, California, USA.

[8] I. Saidi, **L. Zhang**, O. Déforges, V. Barriac. “Machine learning approach for global no-reference video quality model generation”. SPIE Optical Engineering + Applications, August 2018, San Diego, California, USA.

[9] M. Bichon, J. Le Tanou, M. Ropert, W. Hammidouche, L. Morin, **L. Zhang**. “Low Complexity Joint RDO of Prediction Units Couples for HEVC Intra Coding”. Picture Coding Symposium (PCS), June 2018, San Francisco, California, USA.

[10] M. Outtas, **L. Zhang**, O. Déforges, W. Hammidouche, A. Serir. “Evaluation of No-reference quality metrics for Ultrasound liver images”. QoMEX, May 29 - 31, 2018, Sardinia, Italy.

[11] L. Leveque, H. Liu, S. barakovic, J.B. Husic, A. Kumcu, L. Platisa, M. Martini, R. Rodrigues, A. Pinheiro, M. Outtas, **L. Zhang**, A. Skodras. “On the Subjective Assessment of the Perceived Quality of Medical Images and Videos”. QoMEX, May 29 - 31, 2018, Sardinia, Italy.

[12] M. Bichon, J. Le Tanou, M. Ropert, W. Hammidouche, L. Morin, **L. Zhang**. “Low Complexity Joint RDO of Prediction Units Couples for HEVC Intra Coding”. ICASSP, April 2018, Calgary, Alberta, Canada.

[13] S. Tian, **L. Zhang**, L. Morin, O. Déforges. “A full-reference Image Quality Assessment metric for 3D Synthesized Views”. Image Quality and System Performance Conference, at IS&T Electronic Imaging 2018, 28 January - 1 February 2018, Burlingame, California, USA.

[14] M. Outtas, **L. Zhang**, O. Déforges, A. Serir, W. Hammidouche. “Multi-output speckle reduction filter for ultrasound medical images based on multiplicative multiresolution decomposition”. ICIP, September 2017, Beijing, China.

[15] Q. Wang, **L. Zhang**, K. Kpalma. “Fast filtering-based temporal saliency detection using minimum barrier distance”. ICME2017W, July 2017, Hong Kong, China.

[16] T. Xu, **L. Zhang**, Y. Chen, H. Shu, L. Luo. “Quality Assessment Based on PCJO for Low-dose CT Images”. International Conference on Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine (Fully3D), June 2017, Xian, China.

[17] I. Saidi, **L. Zhang**, O. Déforges, V. Barriac. “Evaluation of single-artifact based video quality metrics in video communication context”. QoMEX, May 31 - June 2, 2017, Erfurt, Germany.

[18] M. Bichon, J. Le Tanou, M. Ropert, W. Hammidouche, L. Morin, **L. Zhang**. “Inter-block dependencies consideration for intra coding in H.264/AVC and HEVC standards”. ICASSP, March 2017, New Orleans, USA.

[19] S. Tian, **L. Zhang**, L. Morin, O. Déforges. “NIQSV: A No Reference Image Quality Assessment Metric for 3D Synthesized Views”. ICASSP, March 2017, New Orleans, USA.

[20] M. Outtas, **L. Zhang**, O. Déforges, W. Hammidouche, A. Serir, C. Cavaro-Menard. “A study on the usability of opinion-unaware no-reference natural image quality metrics in the context of medical images”. 8th International Symposium on signal, Image, Video and Communications, November 2016, Tunis, Tunisia.

[21] I. Saidi, **L. Zhang**, V. Barriac, O. Déforges. “Audiovisual quality study for videotelephony on IP networks”. IEEE Workshop on Multimedia Signal Processing (MMSP), September 2016, Montreal, Canada.

[22] I. Saidi, **L. Zhang**, O. Déforges, V. Barriac. “Evaluation of the performance of ITU-T G.1070 model for packet loss and desynchronization impairments”. QoMEX, June 2016, Lisbon, Portugal.

[23] I. Saidi, **L. Zhang**, O. Déforges, V. Barriac. “Interactive vs. non-interactive subjective evaluation of IP network impairments on audiovisual quality in videoconferencing context”. QoMEX, June 2016, Lisbon, Portugal.

[24] **L. Zhang**, C. Cavaro-Ménard, P. Le Callet, D. GE. “A multi-slice model observer for medical image quality assessment”. ICASSP, April 2015, Brisbane, Australia.

[25] T. Wang, H. Shu, H. Jia, B. Li, **L. Zhang**. “Blind Image Quality Assessment Using Natural Scene Statistics in the Gradient Domain”. Asia Modelling Symposium, September 2014, Kuala Lumpur, Malaysia.

[26] C. Cavaro-Ménard, **L. Zhang**, P. Le Callet. “QoE for Telemedicine: Challenges and Trends.” SPIE Optics + Photonics, August 2013, San Diego, USA.

[27] **L. Zhang**, C. Cavaro-Ménard, P. Le Callet. “Evaluation of HVS models in the application of medical image quality assessment.” IS&T/SPIE Electronic Imaging, January 2012, California, USA.

[28] **L. Zhang**, C. Cavaro-Ménard, P. Le Callet. “Key issues and specificities for the objective medical image quality assessment.” VPQM, January 2012, Scottsdale, Arizona, USA.

[29] **L. Zhang**, C. Cavaro-Ménard, P. Le Callet. “Using AUC to study perceptual difference model suitability for the detection task on MR image.” MIPS XIV Conference, August 2011, Dublin, Ireland.

[30] **L. Zhang**, C. Cavaro-Ménard, P. Le Callet. “The effects of anatomical information and observer expertise on abnormality detection task.” Proc. SPIE Medical Imaging, volume 7966, February 2011, San Diego, USA.

[31] C. Cavaro-Ménard, **L. Zhang**, P. Le Callet. “Diagnostic quality assessment of medical images : Challenges and trends.” 2nd European Workshop on Visual Information Processing (EUVIP), July 2010, Paris, France.

French conference papers:

[1] I. Saidi, **L. Zhang**, O. Déforges, V. Barriac. “Evaluation de la qualité audiovisuelle des applications de visiophonie”. CORESA, May 2016, Nancy, France.

[2] M. Outtas, A. Serir, O. Déforges, **L. Zhang**. “Réduction de bruit multiplicatif dans les images ultrasons basées sur la décomposition multiplicative multiresolution (MMD)”. CORESA, May 2016, Nancy, France.

[3] M. Schmidt, C. Cavaro-Ménard, **L. Zhang**, J. Tanguy, Le Callet. “Evaluation par observateur numérique basé tâche de la qualité d’IRM compressées”. CORESA, May 2016, Nancy, France.

[4] **L. Zhang**, C. Cavaro-Ménard, P. Le Callet. “Pertinence des modèles du SVH pour la sensation et la perception des images médicales.” Gretsi, September 2013, Brest, France.

[5] **L. Zhang**, C. Cavaro-Ménard, P. Le Callet. “Modèle numérique pour l’évaluation objective de la qualité d’images médicales 2D.” Gretsi, September 2013, Brest, France.

The regularly updated list of my publications is available on:
`http://luzhang.perso.insa-rennes.fr/publications/`.

TABLE OF CONTENTS

INTRODUCTION

This report, which constitutes a synthesis document of my research in preparation for my Habilitation degree (Habilitation á Diriger des Recherches), presents my research works since September 2013, when I joined the VAADER team of the IETR laboratory at INSA Rennes as a teacher-researcher (“Maître de conférences”).

Since I worked on the quality assessment of medical images based on human perception modeling during my PhD thesis (10/2009-11/2012), I have been going down along this research theme (human perception modeling and its applications), which did not exist yet in the VAADER team. The creation of this new theme is also desired by our team, since it has a strong relationship with two of the four main topics of our team:

- image analysis and understanding;
- video representation and compression.

Note that whatever a new image analysis algorithm or a new video compression method is, it is necessary to use a quality assessment method to validate this new technique and "convince" the end-users. Meanwhile, the saliency information can benefit a wide range of applications related to the main topics of our team, e.g. image quality assessment, image segmentation, image compression, image rendering, object detection and recognition, visual tracking, etc. Thus I strengthened and extended the skills in this theme within the team, by mainly working on the **image quality assessment** and **saliency detection**. I taxonomically present my research themes and works in Figure 1, as well as the postdocs, PhD students, master students I co-supervised (with a supervision rate $> 33\%$) on each topic.

For **image quality assessment,** I have been working on four image/video types:

- Medical images: there are two schools for the medical image quality assessment, while one is the task-based approach, e.g. the anthropomorphic model observer (MO), the core of my PhD work; the other is the adaptation of natural image quality metrics. I then co-supervised three master students to further extend the MOs I proposed in the PhD thesis to more medical modalities and applications;

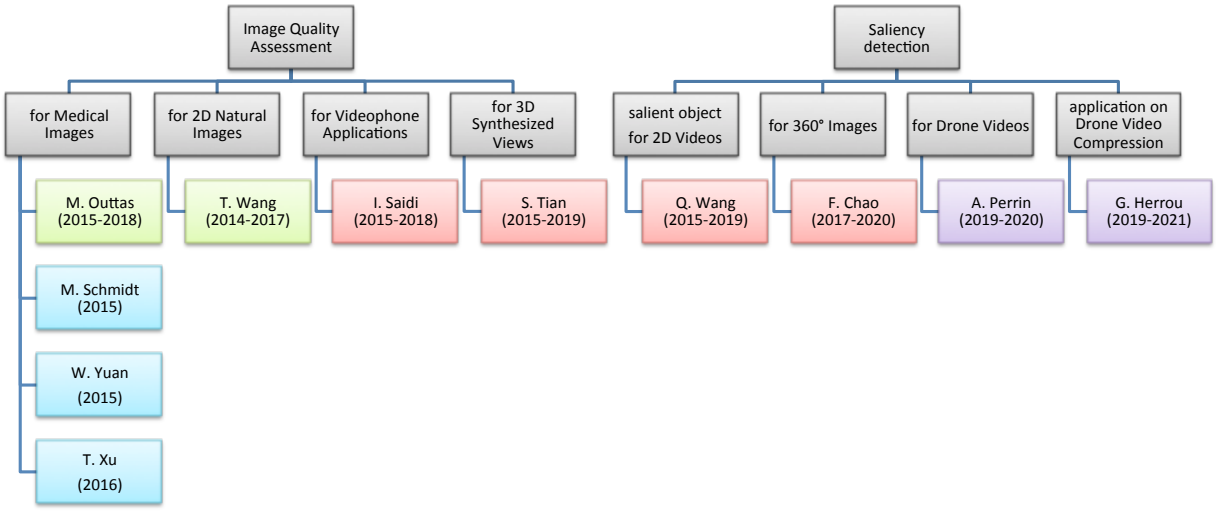


Figure 1: Research work synthesis, where postdoc supervisions are indicated in purple, PhD supervisions are in red, non-official PhD supervisions (or collaborations) are in green, and non-official master supervisions are in blue.

as well as one PhD student to explore the adaptation of natural image quality metrics.

- with M. Schmidt, we worked on the application of the MO on the compression of Magnetic Resonance (MR) images with JPEG2000. Three compression ratios were used and five radiologists had done the detection-localization diagnostic task on the compressed images. The experimental results showed that the model observer worked in a similar way as the radiologists. This work is collaborated with my two PhD thesis supervisors and *Angers Hospital in France*. (1 conference paper is published and 1 journal paper is in preparation.)
- with W. Yuan, we worked on the mathematical extension of the perceptually relevant channelized joint observer (PCJO) to hyposignal task. While the abnormality can appear as a hypersignal or a hyposignal for different imaging modalities, sequences or organs, no MO has been proposed for the hyposignals detection-localization task in the literature. To improve the clinical relevance of the existing MO, we extended the PCJO's capacity from hypersignal-only to both hypersignal and hyposignal. This work is collabo-

rated with Prof. Chen from Southeast University and radiologists from *Nanjing First Hospital in China*. (1 journal paper is published.)

  - with T. Xu, we worked on the application of the MO on the comparison of two different low-dose CT image reconstruction algorithms, a practical need of Prof. Chen from Southeast University and radiologists from *Nanjing First Hospital*. Since the abnormalities often appear as hyposignals on CT images, e.g. the hepatocellular carcinoma (HCC) - the target pathology here, we recurred to the extended PCJO proposed in the previous work. (1 conference paper is published.)
  - A limitation of the MOs I proposed was that they need reference images. In our studies, we took images captured from healthy people as the reference images, which are not always possible and very time-consuming and expensive. With M. Outtas (Algerian PhD, but working within the VAADER team from 2015 to now), we explored firstly the usability of several no-reference metrics proposed originally for natural images in the context of medical images. Then a modified Naturalness Image Quality Evaluator (NIQE) was proposed, and applied on the comparison of different speckle noise reduction methods and the image restoration oriented compression for ultrasound images. The speckle noise reduction methods were also evaluated by radiologists from *Nanjing First Hospital*. Experimental results showed that the modified metric performed better than the direct use of other no-reference metrics proposed originally for natural images, but there is still plenty of room for improvement. (4 conference papers and 1 journal paper are published.)

- 2D natural images: with T. Wang, we collaborated from 2014 when he was still a PhD student in Southeast University until now when he is a lecturer in East China University of Technology. Two full-reference metrics (based on perceptual grouping and multi-scale representation) and two no-reference metrics (based on natural scene statistics) have been proposed during our collaboration. (2 conference papers and 2 journal papers are published.)

- Videos in videophone applications: with I. Saidi, we conducted 2 laboratory experiments and 1 crowdsourcing study for the subjective evaluation of audiovisual quality in the videophone applications. We also compared the state-of-the-art full-reference objective video metrics, as well as distortion-specific no-reference

metrics in this context. In this report, several important works issued from this thesis will be presented in detail in Chapter 2. This is the first industry-oriented doctoral thesis (CIFRE PhDs) funding I got and the 1st thesis I co-supervised officially. (7 conference papers and 1 journal paper are published.)

- 3D synthesized view images: with S. Tian, funded by China Scholarship Council, we constructed a database focusing on the synthesizing distortions with more recent synthesizing algorithms, and proposed 2 full-reference objective metrics and 2 no-reference objective metrics. The details will be presented in Chapter 3. (4 conference papers and 2 journal papers are published.)

Concerning **saliency detection**, I actually worked on both salient object detection and saliency detection (the two notations will be differentiated in Chapter 4) for different applications:

- Salient object detection for 2D videos: with Q. Wang, funded by China Scholarship Council, we firstly proposed a traditional method where the low-level features and priors are hand-crafted; then extended this method using a deep network which achieved much better performances compared to its traditional version. The usage of the deep network is based on our comparative/benchmarking study of the state-of-the-art deep learning based methods. The details will be presented in Chapter 5. (1 conference paper is published and 2 journal papers are submitted.)

- Saliency detection for 360° (or omnidirectional) images: with F. Chao, funded by the Ministry of France, we began the study of saliency detection methods for 360° images from October 2017. In her first year of PhD thesis, we got the 1st place in ICME Grand Challenge "Prediction of Head+Eye Saliency for 360 Images" using a model based on a Generative Adversarial Network. We then extended this model using multi-resolutional Field of View and adaptive weighting for the model training. In the rest one year and a half of her PhD study, we will further exploit the saliency detection for 360° videos. The details will be presented in Chapter 6. (1 conference paper is published and 1 journal paper is submitted up to now.)

- Saliency detection for drone videos and its application on the drone video compression: from 2018 to 2021, I'm leading the project "Automated Saliency Detection from Operators' Point of View and Intelligent Compression of Drone videos",

funded by the ANR (French National Agency for Research) ASTRID (Specific Support for Defence Research Projects and Innovation). This project allows our three partners to recruit a postdoc for each. I co-supervise two of them: with A. Perrin, who started from 01/2019, we are studying the usability of state-of-the-art 2D video saliency detection methods on the drone videos. Considering the specificities of the drone videos (including the bird-point-of-view, the loss of pictorial depth cues...), a new model is certainly necessary and is the objective of this task; with G. Herrou, who will start from October 2019, we will work on the compression of drone video using the saliency as the guidance, i.e. allocate more bits on the saliency parts and less on other parts to optimize the rate-distortion. (1 conference paper and 1 journal paper are submitted up to now.)

This HDR report is organized into three parts, of which the first two correspond to my two research orientations mentioned above. But only four works issued from the four theses I co-supervised officially will be detailed, as exemplars, in this report. Each part begins with a short review of background material and then presents two works related to the research orientation in detail. At the end of each work, we give the list of our scientific contributions, as well as the perspectives. My research project for the further works will be given in the last part. Note that, to ease the reading, parts as well as chapters are self-contained.

PART I

Image quality assessment

CHAPTER 1

IMAGE QUALITY ASSESSMENT BASICS

The image quality assessment (IQA) occupies a very important position in numerous image processing applications, e.g. image acquisition, compression, transmission, restoration, etc. Since human beings are the ultimate receivers of the visual stimulus, the ultimate test is the subjective IQA in which the image quality is evaluated by a panel of subjects (human observers or participants). In this report, we focus on the IQA for natural images, where the average of the values obtained from human observers is known as Mean Opinion Score (MOS). The subjective IQA tests are however costly and time-consuming, thus it is also necessary to develop objective IQA models which can perform similarly to human observers and output quality scores closely related to the MOS.

1.1 Subjective IQA basics

1.1.1 Common elements in test protocols

The implementation of a subjective IQA test must comply with the ITU recommendations [1, 2, 3] to ensure the reliability and reproducibility of the test. Although they are intended for different measurements, the standardized methodologies that we present share some common experimental protocols. These are the panel of observers, the test environment and the global conduct of the sessions.

For the same observed sequence, the evaluation is not stable from one individual to another. Several factors are responsible for this, such as the state of fatigue, knowledge of the images, the observer's general experience in the IQA, or personal appreciation. In our subjective tests mentioned in this report we use only non-expert observers, i.e they are not confronted with the IQA in their professional activity. All participants are firstly examined for their visual acuity through the Snellen test and their color perception defects through the Ishihara test. The observers passing our test should have a visual

acuity of 10/10 for both eyes with or without correction. Moreover, we made sure that all the subjects reported having a normal audition. For greater reliability of the results, a panel larger than 15 participants will give statistically usable results [4, 5].

All the subjective tests mentioned in this report were performed in the laboratory environment, conforming with the ITU recommendations [1, 2, 3], including the general environment, the viewing conditions, and the device calibrations. We placed the display screen in a distance equal to $3 \times H$ (screen height) from the subjects and adapted the ambient brightness of the rooms in order to limit the glare and the visual fatigue of the observers; in order to calibrate our display devices, we have used a tool to neutralize the display defects of the screen and to automatically adjust the hardware settings (brightness, contrast, white point, etc.) so that the display device ensures that it displays the widest range of possible colors.

Before the main test session, we should give an instruction to observers and conduct a training session with them. The instruction is an explanation of the type of methodology, the scoring system, the presentation protocol and any useful elements. A training session is often conducted before the main test session with a few typical conditions to anchor the judgment of the observers. The scores of this training session are not taken into account in the final results. The main test session consists of a variable number of images, which corresponds to the evaluation of a perceived quality under different conditions. For videos, the test sequences have generally a duration between 8 and 10 seconds in order to leave a sufficient time for observers to give a stable score.

1.1.2 Test Methodologies

There are several test methodologies proposed in the ITU recommendations [1, 2, 3]: Absolute category rating (ACR) [P.910]; ACR with hidden reference (ACR-HR) [P.910]; Degradation category rating (DCR) [P.910]; Double-stimulus continuous quality-scale (DSCQS) [BT.500]; Pair Comparison (PC) [P.910]; Subjective Assessment of Multimedia VIdeo Quality (SAMVIQ) [BT.1788]; Single stimulus continuous quality evaluation (SSCQE) [BT.500]... We will only detail the methodologies used in the subjective tests mentioned in this report in this section.

ACR : The ACR (also called single stimulus) method is a category judgment where the test images/sequences are presented one at a time and are rated independently on

a category scale. The method specifies that after each presentation the subjects are asked to evaluate the quality of the sequence shown. This evaluation is performed on a five- or nine-grade categorical scale that is explained by five items (Excellent-Good-Fair-Bad-Poor). An illustration of 5-grade is given in Figure 1.1. The ACR method is an inexpensive method from the point of view of its application, treatment and analysis of the results. It also has the advantage of being able to qualify test systems and obtain their ranking according to the level of quality associated with them.

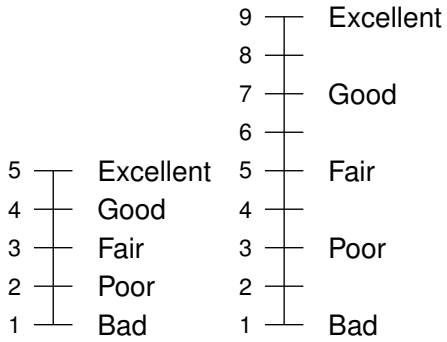


Figure 1.1: Scale of quality assessment (MOS) at 9 and 5 levels.

SAMVIQ : In the SAMVIQ protocol, there is much more freedom for the observers who can view each image several times and correct the notation at any time they want. The observers can compare the degraded versions with each other, as well as with the explicit reference. In each trial, there is also a hidden reference which helps to evaluate the intrinsic quality of the reference when the perceived quality of the reference is not perfect. Each observer moves a slider on a continuous scale graded from 0 to 100 annotated by 5 quality items linearly arranged (excellent, good, fair, poor, bad). The SAMVIQ results have a greater accuracy than the ACR scores for the same number of observers (on average 30% fewer observers were required for SAMVIQ than ACR for the same level of accuracy) [6].

1.1.3 Data Analyses

Subjects screening During a subjective quality assessment, a significant amount of data is collected. It is then necessary to carry out some tests before translating this data into results. Thus, inter-observer coherence is evaluated. As a result of this

verification, the assessment scores of some observers may be rejected. This screening step can therefore be critical in obtaining the results of a methodology since it requires a minimum number of observers. The screening method defined in BT.1788 [3] has been used for our subjective tests in this report.

Opinion Scores Once the tests are performed, the results are analyzed and combined in a single note per image (or video sequence) describing its average quality. In this report, for the subjective tests conducted using ACR, we calculated the MOS, i.e. the average of the quality scores over the total number of participants; for the subjective tests conducted using SAMVIQ, we used the Differential Mean Opinion Score (DMOS), which is the difference between the score of the hidden reference image and the score of the tested image. The confidence interval associated with each MOS/DMOS score was also calculated. In order to easily compare each observer's opinion about the quality of images, a linear transform that makes the mean and variance equal for all observers is often employed. The outcome of such transform is called Z-score.

Statistical test In order to evaluate if there is a significant difference between different test conditions on the quality perception, statistical tests should be used. The data should firstly be checked to see if the samples are normally distributed, e.g. using the Shapiro-Wilk test or Kolmogorov-Smirnov test. If the hypothesis of the normal distribution is not rejected, a parametric test can be used; otherwise, a non-parametric test should be used. In this report, for the former case, the Fisher test is used; for the latter case, the Mann-Whitney test is used.

1.2 Objective IQA basics

1.2.1 Classifications of IQA methods

According to the ITU studies [7, 8], objective metrics may be classified into five main categories depending on the type of input data:

- Media-layer models use the audio or video streams to evaluate the perceived quality. For these models the characteristics of the stream content and decoder

strategies such as error concealment are usually taken into account. The model ITU-T J.247 [9] for video quality assessment belongs to this category.

- Parametric packet-layer model use only the packet header (TCP, RTP, UDP, IP, etc.) information without having access to the media signal. Such models are well suited for in-service non-intrusive multimedia quality monitoring. Among this category we may indicate the Recommendation ITU P.1201 [10].

- Parametric planning models use the quality planning parameters (bandwidth, packet loss rate, delay, frame rate, resolution, etc.) for network and terminals to predict the quality. For example, the models G.1070 [11] and G.1071 [12] are parametric models for estimating video and audio qualities for video-telephony and streaming applications respectively. The E-model (Rec. G.107) is a planning model for audio quality.

- Bitstream-layer models predict the QoE based on both encoded bit stream and packet-layer information without performing a complete decoding. These models can be used in situations where one does not have access to decoded video sequences. The Recommendations ITU P.1202[13] and P.1203 [14] are bitstream layer models for video and audiovisual media streaming quality assessment.

- Hybrid models are a combination of two or more models from the preceding. These models analyze the media signal, the bitstream information and packet header to estimate the perceived quality. For instance, ITU J.343 [15]is on of the developed hybrid models.

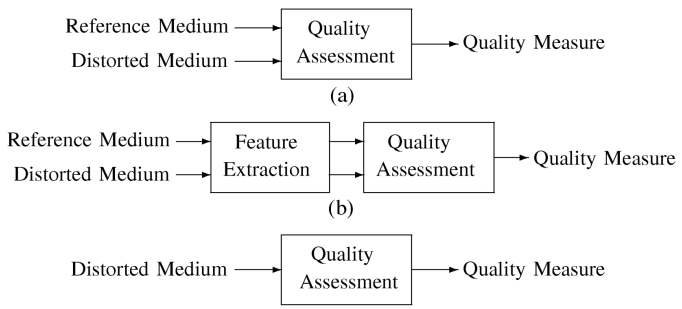


Figure 1.2: Overview of media layer models, figure from [16]

As illustrated in Figure 1.2, the media-layer objective quality assessment methods can be further categorized as full-reference (FR), reduced-reference (RR), and no-reference (NR) depending on whether a reference, partial information about a reference, or no reference is used in assessing the quality, respectively. As explained in [16], full- and reduced-reference methods are important for the evaluation in non-real-time scenarios where both (1) the original (reference) data or a reduced feature data set, and (2) the distorted data are available. For instance, during the development and prototyping process of video transport systems, the original video can be delivered offline for full-reference quality assessment at the receiver, or the received distorted video data can be reliably (without any further bit loss or modifications) delivered back to the sender. In contrast, for real-time quality assessments at the receiver without availability of the original video data, low-complexity reduced-reference or no-reference methods are needed.

1.2.2 Performance evaluation metrics

While there are more metrics can be used for the performance evaluation of the objective IQA metrics, we only introduce here the three commonly used metrics when subjective scores are available [17]:

1. Accuracy prediction: refers to the ability to predict the subjective quality ratings with low error. The Pearson Linear Correlation Coefficient (PLCC) was computed. For two datasets $X = \{x_1, x_2, ..., x_N\}$ and $Y = \{y_1, y_2, ..., y_N\}$ with $\overline{x}$ and $\overline{y}$ the means of the respective datasets, the PLCC is defined by:

$$PLCC = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} \tag{1.1}$$

2. Monotonicity prediction: refers to the degree to which the relationship between the subjective quality ratings and the predicted measure can be described by a monotone function. The Spearman Rank Order Correlation Coefficient (SROCC) was used:

$$SROCC = \frac{\sum(X_i - X')(Y_i - Y')}{\sqrt{\sum(X_i - X')^2}\sqrt{\sum(Y_i - Y')^2}} \tag{1.2}$$

   where X_i and Y_i are the ranks of the ordered data series x_i and y_i respectively; X' and Y' denote the respective midranks.

3. Consistency prediction: measures the ratio of wrong predicted scores by the objective model to the total number of scores. The Root Mean Square Error (RMSE) was computed. For a dataset $\{x_1, x_2, ..., x_N\}$, with $\overline{x}$ is the mean value:

$$RMSE = \sqrt{\frac{1}{N}\sum(x_i - \overline{x})^2} \tag{1.3}$$

The PLCC and RMSE are often computed after performing a non-linear mapping on the objective measures using the cubic polynomial mapping function recommended by the VQEG [17]. This function is used in order to fit the objective model scores to the subjective scores.

CHAPTER 2

PERCEIVED QUALITY FOR VIDEOPHONE APPLICATION

2.1 Introduction

Over the years, multimedia applications have conquered many segments of the telecommunications industry. We are dealing today with multimedia services in many areas, starting with the various digital television systems, video-telephony, video-on-demand (VOD), Internet Protocol television (IPTV) or simply video-sharing services like YouTube or Dailymotion. Multimedia services represent an important part of the global IP traffic that is constantly growing. In the last statistics reported in [18], mobile video services will generate three quarters of mobile data traffic by 2020. Among the most popular multimedia services, the video conversational applications are in full development. In a competitive market, various Over The Top (OTT) players are emerging: Skype, Messenger, Facetime, WeChat, Duo, etc. For example, the statistics show that Skype has more than 300 million monthly active users [19] with 3 billion minutes per day spend on Skype video calls [20].

The development of appropriate methods for measuring and monitoring the perceived quality of these new services becomes thus a major challenge for telecommunication operators. This chapter presents several of our efforts to investigate the issues concerning the perceived quality for videoconferencing application, collaborated with Orange.

- Wa have conducted two subjective experiments to assess the perception of video conference service users under different conditions, and to constitute a sequences database to evaluate the performance of the objective quality metrics. We investigate the video, audio and audiovisual quality and asynchrony perception under two different situations: a non-interactive and an interactive conversational one.

We analyze the effects of network impairments (packet loss, jitter, delay) on perceived audiovisual, audio and video quality. We evaluate the impact of experimental context and scene complexity on the quality perception in case of video calls. Furthermore, we propose new acceptability thresholds of audio-video asynchrony in video telephony context and study the effect of synchronization in the presence and absence of network degradation.

- We surveyed the most recently developed objective full-reference (FR) video quality metrics and we evaluate their prediction accuracy on four different subjective databases: the Live Mobile video quality Database, the EPFL database and two videoconferencing databases developed by Orange Labs. We investigate the assessment of video quality in the context of videoconferencing and video call communications.

2.2 Non-interactive subjective test

In this section, we present a non-interactive audiovisual quality assessment experiment. Before our study, there were few studies addressing the impact of the network settings on perceived multimedia quality [21], and the combination of the network impairments and non-synchronized audio and video had not been well studied.

2.2.1 Experimental set-up and recording

In order to generate our test database, we used a PC-based video conferencing software internally developed in Orange. The reason of this choice is that it allows sharing multimedia contents between two users and separating audio and video IP flows. Thus, we were able to simulate degradations on audio and video independently. We used the audio-visual communication protocol H.323 recommended from the ITU [22] to transmit calls between two users.

To simulate network degradations, we used the NetDisturb software [23] which allows disturbing flows over IP network by generating user-defined impairments (latency, jitter, packet loss ...). The interest of using a network simulator, instead of the real network, in our experience was to totally control the IP network degradation, provide repeatable QoS on audio and video flows using predefined configuration mode and

values, and re-create real world problems in the laboratory. We inserted a machine equipped with NetDisturb between our two clients connected via an Ethernet local network.

Once a conference call was set up, the client sender transmitted the original audio, video or audiovisual files to the receiver (see Fig. 2.1). Then, we controlled the packets transmission between them by adding packet losses, jitter and delay. At the receiver side, we recorded the degraded sequences and captured IP packets traveling over the network (pcap format). To ensure a perfect playback, all recorded multimedia sequences were processed and stored as raw YUV 4:2:0 for the video stream and uncompressed Pulse Code Modulation (PCM) for the audio stream.

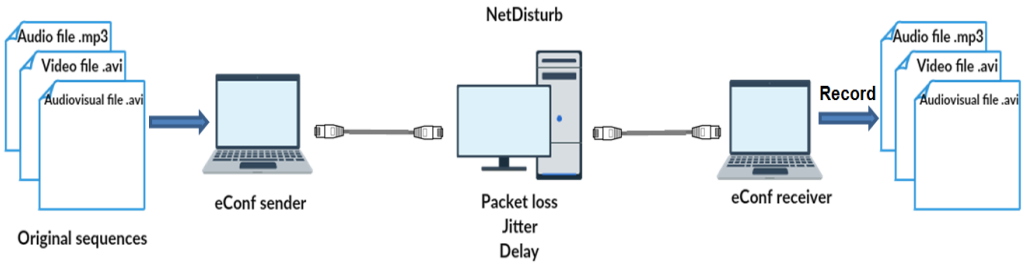


Figure 2.1: Simulation platform design.

2.2.2 Test conditions

The distortions we simulated reflect the range of IP network impairments including packet loss, jitter and delay, of which the tested values are listed in Table 2.1, where a negative value of asynchrony means that the audio stream is delayed according to the video and a positive value means that it is advanced.

Video packet loss VPL (%)	0, 0.5, 1, 2
Audio packet loss APL (%)	0, 2, 5, 20
Video jitter (ms)	0, 60
Audio jitter (ms)	0, 30
Audio-video asynchrony (ms)	-400, -250, -150, 0, +50, +150, +400

Table 2.1: Experiment parameters

Figure 2.2: Frame captures from the original sequences.

2.2.3 Methodology and test protocol

Six sequences (Restaurant, Desk, Sofa, Poster, Hall and Park, cf. Fig. 2.2) were selected to represent different contexts of real life video calls and different levels of spatial and temporal complexities. The duration of the sequences is 8 -10 seconds. The experimental conditions are summarized in Table 2.2.

Video		**Audio**	
Codec	H.264/AVC (constrained baseline)	Codec	AMR Wideband
Bit rate	768 kbps	Bit rate	23.85 kbps
Resolution	VGA (640×480)	Channels	1
Frame rate	15 fps	Sampling frequency	48000 Hz
GOP size	10 frames		
Video color scheme	16 bit YUV (4:2:0)		

Table 2.2: Experimental conditions used in the subjective study

The experiment was organized in three sessions as detailed in Table 2.3. The ex-

perimental method was the 5-grade Absolute Category Rating (ACR). A total of 30 subjects (13 male, 17 female) participated in the experiment. We realized the audio-only and the video-only test with 15 subjects while the audiovisual test was carried out with the other 15 subjects. They were provided with a high quality headphone (Stax SR-404) for sound reproduction. The experiment was performed in an acoustically treated room especially designed for audio and video quality tests. The signals were presented to the subjects via an LCD computer monitor with a 1024×768 resolution. The evaluation score was indicated on a tablet next to the screen on the right of the subjects.

Test	Duration	Sequences	Conditions	Outputs
Audio only	10min	36	5	MOS_A
Video only	10min	36	5	MOS_V
Audiovisual	1h30	176	33	MOS_{AV} MOS_A^{AV} MOS_V^{AV} MOS_{synch}

Table 2.3: Three test sessions, where MOS_{synch} is the audio-video asynchrony score.

2.2.4 Results and discussions

The test results were summarized by computing the averaged MOS values for each test condition over the six sequences. No subject was excluded after the screening. The Mann-Whitney U test was used for the significance test, because our data does not follow the normal distribution.

Audio-video quality Interaction The plots in Fig. 2.3 show the MOS scores averaged over all sequences for both test sessions. They demonstrate that the experiments have been properly designed, as the subjective rates uniformly span over the entire range of quality levels. By plotting MOS_V vs. MOS_V^{AV} and MOS_A vs. MOS_A^{AV}, and calculating their linear correlation coefficients ρ, we noticed that the perceived audio and video qualities are weakly influenced by the audiovisual context. Mann-Whitney test results revealed that there is no significant difference between scores of the two sessions (audio-only and video-only vs. audiovisual).

We are also interested in studying the mutual interaction between the individual audio and video streams. A statistical test revealed that in an audiovisual context the

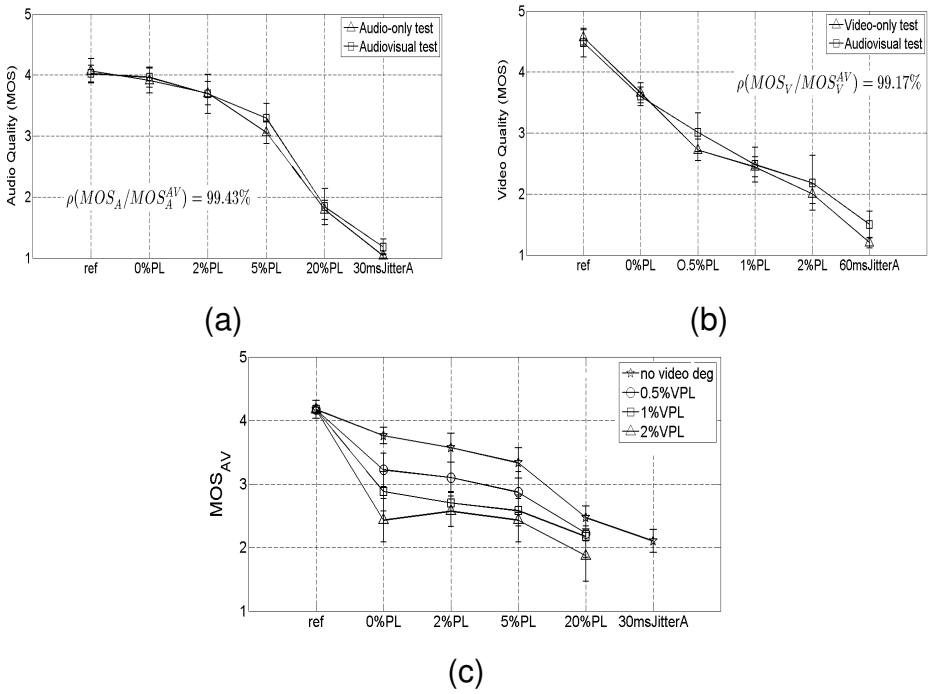
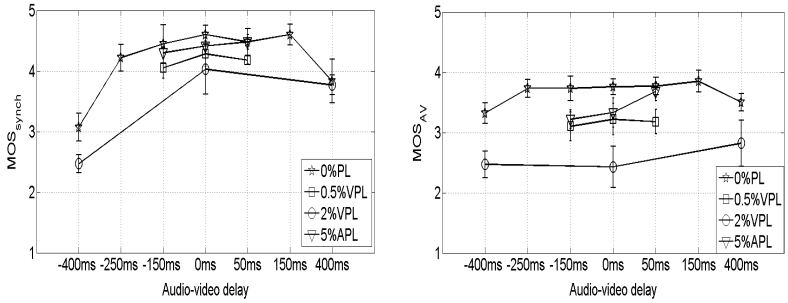


Figure 2.3: Mutual interaction between audio (a) and video (b) qualities and the impact of audio and video quality on overall audiovisual quality (c).

impact of the video impairments on the perceived audio quality is not significant. On another hand, the audio impairments have a small impact on the perceived video quality. For the same video quality level, MOS_V^{AV} values decrease slightly with the percentage of audio packet loss. This drop in MOS scores is more significant in the case of good and average video quality levels (0%VPL, 0.5%VPL). When the video quality is already poor (1%VPL and 2%VPL), quality judgment is not affected by the audio degradation (there is not a significant difference).

Fig. 2.3 (c) shows the interaction between audio and video quality levels in influencing the overall audiovisual quality. The presented results were averaged over all delays (synchronous and not synchronous contents) and over all contents. It reveals that for the same audio quality levels, decreasing the video quality generally results in inferior audiovisual ratings. Alongside, for the same video quality, decreasing the audio quality generally results in inferior audiovisual ratings. The impact of video impairments on audiovisual quality at good audio quality level is more significant than at poor and bad audio quality levels. Concerning the jitter condition, it had the biggest impact on decreasing the audiovisual quality.



(a) (b)

Figure 2.4: Synchronization acceptability chart.

Audio-video Synchronization Limited by the number of conditions and the duration of the test, we could not cross the 6 different values of delay with all the network degradation levels which explained the lack of some points on Figure 2.4. But we are still able to identify thresholds of synchronization acceptability. We set $MOS_{synch} = 4$ and $MOS_{AV} = 3$ as the acceptability limits, corresponding to thresholds when the subjects begin to be disturbed and when the audiovisual quality becomes poor. For audio delayed with more than 250 ms and advanced with more than 150 ms, the desychronisation becomes annoying and the audiovisual quality decreases. We also notice that the presence of video packet loss impairments have a little, but not significant impact on synchronization. The perception of audiovisual quality and synchronization is sensitive to network degradation mainly related to video streams.

2.3 Interactive subjective test

To be closer to a real-life video-conferencing, we present an interactive conversational subjective test in this section. Before our study, few studies have been conducted for evaluating the audiovisual quality in conversational context [24, 25, 26].

2.3.1 Experimental set-up and recording

Figure 2.5 depicts this test platform. User PC1 and user PC2 are two identical systems, placed in two separate rooms and connected via a local Ethernet IP network. Other experimental settings are the same as in the non-interactive test.

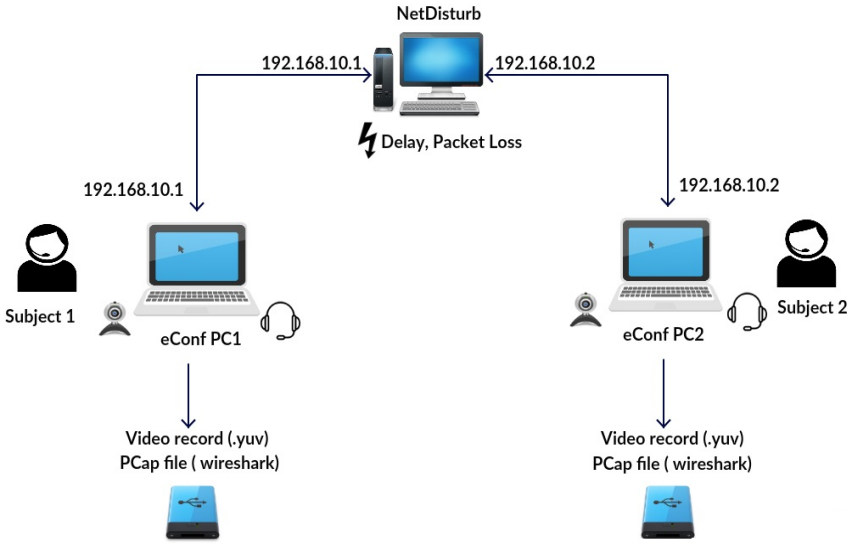


Figure 2.5: Simulation platform design.

2.3.2 Test conditions

Since the interactive takes longer time, we have less conditions here. We only generated two levels of audio and video packet loss which represent the extreme ranges of quality: 1-hardly perceptible, 2-highly annoying. All conditions were symmetric so that the test participants experienced the same quality on both ends of the connection. We randomized the order of the conditions. Table 2.4 provides an overview of the transmission parameters evaluated in this study.

Video packet loss VPL (%)	0, 0.5, 2
Audio packet loss APL (%)	0, 5, 20
Audio Delay AD (ms)	0, 250, 400
Video Delay VD (ms)	0, 150, 400

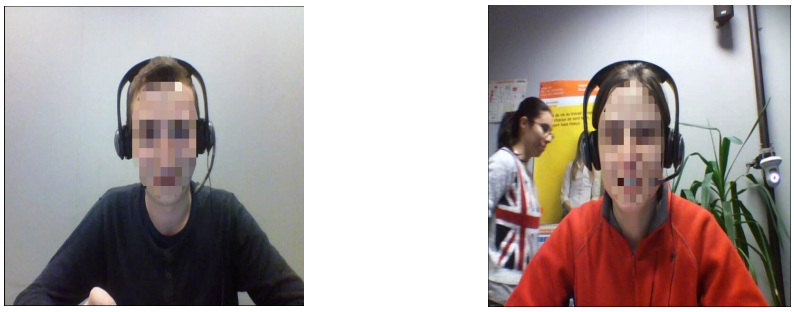
Table 2.4: Experiment conditions

2.3.3 Methodology and test protocol

In order to consider the influence of scene complexity and keep the experiment time within limits, we have configured the two rooms with two different levels of video complexity:

- **Room 1**: where the background behind the subject is a simple white wall (cf. Fig. 2.6.a).
- **Room 2**: where the scene has a certain spatial and temporal complexity. A poster and a plant behind the subject, and one Orange staff walk behind him from time to time (cf. Fig 2.6.b).

The rooms have been acoustically treated and they have a similar audio background.



(a) (b)

Figure 2.6: Screen captures of the conversation in Room 1 (a) and Room 2 (b).

The test has been conducted in an interactive scenario. We proposed a game to stimulate the conversation between the two subjects. For a subject, the objective of the game, was to let its partner guess a word without using the word itself or five additional words listed on a card. We gave each subject 20 cards. This conversation task is similar to the Name-Guessing task from the ITU-T Recommendation P.920 [27]. The subjects could also discuss on their own topic if they prefer. The duration of each conversation was around 3 or 4 minutes. Each discussion corresponds to a specific set of impairments of audio and video. The subjects tested 9 different conditions where the audio and video impairments are independent (limited by the test duration, the interaction between the conditions was not tested).

Twenty subjects (9 male, 11 female) participated in the experiment. They were all inexperienced in evaluating audiovisual quality in such a context, but the majority had already experienced a video-conference call. Subjects were asked to rate the perceived overall audiovisual quality (MOS_{AV}), audio quality (MOS_A) and video quality (MOS_V) and the audio-video synchrony annoyance (MOS_{synch}), using the 5-grade ACR.

2.3.4 Results and discussions

The results of the subjective experiment are summarized by averaging the scores assigned by the panel of participants for each conversation. For the comparison between the interactive and non-interactive experiments and between the scenes with different complexity, the Mann-Whitney U test [28] is used since the data does not follow the normal distribution. We set the significant difference level to $\alpha = 0.05$. Our screening results show that no subject has to be excluded.

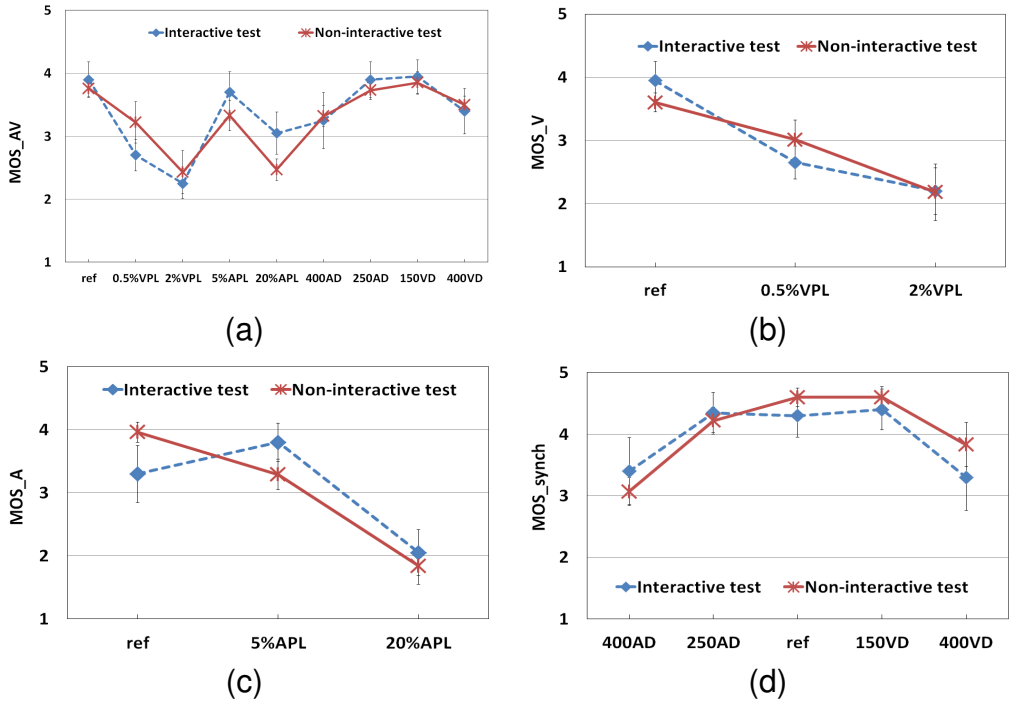


Figure 2.7: Interactive vs. non-interactive MOS scores

Influence of the experiment context: interactive vs. non-interactive We firstly investigate the influence of the experimental context (non-interactive vs. interactive)

on audiovisual quality (AVQ), video quality (VQ), audio quality (AQ) and audio-video synchronization acceptability. Figure 2.7 shows the MOS scores obtained in the two contexts averaged over all scenes.

By comparing the two plots in Figure 2.7.a, we observe that there is a significant difference between MOS_{AV} scores in case of 0.5% video packet loss and 20% audio packet loss. This may indicate that subjects are more sensitive to low video impairments when they communicate than when they passively watch an audiovisual sequence. The interactive task may make the subjects discriminant and severe in the assessment of the audiovisual quality since the video impairments may have more psychological impact on the visual communication they are involved in. However, for important VPL (2%) the quality is poor enough that there is not a significant difference between subjective scores in the two contexts. The subjects give a significantly higher quality note in the interactive context than in the non-interactive context when the audio quality is very low. This may indicate that their attention on the audiovisual quality judgment may be diverted by the guessing game.

For the perceived video quality, Figure 2.7.b shows no significant difference between the two contexts. Subjects perception of the video quality and concentration on the artifacts are the same. Nevertheless, we report a significant difference of perceived audio quality between the two contexts (Figure 2.7.c). Considering the variances, we note that for the interactive test there was not a significant difference between reference and 5%APL condition, while for the non-interactive there was this significant difference – indicating that the impairments are more noticeable in the non-interactive context. The reason of this variance may be that the audio impairments are more noticeable when the subjects are just viewing and listening to an audiovisual content. Then, they are more concentrated and they are more able to notice the impairments.

Concerning the thresholds of desynchronization acceptability, as it can be seen in Figure 2.7.d, there is not a significant difference between the two test contexts.

These results may be true for our tested conditions where only one source (video or audio) was impaired at the time (no interaction between the conditions). Previous studies showed that when both audio and video were impaired, differences in MOS ratings were found [25].

Influence of scene complexity for interactive test We are also interested in the influence of the scene complexity on multimedia quality and audio-video synchronization

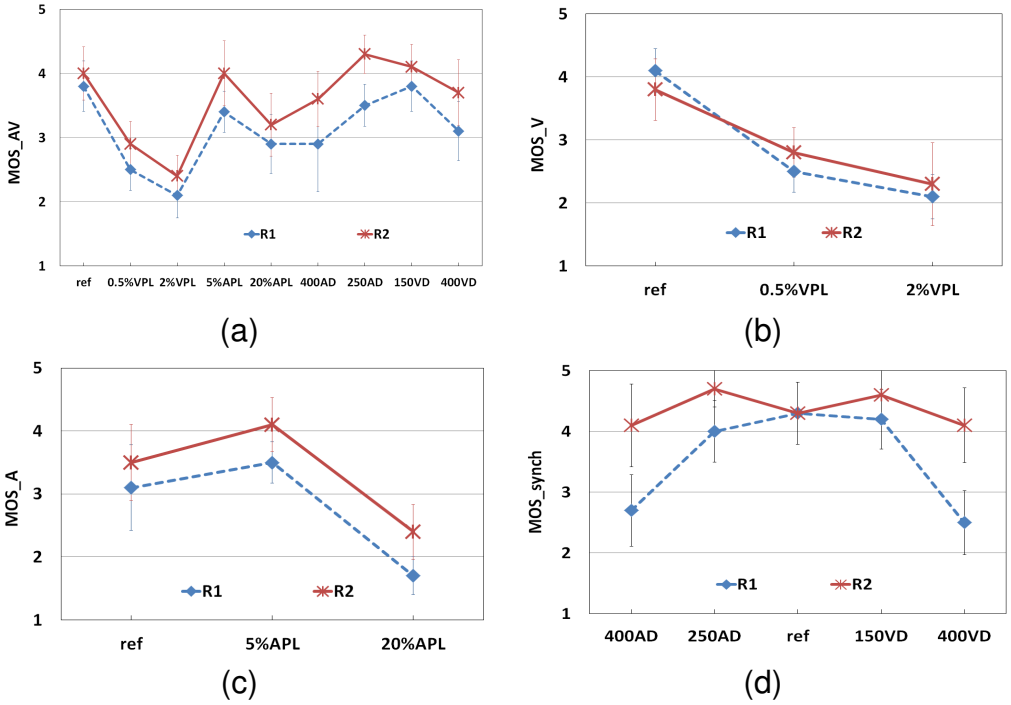


Figure 2.8: Impact of scene complexity for interactive experiment context.

acceptability for the interactive test.

Figure 2.8 shows the MOS_{AV}, MOS_V, MOS_A and MOS_{synch} scores associated to 95% confidence intervals, according to the quality condition and scene complexity. "R1" denotes the perception of the complex scene of Room 2 from Room 1; and "R2" denotes the perception of the simple scene of Room 1 from Room 2.

We can see that generally the perceived AVQ is higher in a simple scene than that in a complex scene at the same degradation levels (an average drop of MOS_{AV} score is about 0.5). The statistical test reveals that there is a significant difference between subjective MOS_{AV} scores for the two rooms. This may indicate that when a scene is composed of complex spatial and temporal elements (presence of high frequencies in the picture and high amount of temporal activity), the network impairments would have a greater impact, and the artefacts (block loss and blockiness) would be more visible. In fact, scenes with high temporal and spatial complexities require more bit rate to be encoded. At a constant bandwidth, more encoding artefacts will occur and the efficiency of the packet loss concealment algorithm is reduced [29].

For the video quality (Figure 2.8.b), there is not a significant difference between Room1 and Room2. We have expected to have a significant difference in the results

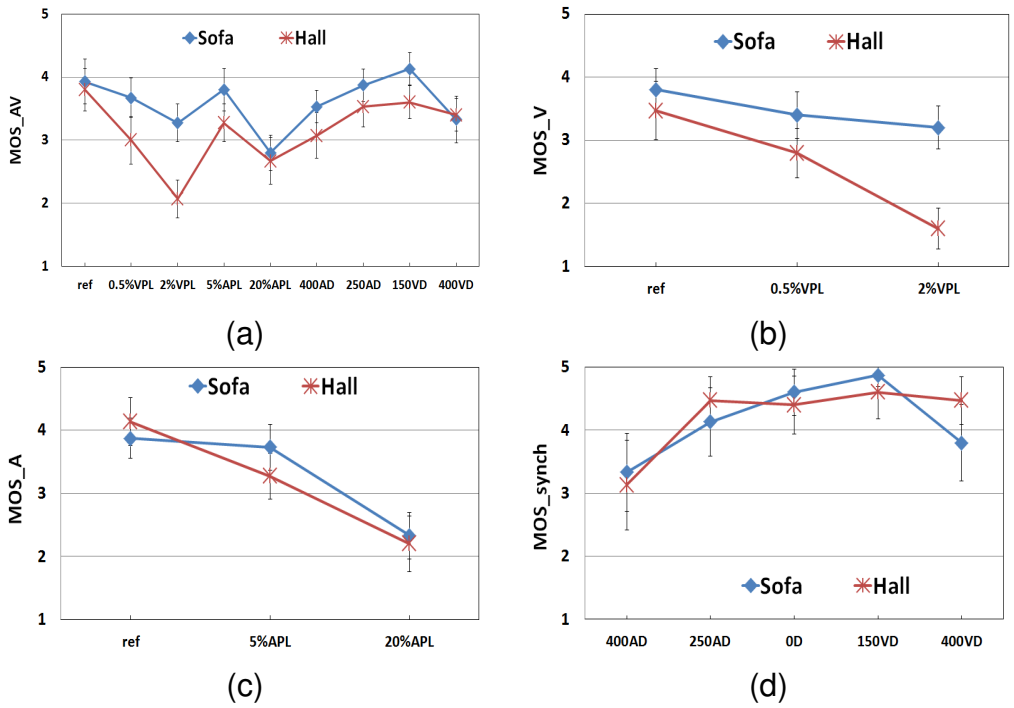


Figure 2.9: Impact of scene complexity for non-interactive experiment context.

because the complexity of the scenes is guessed to have a stronger impact on video quality than on audio quality. This may be explained by the fact that the difference of complexity between the scenes is not sufficient to have an impact on the perceived video quality.

For the perceived audio quality (Figure 2.8.c), there is no significant difference between the results for the two rooms. This is logic since the spatial complexity is not expected to have an effect on audio quality. Furthermore, the audio background deployed in our experiment was the same when it comes to the both rooms used. Thus, the used audio background did not allow to reveal any impact in this case.

Figure 2.8.d shows that the synchronization annoyance of the subjects is also influenced by the spatial and temporal complexity of the perceived scene. The differences in MOS_{synch} are statistically significant. These plots of synchronization acceptability are coherent with the MOS_{AV} results even if the difference of MOS_{synch} between the two rooms is more important. This may indicate that an increased temporal activity has a direct impact on perceived lip synchronization since the movements disturb subject concentration.

Influence of scene complexity for non-interactive test In order to stay coherent with the conversational test we present in this part a comparison between subjective results of the "Sofa" and the "Hall" scene. We chose these sequence scenes due to the difference of spatial and temporal complexity between them and to the similarity they have with the interactive scene content.

In Figure 2.9, MOS_{AV}, MOS_V, MOS_A and MOS_{synch} scores are represented and associated with 95% confidence intervals, according to the quality condition for each scene. We report a significant difference in MOS_{AV} scores between the sequences for all the test conditions except the reference, 20%APL and 400 ms video delay. Thus, compared with Figure 2.9.a, we deduce that overall quality perception is influenced by the complexity of the perceived scene in both interactive and non-interactive context. This observation affirms that the environment and the position of the person on the video call is a parameter to take into account to evaluate the perceived communication quality. This complexity impact could be studied through a non-interactive experiment.

For the video quality, there is a significant difference in MOS_V (Figure 2.9.b). In fact, the subjective scores of the complex scene ("Hall") are lower than that of the simple scene. This observation is justified by the fact that video artifacts caused by packet losses are more visible with sequence complexity. We notice that the SI difference between the two scenes here is much bigger than that in the interactive context. This may explain why we did not observe a significant difference in MOS_V in the interactive context.

As it can be seen in Figure 2.9.c there is not a significant difference of audio score between the two scenes. This result is expected since scene complexity has not an effect on audio quality perception, and consistent with the finding in the interactive context.

Figure 2.9.d shows that the subjects' reaction to desynchronization annoyance is the same for the two scenes, no significant difference is noticed. Thus, unlike the interactive context, in a non-interactive context the scene complexity does not impact audio-video synchronization perception. Previous studies have shown that in a passive context, large delay in the audiovisual signals does not necessarily impact the quality perception as test subjects accommodate for it [30].

2.4 Comparative study of existing FR VQA metrics

While the POLQA model has been widely accepted within Orange as the objective audio quality metric, objective video metrics have not yet been tested for the videoconferencing application. In this section, we present a survey of recently developed objective FR video quality assessment (VQA) metrics and we evaluate their prediction accuracy on four different subjective databases: the Live Mobile video quality Database, the EPFL database and two videoconferencing databases developed by Orange Labs. We investigate the assessment of video quality in the context of videoconferencing and video call communications. We choose these subjective database because of the nature of the contents and the variety of the simulated impairment types: transmission error (packet loss, jitter, freezing, etc.), coding (variable bit rates), and frame rate. Note that these metrics had not been compared with each other before this study.

2.4.1 Tested FR VQA metrics

The selected algorithms which we studied have been widely cited in the literature and reported to have good performances. Moreover, the authors of the selected metrics have released the source codes. Therefore, the presented results are easy to reproduce. The ten FR video quality assessment metrics described in the following subsections include Peak Signal to Noise Ratio (PSNR), structural similarity index (SSIM) [31], multi-scale structural similarity index (MS-SSIM) [32], video quality metric (VQM) [33] (including its general model and videoconferencing model), Open Perceptual Video Quality metric (OPVQ) [34], motion-based video integrity evaluation (MOVIE) [35] , ViS3 [36], SSIMplus [37] and video multi-method assessment fusion (VMAF) [38]. Objective MOS prediction metrics are also standardized by the ITU (J. series recommendations) to assess the video quality. It would be interesting to compare their prediction accuracy with the diverse full reference metrics. However, we do not have access to these models because of their commercial licenses. For instance, the model J.247 is owned by the company OPTICOM. We introduce in our study an open source implementation of this model named OPVQ [34] which does not implement the temporal alignment part of J.247 due to its patent. In Tab. 2.5 we summarize the characteristics of the surveyed metrics.

Table 2.5: Comparison of the charasterisctics of the full reference objective metrics

Metric	**Approach**	**Temporal Pooling method**	**Value Range**	**Execution time (normalized based on PSNR)**	**Implementation**
PSNR	Mean square error measurement	Mean over the frames	[0, 100]	1	MSU software [39]
SSIM [40]	Structural distortion measurement	Mean over the frames	[0, 1]	1.05	MSU software [39]
MS-SSIM [32]	Multi-scale structural distortion measurement	Mean over the frames	[0, 1]	2	MSU software [39]
VQM [33]	Edge impairment filter	Compute Temporal Information (TI)	[0, 1]	30	NTIA software [41]
MOVIE [35]	Gabor filter bank	Temporal distortions index	[0, 1]	456	Source Code [42]
ViS3 [36]	detection-based and appearance based strategies of the MAD algorithm	Spatiotemporal dissimilarity index	[0, 100]	23	Matlab code [43]
SSIMplus [37]	Contrast sensitivity function	Mean over the frames	[0, 100]	4	SSIMwave software [44]
VMAF [38]	Machine Learning	Temporal information among the elementary metrics	[0, 100]	26	Source code [45]
OPVQ [34]	ITU-T J.247	Mean over the frames	[1,5]	19	OpenVQ Toolkit [46]

2.4.2 Test databases

To investigate the performance of the above mentioned video quality metrics, we compare the results on different subjective databases. With the aim of having large test database with diversified conditions and impairment types which may be encountered in videoconferencing applications, we selected four databases to work on: The LIVE

mobile video quality database [47, 48, 49], the EPFL video database [50, 51] and two Orange internal subjective test databases [52].

The Orange 1 database is obtained from our non-interactive test (cf. section 2.2). The Orange 2 database is an internal Orange database, of which the initial aim is to study the influence of coding bit rate and frame rate on the perception of video quality in the context of videoconferencing and video-calling services. Three reference audio-visual sequences are used. They were uncompressed (YUV 4:2:0 with pixel depth of 8 bits) in VGA resolution. These sequences have different levels of complexity in terms of movement, details and texture. The video part of these sequences was encoded using H.264 (Baseline and High profiles) and H.265/HEVC codecs. Six video coding bit-rates (64, 128, 256, 384, 576 and 768 kbps) were adopted in order to be representative of the common use case conditions. The test sequences were displayed on a Nexus 6 phablet (5.96", 1440×2560).

Table 2.6: Properties of subjective VQA databases

	Live Mobile	**EPFL**	**Orange 1**	**Orange 2**
Nbr. of sequences	200	78	30	95
Nbr. of references	10	6	6	3
Resolution	HD 1280×720	CIF	VGA	VGA
Duration	10 s	8 to 10 s	8 to 10 s	10 s
Frame rate	30 fps	30 fps	15 fps	15 and 30 fps
Distortion types	H.264 - 4 diff. bitrates wireless packet loss frame freezes rate adaptation temporal dynamic	packet loss	Packet loss Jitter	H.264 High profile H.264 Baseline profile H.265 encoding
Encoder	H.264 AVC	H.264 AVC	H.264 AVC	H.264 and HEVC
Assessment method	SSCQE with HR	SS	ACR	ACR
Subjective scores	DMOS $[0, 5]$	MOS $[0, 5]$	MOS$[1, 5]$	MOS$[1, 5]$
Nbr. of subjects	36	40	15	22

2.4.3 Performance comparison

We evaluate the performance of the full reference metrics under study using three statistical indicators (PLCC, SROCC and RMSE). The performance of all metrics are summarized in Table 2.7.

All three statistical measures (PLCC, SROCC and RMSE) show that generally three metrics (i.e. SSIMplus, ViS3 and VMAF) outperform the other metrics. The common characteristic of these metrics is that they are video metrics that include the movement

Table 2.7: Statistical correlations of full reference metrics with the MOS scores (the best three performing metrics are highlighted in bold font for each test database and each criterion).

	PSNR	SSIM	MS-SSIM	VQM-G	VQM-V	OPVQ	MOVIE	Vis3	SSIMplus	VMAF
EPFL database										
PLCC	0,88	0,89	0,89	0.90	0.89	0.91	0,87	**0,92**	**0,93**	**0,91**
SROCC	0,87	0,91	0,92	0.88	0.90	0.89	0,87	**0,90**	**0,92**	**0,92**
RMSE	0,68	0,66	0,65	0.61	0.65	0.60	0,71	**0,58**	**0,54**	**0,55**
Live Mobile database										
PLCC	0,71	0,65	0,65	0.83	0.82	**0.85**	0,71	0,84	**0.84**	**0,86**
SROCC	0,65	0,60	0,65	0.79	0.77	**0.82**	0,64	0,75	**0.76**	**0,77**
RMSE	0,62	0,66	0,66	0.50	0.52	**0.52**	0,61	0,52	**0.46**	**0,45**
Orange database 1										
PLCC	0,72	0,79	**0,81**	0.69	0.72	0.66	0,74	**0,85**	**0,79**	0,22
SROCC	0.68	0,71	**0,77**	0.72	0.74	0.67	0,72	**0,82**	**0,74**	0,23
RMSE	0.45	0.46	**0,46**	0.49	0.46	0.51	0,53	**0.42**	**0,48**	0,68
Orange database 2										
PLCC	0.48	0.52	0.48	0.55	0.58	0.57	0.73	**0.74**	**0.81**	**0.82**
SROCC	0.57	0.63	0.62	0.32	0.37	0.54	0.53	**0.91**	**0.75**	**0.76**
RMSE	0.61	0.60	0.61	0.53	0.51	0.61	0.54	**0.52**	**0.41**	**0.43**

information in their quality assessment algorithms. On the other hand, classic image based metrics (PSNR, SSIM and MS-SSIM) are the least correlated with the subjective video quality judgment.

By comparing the two VQM models (NTIA general and videoconferencing model) there is no significant difference between the correlation values when they are applied on typical distortion errors in video transmission (appearing in the Live Mobile and EPFL databases). But VQM Videoconferencing model outperforms the NTIA General model when the video contents are close to a videoconferencing context where subjective scores are more influenced by this context (cases in the Orange1 and Orange2 databases). On the other hand, we note that both VQM_G and VQM_V models performs worse when video sequences are encoded with H.265/HEVC (cf. results on the Orange2 database). This may be interpreted by the optimization of these models for video sequences encoded with H.263 and MPEG-4 [53].

Even though the OPVQ model provides support for only a limited set of spatial resolutions (VGA, CIF and QCIF) and has been tested and validated for VGA resolution only, our correlation results prove that it could be applied on HD sequences(cf. their good performances on the EPFL and LIVE databases). Furthermore, the coefficient parameters of the OPVQ model are trained on a data set containing quality impairments related to H. 264, H. 264/SVC and MPEG-4 coding, transmission errors, temporal dynamics (switches in video coding bit rates during the sequence). We find thus that it has a good performance (a correlation value of 85%) for the cases including

these degradations, such as in the Live Mobile database.

However, the model has lower performances for cases with new degradation types (e.g. jitter, HEVC coding, frame rate changes) for which the model was not trained, cf. the results on the two Orange databases.

The main strength of MOVIE is video quality estimation according to motion trajectories. The metric is accurate in detecting distortions that appear in regions containing movements. This explains its good performances for the EPFL and Orange 1 databases. In fact, it is known that unlike application distortions (coding, frame rate, resolution, etc.) independent from the content, the transmission impairments (in particular the packet loss) infect objects on movement (which do not belong to the scene background).

A previous review [16] in 2011 showed that MOVIE had the best correlation with subjective opinions on LIVE video quality database, before the appearance of Vis3, SSIMplus and VMAF. The major drawback of MOVIE is its extremely high calculation complexity. MOVIE is the most complex metric in our experiment, which needs much more time than any other metric. This prevents its practical use in operational context.

Results shown in Table 2.7 reveal that Vis3 is competitive against the other metrics. The spatio-temporal dissimilarity estimation based on the video decomposition into spatio-temporal slices (STS) makes the algorithm less sensible to the temporal alignment between the reference and the degraded sequences. In fact, due to the videoconferencing software and the recording process used to generate the Orange 1 database, we notice a slight misalignment in frames of the reference and those of the test videos. This difference impacts all the other objective metrics scores that are based on frame by frame comparison except ViS3 which is based on the Group Of Pictures (GOP) comparison. Thus, the most correlated metric for Orange 1 database (in terms of PLCC and SROCC) is ViS3. Furthermore, the performance comparison of ViS3 with the state-of-the-art video quality metrics by P.V. Vu and D.M. Chandler [36] reveals that for IP packet loss impairments, VQM General model and MOVIE outperform Vis3 for some databases. However, we prove that for videoconferencing contents ViS3 may be a good indicator for video quality in transmission error conditions too (cf. results on EPFL and Orange 1 databases).

By comparing all the results we notice that generally, for all degradation types and all the databases, SSIMplus is one of the most competitive metrics. Despite the fact that we did not contain impairments in the range of device variability and viewing conditions

in our experimental tests, the SSIMplus shows an accurate video quality prediction ability. Note that the SSIMplus software version that we used was not designed to handle the cases with freezes frames where there is a large temporal misalignment between the reference and the degraded sequences. This explains why the SSIMplus has lower performances on the LIVE Mobile database. But there is a feature built in a commercial SSIMplus LIveMonitor software that automatically align frames up to 10 second difference.

Concerning the VMAF metric, it is highly correlated with the subjective results for all the tested cases except when only network impairments (packet loss and jitter) were simulated (the case of the Orange1 database) . We recall that the VMAF metric approach is based on a machine learning algorithm. Consequently its prediction accuracy largely depends on the characteristics of the training database: impairment types, codec configuration, resolution, frame rate, etc. Indeed, this model has been currently learned on sequences with only degradation caused by changes in resolution and different encoding bit rates. The EPFL database also contains only transmission errors but VMAF shows a good prediction accuracy (PLCC=91%, SROCC=92%, RMSE=0.55). In fact, IP network video packet loss depends highly on the used degradation simulator, the test bed and especially the video decoder and the jitter buffer. For the Orange1 database, some experts visualized the sequences and chose those with more perceived and annoying packet loss (degradation in regions of interest). Furthermore, a random model was used to simulate packet loss degradation for Orange1 database while the Gilbert-Elliot model was used for EPFL database. This difference between the models may explain the difference of the degradation perception.

Table 2.8 reports the statistical significance results of the F-test on the variance of the objective models at a 95% significance level. Each entry in the table consists of 4 symbols corresponding to the databases "EPFL", "LIVE Mobile", "Orange1" and "Orange2". The symbol "+" indicates that the statistical performance of the VQA metric in the column is superior to that of the metric in the row. The symbol "-" means the opposite, while "0" indicates that the statistical performance of the metric in the row is equivalent to that of the metric in the column. Generally, the statistical analysis shows that at a 95% confidence interval, all other metrics outperform PSNR and SSIM. It also proves that the most consistent results with a high accuracy have been achieved by three metrics, i.e. ViS3, SSIMplus and VMAF.

Table 2.8: Statistical significance table based on residuals between model predictions and the MOS values for respectively the EPFL, LIVE Mobile, Orange1 and Orange2 databases. The symbol "+" indicates that the statistical performance of the VQA metric in the column is superior to the one in the row. The symbol "-" means the opposite, while "0" indicates that the statistical performance of the metrics in the row and in the column are equivalent.

	PSNR	SSIM	MS-SSIM	VQM-G	VQM-V	OPVQ	MOVIE	ViS3	SSIMplus	VMAF
PSNR	0 0 0 0	0 0 0 0	0 0 + 0	0 + 0 +	0 + 0 +	0 + 0 +	0 0 0 +	0 + + +	0 + + +	+ + - +
SSIM	0 0 0 0	0 0 0 0	0 0 0 0	0 + 0 0	0 + 0 +	0 + 0 +	0 + 0 +	0 + + +	+ + 0 +	0 + - +
MS-SSIM	0 0 - 0	0 0 0 0	0 0 0 0	0 + - +	0 + - +	0 + - +	0 + 0 +	0 + + +	0 + 0 +	0 + - +
VQMG	0 - 0 -	0 - 0 0	0 - + -	0 0 0 0	0 0 0 0	0 0 0 0	0 - + +	0 0 + +	0 + + +	0 + - +
VQMV	0 - 0 -	0 - 0 -	0 - + -	0 0 0 0	0 0 0 0	0 0 - 0	0 - 0 +	0 0 + +	0 + + +	0 + - +
OPVQ	0 - 0 -	0 - 0 -	0 - + -	0 0 0 0	0 0 + 0	0 0 0 0	0 - + +	0 0 + +	0 0 + +	0 0 - +
MOVIE	0 0 0 -	0 - 0 -	0 - 0 -	0 + - -	0 + 0 -	0 + - -	0 0 0 0	+ + + +	+ + + +	+ + - +
ViS3	0 - - -	0 - - -	0 - - -	0 0 - -	0 0 - -	0 0 - -	- - - 0	0 0 0 0	0 0 - +	0 0 - +
SSIMplus	0 - - -	- - 0 -	0 - 0 -	0 0 - -	0 0 - -	0 0 - -	- - - -	0 0 + -	0 0 0 0	0 0 - 0
VMAF	- - + -	0 - + -	0 - + -	0 - + -	0 - + -	0 0 + -	- - + -	0 0 + -	0 0 + 0	0 0 0 0

2.5 Conclusions and Perspectives

This chapitre presents two subjective audiovisual quality tests investigating audiovisual quality in both interactive and non-interactive contexts and under different scene complexities, as well as an updated survey of the state-of-the-art media-layer full reference objective video quality models for videoconferencing application.

By comparing non-interactive vs. interactive test results, we found that statistically there is no significant difference for MOS_A, MOS_V and MOS_{synch} scores between the two experimental contexts. Nevertheless, considering MOS_{AV} scores we noted a significant difference between the two contexts. Thus, in future experiments we may rely on non-interactive test only and apply their results (with the exception of the evaluation of AV quality, for which interactive tests remain mandatory) to a conversational context. But this result needs to be further validated. Because during the thesis work, non-interactive test was conducted before the interactive test, the source sequences are different in the two tests. For a more fair comparison between the two tests, it would be better to redo a non-interactive test using the same sequences recorded in the conversational context.

The subjective test results show that the scene complexity has an impact on the perceived audiovisual quality in both contexts and on the perception of audio-video synchronization in the interactive context. To add a precision detail and explain this observation, we take two indicative sequences from the recorded conversations and we calculate the SI and TI indexes: for the complex scene TI= 47 and SI= 79; for the

simple scene TI= 29 and SI= 61. Thus, from this observation we might open a question to discuss in a future study: from which difference of scene complexity we could detect a significant difference in perceived video quality?

In all our subjective tests, we have limited ourselves to the evaluation of application type and transmission impairments. It is obvious that a video telephony service is impacted by other factors, such as context, psychological situation, type of terminal, OS .... Enlargement to a wider spectrum of impairments and conditions would allow a finer characterization of the quality of a video call service.

2.6 Contributions in this field

In general, our work led to a better understanding of audiovisual quality assessment processes for videoconferencing services. The contributions are twofold: the constitution of subjective databases of audio visual sequences corresponding to a real scenario of video call; the evaluation of the existing objective quality assessment tools for this specific application.

Actually some other results for this field are not described in this manuscrit due to limited space, i.e. the comparison between laboratory and crowdsourcing subjective tests, combinations of no-reference but distortion-specific metrics for global video quality assessment using basic machine learning approaches, further analysis on the perception of asynchrony. But these (as well as the described works) can be found in the following related publications:

[1] I. Saidi, **L. Zhang**, O. Déforges, V. Barriac. “Laboratory and crowdsourcing studies of lip sync effect on the audio-video quality assessment for videoconferencing application”. ICIP, September 2019, Taipei, Taiwan.

[2] **L. Zhang**, I. Saidi, S. Tian, V. Barriac, O. Déforges, “Overview of full-reference video quality metrics and their performance evaluations for videoconferencing application”. J. Electron. Imaging; March 2019; 28(2), 023001.

[3] I. Saidi, **L. Zhang**, O. Déforges, V. Barriac. “Machine learning approach for global no-reference video quality model generation”. SPIE Optical Engineering + Applications, August 2018, San Diego, California, USA.

[4] I. Saidi, **L. Zhang**, O. Déforges, V. Barriac. “Evaluation of single-artifact based video quality metrics in video communication context”. QoMEX, May 31 - June 2, 2017, Erfurt, Germany.

[5] I. Saidi, **L. Zhang**, V. Barriac, O. Déforges. “Audiovisual quality study for videotelephony on IP networks”. IEEE Workshop on Multimedia Signal Processing (MMSP), September 2016, Montreal, Canada.

[6] I. Saidi, **L. Zhang**, O. Déforges, V. Barriac. “Evaluation of the performance of ITU-T G.1070 model for packet loss and desynchronization impairments”. QoMEX, June 2016, Lisbon, Portugal.

[7] I. Saidi, **L. Zhang**, O. Déforges, V. Barriac. “Interactive vs. non-interactive subjective evaluation of IP network impairments on audiovisual quality in videoconferencing context”. QoMEX, June 2016, Lisbon, Portugal.

[8] I. Saidi, **L. Zhang**, O. Déforges, V. Barriac. “Evaluation de la qualité audiovisuelle des applications de visiophonie”. CORESA, May 2016, Nancy, France.

CHAPTER 3

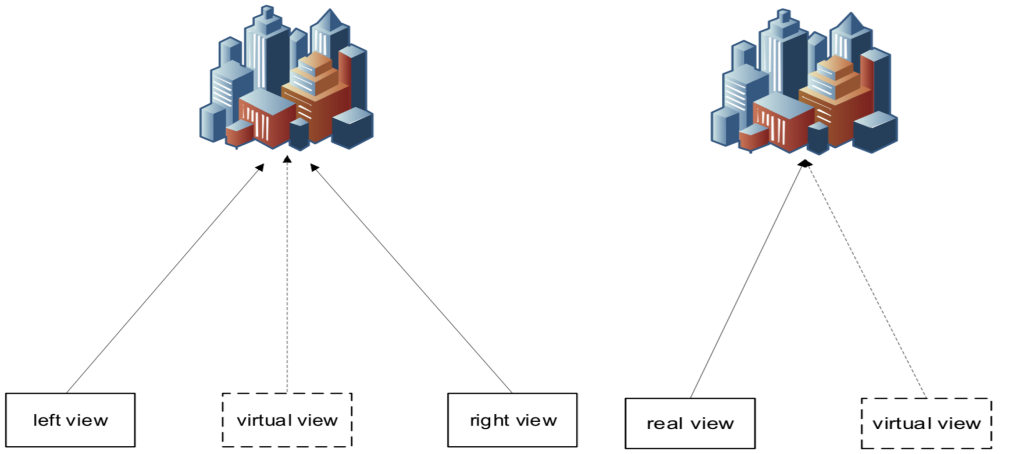
DIBR-SYNTHESIZED VIEW QUALITY ASSESSMENT

3.1 Introduction

The past decade has witnessed the fast development of the 3D movie market. However, this stereoscopic video can only provide two viewpoint videos, the observer can not get a stereoscopic perception at another viewpoint. On the contrary, Free-viewpoint Video (FVV) allows the users to view a 3D scene by freely changing the viewpoints. For example, Canon announced on September 2017 its Free Viewpoint Video System that gives the users a better Quality of Experience (QoE) where they can view sporting events from various different angles and viewpoints. However, containing much more views, these applications require a huge size of data. At the same time, it is also practically impossible to acquire images at all the viewpoints of a particular 3D scene, which is instead captured by multiple cameras at different viewpoints. Thus, some of the views have to be synthesized, often by using the Depth-Image-Based-Rendering (DIBR) technique [54].

The idea of DIBR is to synthesize the virtual views by using the texture and depth information at another viewpoint. There are two main kinds of DIBR view synthesis algorithms (cf. Fig. 3.1): the single view based synthesis and the interview synthesis: for the single view based synthesis, we use the one base view to synthesis another; for the interview synthesis, we use two base views to render the middle one.

Recently, the DIBR becomes also a promising solution for synthesizing virtual views in many other recent popular immersive multimedia applications, such as Virtual Reality (VR) [55], Augmented Reality (AR) [56] and Light Field (LF) multi-view videos [57], etc. For example, DIBR has already been used in a light field compression scheme where only very sparse samples (four views at the corners) of light field views are transmitted while the others are synthesized. This new scheme significantly outperformed High



(a) interview synthesis (b) single view based synthesis

Figure 3.1: DIBR view synthesis

Efficiency Video Coding (HEVC) inter coding for the tested LF images [58]. Another example concerns 360-degree and volumetric videos: two developing areas pointing to how video will evolve as VR/AR technology becomes the mainstream [59]. Current 360-degree videos allow viewers to look around in all directions, but only at the shooting location: they do not take into account the translation (changes in position) of the head. To achieve more immersive QoE, some companies propose to use DIBR to synthesize the non-captured views when users move from the physical camera's position, as proposed in Technicolor's volumetric video streaming demonstration [60]. In the social and embodiment VR media applications, where a VR media designed for 360-degree videos mixed with real-time objects for multiple users, an eye-contact technique based on the DIBR [60] can provide the users the viewpoint according to their eye positions, which gives the users a better interactive QoE.

Although DIBR has a great potential, current DIBR algorithms may introduce some new types of distortions (e.g. object shifting, ghosting effect, object warping, slight geometric distortion, stretching, blurry regions, crumbling, flickering, black holes, etc.) which are quite different from those caused by image compression. Most compression methods can cause specific distortions [61], eg. blur [62], blockiness [63] and ringing [64]. These distortions are often scattered over the whole image, while the DIBR-synthesized artifacts (caused by distorted depth map and imperfect view synthesis method) mostly occur in the disoccluded areas. Since most of the commonly

3. Source: `https://developer.att.com/blog/shape-future-of-video`

used 2D objective quality metrics are initially designed to assess common compression distortions, they may fail in assessing the quality of DIBR-synthesized images [65, 66].

This chapitre presents our contributions to the improvement of quality assessment of DIBR-synthesized views, including two No-reference (NR) objective metrics and two Full-reference (FR) objective metrics, as well as a new DIBR image database.

3.2 FR metrics: SC-DM and SC-IQA

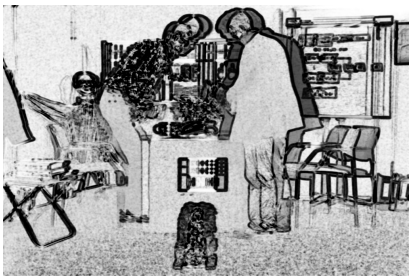
3.2.1 SC-DM

In FR pixel-wise metrics (eg. PSNR, SSIM), the global shift in DIBR-synthesized views is often easily penalized. To solve this problem, we propose a Shift Compensation and Dis-occlusion based Model (SC-DM), which firstly compensating the global object shift, and then using a disparity map as a mask to weight the final distortion. This model can be combined with any pixel based FR metrics, but we only tested it on the commonly used PSNR and SSIM. The SC-DM can be divided into two parts: global shift compensation and dis-occlusion mask weighting.

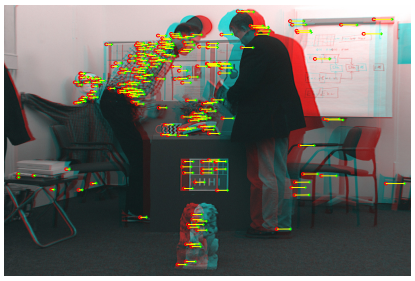
Global shift compensation Fig. 3.2 (a) gives an example of the SSIM map between the synthesized image and the reference image in the adopted database [67], it can be observed that there is a great global shift between the synthesized image and the reference image.

In this part, the global geometric shift is compensated roughly by a SURF [68] + RANSAC [69] homography approach. Firstly, SURF feature points in the reference and synthesized images are detected and matched. Then, to be robust, the RANSAC algorithm is used to refine the matching and estimate the homography matrix H. After that, the pixels of the synthesized image are warped to the corresponding positions in the reference image by H. The SSIM map before transform, the matched feature point pairs and the SSIM map after transform are shown in Fig. 3.2.

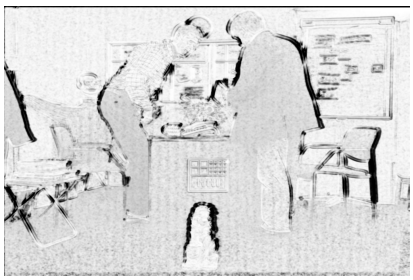
We can observe that the global shift between the synthesized and the reference images has been roughly compensated since only a limited number of regions gets very low SSIM value (the black regions in Fig. 3.2(c)). Compared to the SSIM map



(a) SSIM map before transform



(b) optimized matched feature point pairs



(c) SSIM map after transform

Figure 3.2: Example of feature points matching and transform

before transform (Fig. 3.2(a)), the SSIM map after transform Fig. 3.2(c) shows that most of the ghost effect in the SSIM map has been removed.

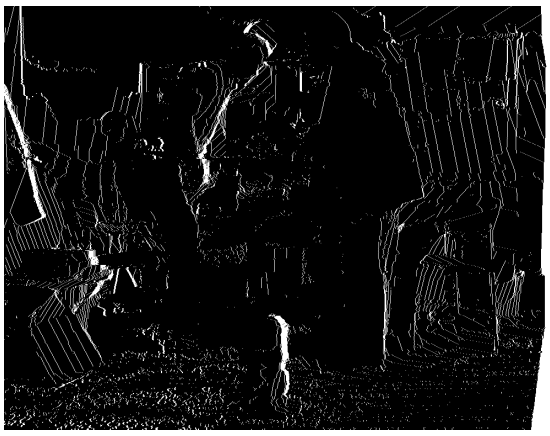


Figure 3.3: Example of dis-occluded mask

Dis-occlusion Mask In this part, we use a dis-occlusion mask to weight the difference between the synthesized image and the reference image. As introduced in section 3.1, the major problem of the DIBR method is the dis-occlusion: regions which are occluded in the captured views become visible in the virtual ones. Due to the lack of original texture information, a synthesized image often contains dis-occlusion holes which significantly degrade the quality. Thus, we utilize a dis-occlusion mask to weight the final distortion. The depth map in the original view-point ($Depth_o$) is used to calculate the dis-occlusion mask. In a rectified configuration, 3D warping process, the horizontal disparity which is the horizontal displacement for each pixel can be obtained by Eq. (3.1):

$$d = \frac{f \times l}{Z} \tag{3.1}$$

where f, l, Z represent the camera focal length, the baseline distance between these two views and the depth value of this pixel respectively.

The depth map in the synthesized view-point ($Depth_s$) given initial value to -1, then the depth map in the original view-point ($Depth_o$) is warped to the synthesized view-point by Eq. (3.2):

$$Depth_s(i+d,:) = Depth_o(i,:); \quad (i+d), i \in [1, W] \tag{3.2}$$

where W is the image width, the colon ":" indicates all subscripts in this array dimension.

The dis-occluded mask dis_mask can then be obtained by extracting all the pixels with value -1 in $Depth_s$, which is shown in Fig. 3.3. This mask is a binary image, the while pixel's value equals "1", while the dark pixel's value equals "0".

Weighted PSNR and Weighted SSIM Generally speaking, the dis-occlusion mask dis_mask, can be integrated into any existing full-reference metric as a weighting mask since the DIBR view synthesis distortion mainly occur in the dis-occluded regions. We propose and test the weighted $PSNR$ ($PSNR'$) and $SSIM$ ($SSIM'$) as defined in the following equations:

$$MSE' = \frac{\sum_{(i,j)\in I}(I_{syn}(i,j) - I_{ref}(i,j))^2 \cdot dis\_mask(i,j)}{\sum_{(i,j)\in I} dis\_mask(i,j)} \tag{3.3}$$

$$PSNR' = 10 \cdot log_{10}(\frac{255 \times 255}{MSE'}) \tag{3.4}$$

$$SSIM' = \frac{\sum_{(i,j)\in I} SSIM(i,j) \cdot dis\_mask(i,j)}{\sum_{(i,j)\in I} dis\_mask(i,j)} \tag{3.5}$$

where I_{syn} and I_{ref} denote the the compensated synthesized image and the reference image respectively; dis_mask denotes the obtained disocclusion mask; $SSIM$ denotes the $SSIM\ map$ between the compensated synthesized image and the reference image.

3.2.2 SC-IQA

In the previous model SC-DM, the global shift has not been compensated precisely, thus later we proposed a Shift Compensation based Image Quality Assessment metric (SC-IQA) for DIBR-synthesized views by using a multi-resolution block matching method. Besides, it does not need the depth map. The block diagram is shown in Fig. 3.4, in addition to the SURF + RANSAC homography transform, a multi-resolution block matching is proposed to precisely compensate the object shift and penalize the local artifacts. Besides, a saliency map is used as a weighting function to improve the performance. The final overall quality scores are obtain by measuring the γ% worst blocks since human observers are more sensitive to poor quality regions rather the good ones.

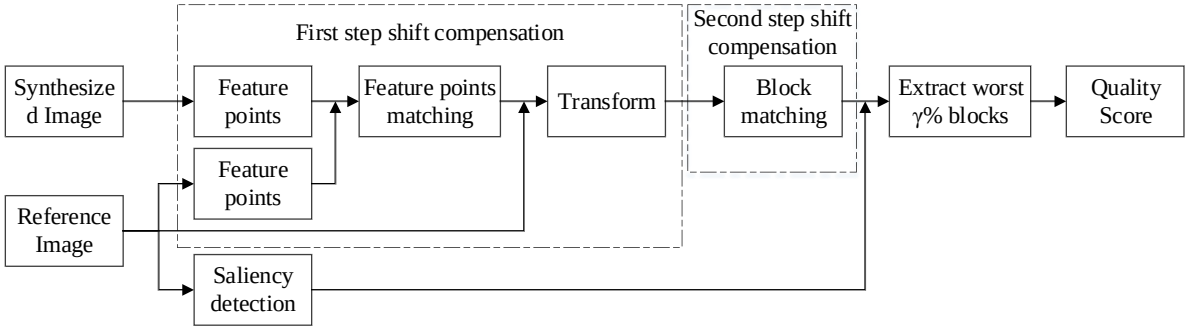
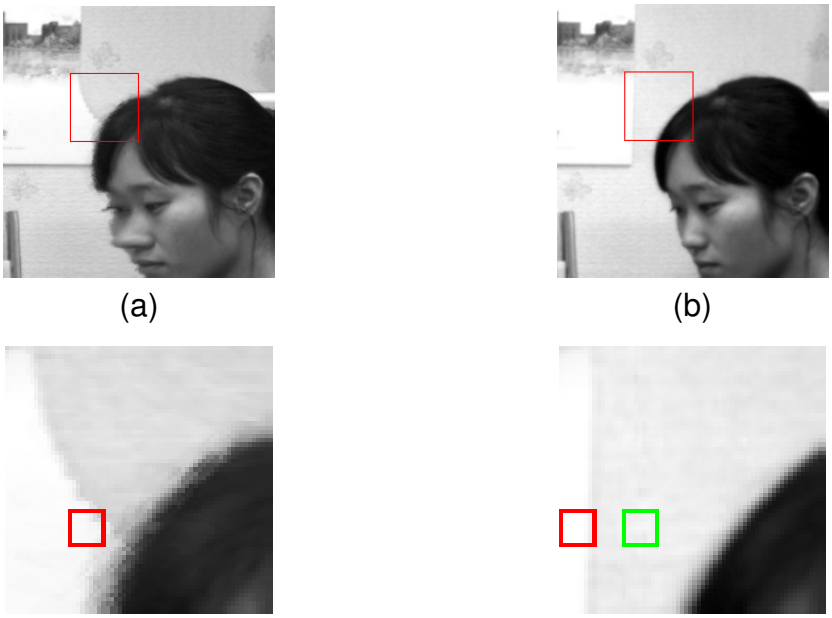


Figure 3.4: Block scheme of the SC-IQA metric

Multi-resolution block matching In this part, a multi-resolution block matching algorithm is used to precisely compensate the shift and also to detect the large geometric



(a) (b) (c) (d)

Figure 3.5: Block matching: (a), (b) are the patches in the synthesized and the reference image; (c) block in the synthesized image; (d) matched block in the reference image: for direct 8x8 block-matching (red block), or multiresolution block-matching (green)

distortions. We will see bellow on an example why a regular block-matching would not be adequate. In the first step, we use a large block $N1 \times N1$ (N1 = 64) for primary matching; then we use a small block $N2 \times N2$ (N2 = 8) for final matching. The matching process can be described by the following steps:

1. Divide the synthetized view into a regular grid of $N1 \times N1$ blocks;

2. For each $N1 \times N1$ block, search for the best matching block in the reference view. The best matching block is the one showing the largest following similarity criterion:

$$sim(s,r) = \frac{cov(s,r) + \epsilon}{var(s) + var(r) + \epsilon} \tag{3.6}$$

   where s, r denote the blocks in the synthesized image and the reference image; the operation cov and var denote the co-variance and variance respectively; ϵ is a constant value to stabilize the division with weak denominator.

3. Each $N1 \times N1$ block is divided into smaller $N2 \times N2$ blocks and the process is repeated with a smaller search window.

Since the shift only occurs in the horizontal direction, we only search the blocks in this direction for matching. We assume the biggest shift in the synthesized image to be 30, the search windows of N1 and N2 are restricted to 30 and 5 respectively.

The goal of this multi-resolution block matching algorithm is to compensate the global shift and not compensate local geometric distortions, so that they will be penalized. Now, if we directly use $N2 \times N2$ block for matching, and set the search window to 30 (the biggest shift range in the synthesized image), the computational complexity will be much higher. Besides, as shown in Fig. 3.5, there exists great geometric distortion in the red block ($N1 \times N1$) in Fig. 3.5 (a) compared to its matched block in the reference image (the red block in Fig. 3.5 (b)). If we directly use 8×8 block for matching and set the searching window to 30, the best matched block for the block in Fig. 3.5 (c) is the red block in Fig. 3.5 (d). There exists little difference between these two red blocks, so the geometric distortion will not be penalized. On the contrary, if we use the proposed multi-resolution block matching method, the matched block is the green one, this geometric distortion will be surely penalized. The multi-resolution approach is thus more efficient to find the real physically matching block, and detect wether there is local distorsion within this block..

Saliency weighting In addition, a saliency detection [70] is also used as a weighting map to improve the performance of the proposed metric. The distortion of each $N2 \times N2$ block is measured by averaging the weighted mean square errors between the blocks of the synthesized and the reference images, as shown in:

$$MSE_B = \frac{\sum_{(i,j)\in B} (syn(i,j) - ref(i,j))^2 \times Sal\_map(i,j)}{\sum_{(i,j)\in B} Sal\_map(i,j)} \tag{3.7}$$

where B means the matched $N2 \times N2$ blocks; (i, j) denotes the pixel in the block; syn and ref represent the blocks in the synthesized image and reference image respectively; Sal_map represents the saliency map in this block.

Quality pooling Since humans tend to perceive poor regions in an image with more severity than the good ones [71, 72], we only use the blocks with the worst quality to calculate the final quality as shown in Eq. 3.8.

$$MSE_W = \frac{1}{N_W} \sum_{i\in W} MSE_B(i) \tag{3.8}$$

where W represents the set of the worst γ% blocks in the image, N_W is the number of items in the set W. The final quality score is computed as the following equation:

$$Score_{SC-IQA} = 10 \times log_{10}(255 \times 255/MSE_W) \tag{3.9}$$

where a higher quality score indicates a better quality.

3.3 NR metrics: NIQSV and NIQSV+

FR metrics always need the reference view which may be unavailable in some circumstances. Thus we also proposed two NR metrics for DIBR-synthesized views (called NIQSV and NIQSV+).

3.3.1 NIQSV

The NIQSV (No-reference Image Quality assessment of Synthesized Views) metric assume that a good quality synthesized view should present sharp and regular object borders, smooth values inside the object and large discontinuities at the object borders. Such “perfect” images are insensitive to opening and closing morphological operations while some artifacts such as blurry regions around the object edges and crumbling in the synthesized views are sensitive to such morphological operations. The crumbling is small-sized artifacts which can be easily detected by the morphological operations with Structural Element (SE) larger than their size; the blurry regions change much more significantly after the opening and closing morphological operations compared to the good quality images with sharp edges and flat areas. Thus, these properties could be used to detect these artifacts.

The NIQSV’s block scheme is presented in Fig 3.6. It quantifies the distortions in luminance component Y and chrominance components U, V using a set of morphological operations. Then the 3 obtained distortions are pooled into one global distortion by a weighted average. Furthermore, an edge image is utilized to weight the final distortion since the distortions of synthesized views mainly happen around object edges.

The opening operation used on the synthesized image can help to remove some thin blurry regions, and the following closing operation with a relatively larger Structural Element (SE) can fill the holes in the disoccluded areas. The distortion of each

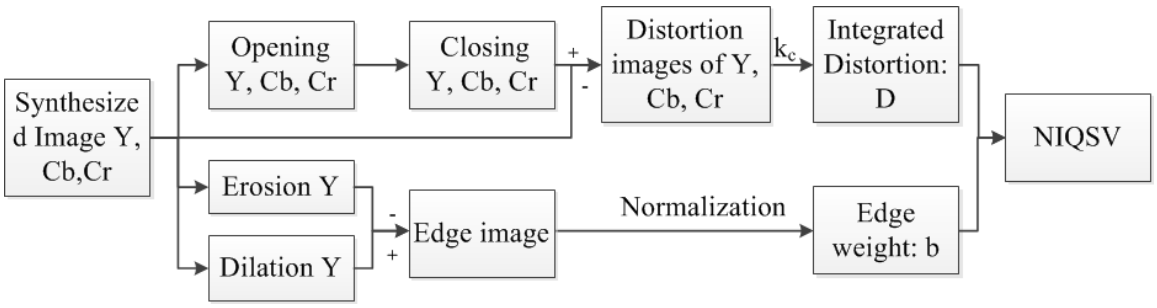


Figure 3.6: Block scheme of NIQSV

component is obtained by measuring the difference between the original component I_X and the processed component I'_X after the opening and closing operation. It can be computed as follows:

$$I'_X = (I_X \circ SE_o) \bullet SE_c, X \in (Y, Cb, Cr) \tag{3.10}$$

$$D_X = |I'_X - I_X|, X \in (Y, Cb, Cr) \tag{3.11}$$

where D_X denotes the difference of each color component, I_X is the corresponding color component of the synthesized image, SE_o is the SE used for opening and SE_c is the SE used for closing. In this paper, the shape of SE_o and SE_c is a circle, the size of SE_o and SE_c is 3 and 8 respectively.

The overall distortion is obtained by the following equation:

$$D = (1 - w_c) \cdot D_Y + \frac{w_c}{2} \cdot (D_{Cb} + D_{Cr}) \tag{3.12}$$

which is a weighted sum of the distortions computed on each color component where the weight is related to the parameter w_c. The value of w_c is set to 0.5 which means that the distortion in luma component weights 50% in the overall distortion.

To reduce computational complexity, the edge image is firstly extracted by a pair of morphological operators as described in Eq. (3.13). Then, they are normalized to $[0, 1]$ using Eq. (3.14):

$$E = (I_Y \oplus SE) - (I_Y \ominus SE) \tag{3.13}$$

$$e = E/Vmax; Vmax = 255, e \in (0, 1) \tag{3.14}$$

where SE is the structural element used for erosion and dilation, the symbols $\oplus$ and $\ominus$ denote the morphological dilation erosion operation respectively. The shape of SE is

a circle and its diameter is set to 4. $E/Vmax$ (where Vmax is the maximum value that an edge-detector may provide for 8-bit images: 255) is used as the edge weight. The final edge weight e is used to weight the overall difference D in the whole image. The pixels with higher edge value have more weight on the distortion map. Especially, for the pixels with no edge, the distortion on it will not be considered.

Finally, the overall image quality score $NIQSV$ is computed as follows:

$$MSE' = \frac{\sum_{(i,j)\in I} e(i,j) \cdot D(i,j)^2}{\sum_{(i,j)\in I} e(i,j)} \tag{3.15}$$

$$NIQSV = 10 \cdot log_{10}(\frac{255 \times 255}{MSE'}) \tag{3.16}$$

Fig. 3.7 shows the processed images of one synthesized view in the “Newspaper” sequence as an example.

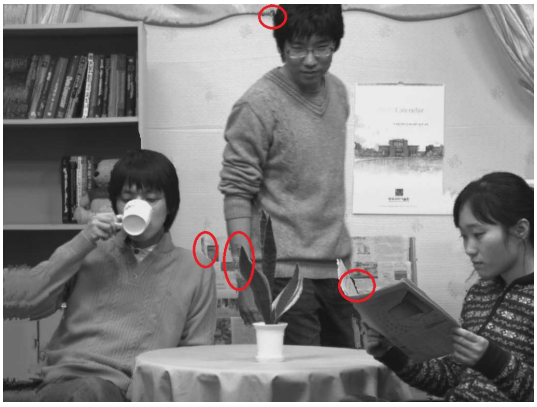
3.3.2 NIQSV+: an extension of NIQSV

The NIQSV+, as an extension of NIQSV, is also based on the assumption that the images with good quality are composed of flat regions separated by sharp and regular edges. A block diagram of NIQSV+ is presented in Fig. 3.8. The proposed method can be divided into three parts. Part A is designed to detect the blurry regions and crumbling around the object edges, which has been introduced in the previous section NIQSV; part B is related to the unfilled black holes in the dis-occluded areas; and part C is the detection of stretching distortion which always occurs in the left or right side of the synthesized view.

Detection of black holes In part B, the distortion of unfilled black hole pixels is taken into consideration. Normally, most natural images do not contain pixels with 0 luminance value. Thus, we use the proportion of black hole pixels in the whole image to measure this type of distortion, as defined in Eq. 3.17:

$$Zrate = NumofBHpixels/(W \times H) \tag{3.17}$$

where $NumofBHpixels$ denotes the number of black hole pixels in the whole images, W and H are the width and the height of the image.



(a) Synthesized image I_Y

(b) Open and closed image $(I_Y \circ SE_o) \bullet SE_c)$

(c) Normalized edge weight b

(d) Overall Difference D

Figure 3.7: Examples of intermediate results in the NIQSV measurement for one synthesized view in the “Newspaper”sequence. The distortions marked in (a) are well detected in (d), while in the non-distortion regions, such as the girl’s hair, the distortion values are very low.

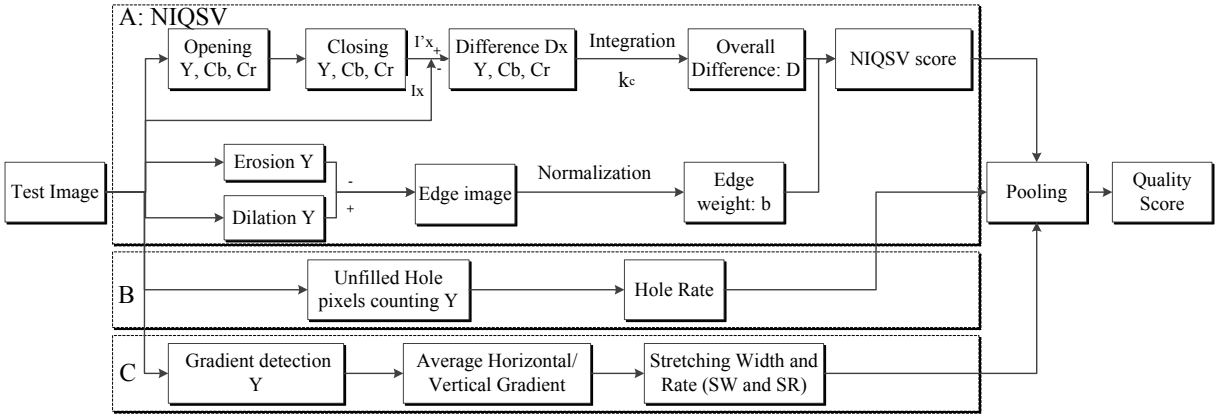


Figure 3.8: Block Diagram of NIQSV+

Detection of Stretching In part C, a stretching measurement is defined, based on the observation that the stretching may happen around the left or right side of the image due to lack of the corresponding texture information, to estimate the level of stretching in the synthesized image. The stretching is detected by measuring the crash of horizontal gradient in the stretching area. Firstly, the horizontal and vertical gradients are calculated with the Sobel operator.

$$\begin{cases} \nabla_{ver} = I_y * G_{ver} \\ \nabla_{hor} = I_y * G_{hor} \end{cases} \tag{3.18}$$

where I_y is the Y component of the synthesized image, G_{hor} and G_{ver} denote the Sobel horizontal and vertical gradient operator. The Average Horizontal / Vertical gradient ($\bar{g_H}$ / $\bar{g_V}$) in column are defined in Eq. 3.19.

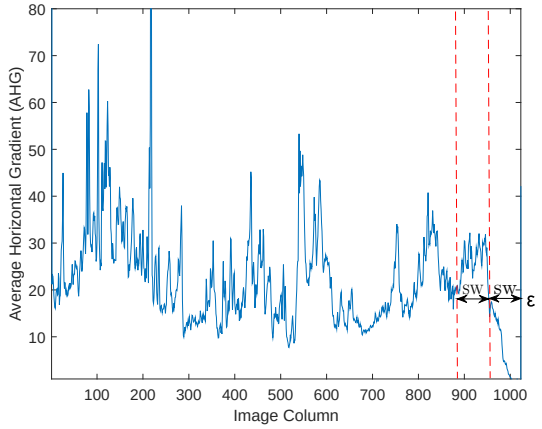
$$\begin{cases} \bar{g_H} = \sum_{j=1}^{H} \nabla_{hor}/H \\ \bar{g_V} = \sum_{j=1}^{H} \nabla_{ver}/H \end{cases} \tag{3.19}$$

where H denotes the height of the image. Since stretching mainly happens in the horizontal direction, the average horizontal gradient in the column can be used to detect the stretched regions.

As shown in Fig. 3.9, the average values of the horizontal gradient in the stretched area (in the right side) are very low. The Stretching Width (W_s) is obtained by calculating the width of this area.

$$W_s = \sum_{j=1}^{0.1\times W} S + \sum_{j=0.9\times W}^{W} S \tag{3.20}$$

(a) Synthesized Image ($W = 1024$, $H = 768$)



(b) $\bar{g_H}$ distribution

Figure 3.9: Synthesized Image and its corresponding average horizontal gradient

where S is a index to mark the stretching areas. Since the stretching artifacts only occur in the left or the right side of the image, $0.1 \times W$ and $0.9 \times W$ are used to take into account only the side portions of the image.

$$S = \begin{cases} 1, \bar{g_H} < \varepsilon \\ 0, \ else \end{cases} \tag{3.21}$$

where ε is a threshold used to extract the stretching regions, its value is set to 50% of the mean value of $\bar{g_H}$. However, even with the same stretching width, the perceptual annoyance could be different in different textures. In order to handle this issue, a Stretching Rate (R_s) is defined by comparing the average gradient in the stretching regions and those in the adjacent non-stretching regions with the same width, as shown in Fig. 3.9. The more similar these two regions are, the less the stretching will be per-

ceptible. Since the mean horizontal gradient of these two regions is quite different, we only compare the vertical gradients.

$$R_s = \frac{\nabla_{ref} - \nabla_{str}}{\nabla_{ref}} \tag{3.22}$$

where ∇_{str} presents the average $\bar{g_V}$ value in the stretching area, ∇_{ref} is the average $\bar{g_V}$ value in the adjacent non-stretching regions with the same width. When the ∇_{str} and ∇_{ref} values are closer, this type of distortion is less significant, and the SR value is lower. The final stretching distortion is calculated as follows:

$$S\_index = (log_{10}(W_s + 1) + 1) \times (R_s + 1) \tag{3.23}$$

Overall quality measurement Finally, the integrated overall quality score is computed as Eq. 3.24 since higher stretching and black hole rate indicate bad image quality.

$$NIQSV+ = \frac{NIQSV}{S\_index \times (1 + k_z \times Zrate) + C} \tag{3.24}$$

where k_z denotes the weight of black hole distortion to the final measurement. Since the black hole pixels hold a very low proportion in the whole image, k_z should be a large value. C is a constant used to adjust the difference between the images with "black hole" or "stretching" artifacts and those without. The dependency of these two parameters (k_z, C) are discussed in Section 2.3. The evaluation of the NISQ and NIQSV+ proposed metrics is presented in the next section.

3.4 Performance evaluation of four proposed metrics

The performances of the four proposed metrics are evaluated and compared to other state-of-the-art metrics using IRCCyN/IVC DIBR image database [73, 67]. This database contains frames from 3 different MVD sequences: Book Arrival (1024×768, 16 cameras with 6.5 cm spacing), Lovebird1 (1024×768, 12 cameras with 3.5 cm spacing) and Newspaper (1024×768, 9 cameras with 5 cm spacing). For each sequence, there are four virtual views generated from another viewpoint using the following seven DIBR synthesis algorithms A1-A7:

- A1 [54]: the depth map is filtered to remove depth discontinuities; borders are

cropped and then the image is interpolated to reach its original size. This may lead to shifting and global radial artifacts.

- A2: the depth map is pre-processed in the same way as in A1, and the borders are in-painted as described in [74] instead of being cropped. This may induce blurring and geometry distortions around the object discontinuities since the depth map is pre-processed by a low-pass filter.
- A3: Tanimoto et al. [75] proposed a 3D view generation system which is adopted as a reference software by the MPEG 3D video group. The blended mode was not used, thus meaning only one image was used to interpolate the virtual view. The in-painting method[74] is also used in A3, which may induce blur into the disoccluded regions.
- A4: Muller et al.[76] proposed a hole filling method aided by depth information. The corresponding depth values at the hole boundary are examined row-wise to find background color samples to be copied into the hole. This may fail to reconstruct the vertical or oblique structures and complex textures. Some foreground color may be propagated into the hole owing to the depth estimation errors.
- A5: Ndjiki-Nya et al. [77] used a patch-based texture synthesis method to fill the missing part in the virtual view. Since the used patches are rectangular, which may lead to block artifacts and only straight edges could be accurately reconstructed.
- A6: Koppel et al. [78] extended A5 by a background sprite which takes the temporal information into consideration to improve the synthesis.
- A7: holes in virtual views are left unfilled.

For each of the synthesized viewpoints a reference view is available as the chosen virtual viewpoints correspond to viewpoints also acquired with a real camera.

Here the objective quality scores are firstly mapped to the subjective scores using:

$$DMOS_p = a \cdot scores^3 + b \cdot score^2 + c \cdot score + d \tag{3.25}$$

where $score$ is the score obtained by the objective metric and a, b, c, d are the parameters of the cubic function. They are obtained through regression to minimize the difference between $DMOS_p$ and $DMOS$.

For a fair comparison with the other metrics, we use a cross-validation scenario to obtain the performance of the proposed metric: the adopted database is partitioned into two non-overlapping sets with randomly selected 50% images as a training set and the other 50% as a test set. This random train-test procedure was repeated 100 times and the average performance on the test set across the 100 iterations was reported as the performance of our proposed method.

The PLCC, RMSE, SROCC values are shown in Table 3.1, from which we can see that NIQSV performs much better than PSNR and SSIM and achieves very closely to the other three FR metrics: 3DSwIM, MW-PSNR and MP-PSNR, the SROCC value is even a little better than 3DSwIM. Compared to NIQSV, NIQSV+ improves further the performance a lot with additional steps (detection of black holes and stretching).

Table 3.1: PLCC, RMSE and SROCC between DMOS and objective metrics. (The best four results are marked in bold), NIQSV+_s means NIQSV with stretching detection, NIQSV+_b means NIQSV with black hole detection

Metric		PLCC	RMSE	SROCC
NR 3D metrics	**NIQSV+**	**0.7114**	**0.4679**	**0.6668**
	NIQSV+_s	0.6886	0.4828	0.6497
	NIQSV+_b	0.6423	0.5103	0.4806
	NIQSV	0.6346	0.5146	0.6167
	APT [79]	**0.7307**	**0.4546**	**0.7157**
FR 3D metrics	**SC-IQA**	**0.8496**	**0.3511**	**0.7640**
	PSNR'	**0.8242**	**0.3771**	**0.7889**
	SSIM'	0.5681	0.5478	0.5475
	3DSwIM [80]	0.6864	0.4842	0.6125
	MP-PSNR [81]	0.6729	0.4925	0.6272
	MP-PSNRr [81]	0.6954	0.4784	0.6606
	MW-PSNR [82]	0.6200	0.5224	0.5739
	MW-PSNRr [82]	0.6625	0.4987	0.6232
	VSQA [83]	0.6122	0.5265	0.6032
FR 2D metrics	PSNR	0.4557	0.5927	0.4417
	SSIM [84]	0.4348	0.5996	0.4004
	MS-SSIM [85]	0.5406	0.5602	0.5021
	IW-PSNR [86]	0.3608	0.6210	0.3460
	IW-SSIM [86]	0.5337	0.5631	0.4795
NR 2D metrics	NIQE [87]	0.4022	0.6096	0.3673
	BIQI [88]	0.5273	0.5657	0.3555
	BliindSII [89]	0.5331	0.5633	0.1800

We can see that the proposed weighted PSNR (PSNR') and SC-IQA (γ = 1) per-

form significantly better than other tested metrics. The PLCC gain of $PSNR'$ achieves 36.85% compared to the PSNR. The weighted SSIM ($SSIM'$) achieves a gain of PLCC 13.33% compared to the SSIM.

Table 3.2: Execution time of each IQA metric normalized base on PSNR. The metrics A-Z indicate NIQSV+, NIQSV, APT, 3DSwIM, MP-PSNR, MP-PSNRr, MW-PSNR, MW-PSNRr, VSQA, PSNR, SSIM, IW-PSNR, IW-SSIM, NIQE, BIQI and BliindS2 respectively.

Metric	A	B	C	D	E	F	G	H
time	21	18	13k+	90	100	35	12.4	9.6
Metric	I	J	K	L	M	N	O	P
time	140	1	7.4	75	75	45	67.5	6.8

As NR metrics, the proposed metric NIQSV+ and APT have good performances, even better than most state-of-the-art FR 3D metrics (except PSNR' and SC-IQA). APT performs a little better than our proposed method (PLCC 0.019 higher), but the proposed method executes much faster cf. Table 3.2.

3.5 A new database for benchmarking DIBR algorithms

3.5.1 Motivation of the new database

There are several DIBR related databases, cf. Table 3.3. Each database has its own focus. The IVC databases focus on the distortions caused by different DIBR synthesis algorithms, the MCL-3D and SIAT database investigate the influence of traditional 2D distortions of original texture and depth map on the DIBR-synthesized views.

Our IETR database focuses on the distortions only caused by DIBR algorithms (like the IRCCyN/IVC DIBR database), but with state-of-the-art DIBR algorithms. The “view" or stimuli in this subjective test indicates an individual synthesized image. In total, we tested seven DIBR algorithms, including both the interview synthesis and the single view synthesis methods. We selected those DIBR algorithms which produce no longer “old-fashioned” artifacts and of which the code sources were provided by their authors. Note that the SIAT database focuses on the effect of texture and depth compression on the synthesized views and it contains only one DIBR algorithm. Compared to the MCL-3D and the IVY databases, the proposed new database (1) includes not only virtual views generated by view extrapolation, but also by view interpolation; (2) tests

more and newer DIBR algorithms; (3) shows the views on a 2D display to avoid the 3D display settings and configurations influences (same approach was used in the IRCCyN/IVC DIBR database). The IRCCyN/IVC DIBR database also focuses on the comparison of different DIBR algorithms, but it contains some "old-fashioned" DIBR artifacts (eg. black holes) and it contains less source images than ours. In addition, since machine learning (especially deep learning) based methods become more and more popular for image quality assessment recently, larger data is essential for the development of these methods and different databases are usually needed for the cross-validation. The proposed database can be used along with the IVC database for this type of usage.

3.5.2 Description of the new database

Ten MVD test sequences provided by MPEG for the 3D video coding are used in this experiment. The $Balloons$, $BookArrival$, $Kendo$, $Lovebird1$, $Newspaper$, $Poznan\ Street$ and $PoznanHall$ sequences are natural images while the $Undo\ Dancer$, $Shark$ and $Gt\ Fly$ are computer animation images, as shown in Fig. 3.10. The characteristics of the sequences are summarized in Table 3.4.

Eight DIBR algorithm are used in this database, cf. Table 3.5. For each single view based DIBR algorithm, a single virtual viewpoint is extrapolated from the neighboring two views separately. For the interview DIBR algorithms, the virtual viewpoint is synthesized based on both neighboring views. We consider thus for each reference image, 2 virtual views synthesized by 2 interview synthesis algorithms and 12 virtual views synthesized by 6 single view based DIBR algorithm, which leads to 14 degraded images.

Test protocol We choose to follow the SAMVIQ protocol because of its stability, reliability and relatively higher discriminability. The experiment was conducted on a NEC MultiSync PA322UHD monitor with resolution 3840×2160. The environment of the subjective experiment was controlled as recommended in the ITU-R Rec. BT.1788 [90]. Altogether, 42 naive observers (28 males and 14 females with an age varying from 19 to 52 years old) participated in the subjective assessment experiment. One observer is eliminated after the observer screening using the method recommended in the ITU-R Rec. BT.1788 [90]. That leads to 41 observers finally for this database.

Table 3.3: Summary of existing DIBR related database

Name	No. seq.	No. DIBR algos	DIBR algos		other distortions	size	Reference	display
			Name	Year				
IVC DIBR-image	3	7	Fehn's	2004	No	84	original	2D
			Telea's	2003				
			VSRS	2009				
			Müller	2008				
			Ndjiki-Nya	2010				
			Köppel	2010				
			Black hole	—				
IVC DIBR-video	3	7	idem		H.264	84	original	2D
MCL-3D	9	4	Fehn's	2004	Additive White Noise	693	synthesized	Stereo.
			Telea's	2003	Blur			
			HHF	2012	Down sampling			
			Black hole	—	JPEG			
					JPEG2k			
					Translation Loss			
SIAT video	10	1	VSRS	2009	3DV-ATM coding	140	original	
IVY	7	4	Criminisi	2004	No	84	original	Stereo.
			Ahn's	2013				
			VSRS	2009				
			Yoon	2014				
IETR (proposed)	10	7	Criminisi	2004	No	140	original	2D
			VSRS	2009				
			LDI	2011				
			HHF	2012				
			Ahn's	2013				
			Luo's	2016				
			Zhu's	2016				

(a) BookArrival (b) Lovebird1 (c) Newspaper

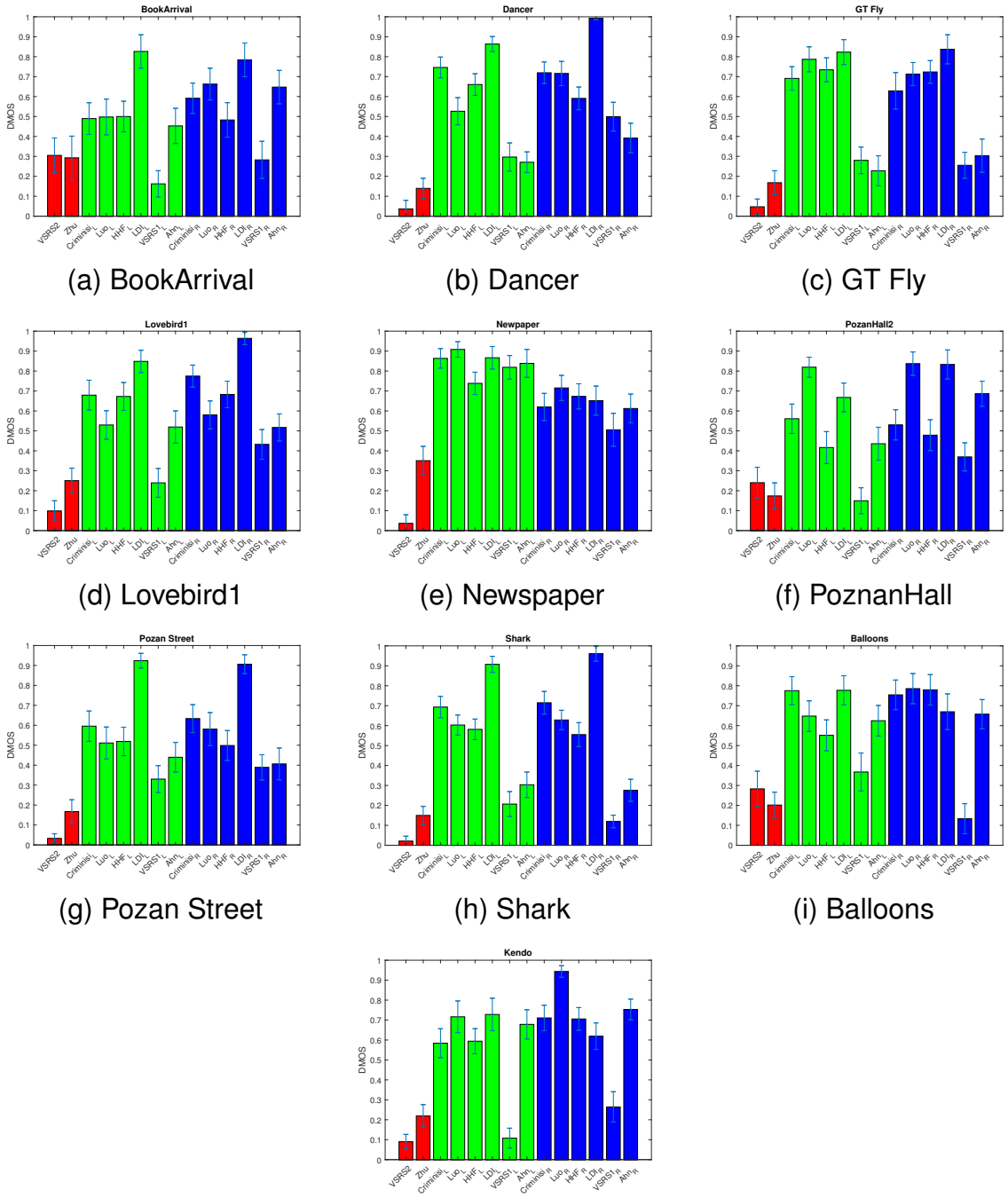
(d) Balloons (e) Kendo

(f) Dancer (g) GT Fly (h) PoznanHall

(i) Pozan Street (j) Shark

Figure 3.10: The used MVD sequences

Subjective scores The obtained $DMOS$ score distributions and their confidence intervals are shown in Fig. 3.11. Generally, the interview synthesis methods outperform the single view based synthesis methods in most sequences. However in some sequences, such as $BoolArrival$, the VSRS1 get better results than VSRS2 and Zhu's



(a) BookArrival (b) Dancer (c) GT Fly

(d) Lovebird1 (e) Newspaper (f) PoznanHall

(g) Pozan Street (h) Shark (i) Balloons

(j) Kendo

Figure 3.11: DMOS distribution and confidence intervals of the synthesized views of different MVD sequences and different view synthesis methods. The x-labels are $VSRS2$, Zhu, $Criminisi_L$, Luo_L, HHF_L, LDI_L, $VSRS1_L$, Ahn_L, $Criminisi_R$, Luo_R, HHF_R, LDI_R, $VSRS1_R$, Ahn_R ordinally. The subscript L means this virtual view is synthesized from the neighboring left view, while the subscript R means from the right. $VSRS2$ denotes the view interpolation inter-view mode of VSRS. The error bars indicates the corresponding confidence intervals of the tested images. The bars referring to inter-view synthesize views are marked by red, the left-extrapolated views marked by green, and the right-extrapolated views marked by blue.

methods, but not very significantly according to the corresponding confidence intervals. One reason could be that, owing to the inaccuracy of depth map, the same object in the two base views are rendered to different positions which results in a "ghost" effect in the synthesized view. However, this situation does not happen in single view based synthesis method VSRS1.

A statistical analysis (student T-test here) was also made over the obtained $DMOS$ scores, to show the statistical equivalence information of the tested algorithms. The scores of single view based methods are obtained by averaging the scores of the two images synthesized from the viewpoints at the two sides. The t-test results show that the view interpolation methods (VSRS2 and Zhu's), which use the two neighboring views as reference views, perform much better than the single view based methods. Among the single view based approaches, VSRS1 and Ahn's methods are significantly superior to the others.

Table 3.4: Introduction of the tested MVD sequences

Sequence	Resolution	Frame No.	View ref. Position	View sys. Pos.	SI
BookArrival	1024 $\times$ 768	58	8, 10	9	60.2348
Lovebird1	1024 $\times$ 768	80	4, 8	6	64.9756
Newspaper	1024 $\times$ 768	56	2, 6	4	61.1012
Balloons	1024 $\times$ 768	6	1, 5	3	47.6410
Kendo	1024 $\times$ 768	10	1, 5	3	48.6635
Undo Dancer	1920 $\times$ 1088	66	1, 9	5	64.1033
GT Fly	1920 $\times$ 1088	150	1, 9	5	55.5549
Poznan street	1920 $\times$ 1088	26	3, 5	4	61.3494
Poznan Hall2	1920 $\times$ 1088	150	5, 7	6	23.5174
Shark	1920 $\times$ 1088	220	1, 9	5	48.6635

Table 3.5: Type of DIBR method

DIBR method	inter-view or single view (extrapolation)
VSRS2 [91]	inter-view
Zhu's [92]	inter-view
Criminisi's [93]	single view (extrapolation)
Luo's [94]	single view (extrapolation)
HHF [95]	single view (extrapolation)
LDI [96]	single view (extrapolation)
VSRS1 [91]	single view (extrapolation)
Ahn's [97]	single view (extrapolation)

3.5.3 DIBR algorithms benchmarking using our database

We also compared the performances of several existing objective IQA metrics on the proposed database. The obtained PLCC, RMSE, SROCC values are given in Table 3.6. It can be noticed at once that the performances of these metrics on the presented database are quite bad (no PLCC value more than 70%). Among which, the proposed metrics $PSNR'$ (SCDM), SC-IQA and the side view based FR metric LOGS perform the best in terms of the PLCC on this database. Especially for NIQSV+, NIQSV, NIQE and BliindS2 NR metrics, they show weak correlations with the subjective results.

Table 3.6: PLCC, RMSE and SROCC between DMOS and objective metrics, where "SV FR metric" indicates the side view based FR metric

Metric		PLCC	RMSE	SROCC
FR 2D metrics	PSNR	0.6012	0.1985	0.5356
	SSIM	0.4016	0.2275	0.2395
	MS-SSIM	0.6162	0.1957	0.5355
	IW-PSNR	0.5827	0.2019	0.4973
	IW-SSIM	0.6280	0.1933	0.5950
	UQI	0.4346	0.2237	0.4113
	PSNR-HVS	0.5982	0.1991	0.5195
FR 3D metrics	MP-PSNR	0.5753	0.2032	0.5507
	MP-PSNRr	0.6061	0.1976	0.5873
	MW-PSNR	0.5301	0.2106	0.4845
	MW-PSNRr	0.5403	0.2090	0.4946
	VSQA	0.5576	0.2062	0.4719
	PSNR'	0.6685	0.1844	0.5903
	SC-IQA	0.6856	0.1805	0.6423
SV FR metric	LOGS	0.6687	0.1845	0.6683
NR 3D metrics	NIQSV	0.1759	0.2446	0.1473
	NIQSV+	0.2095	0.2429	0.2190
	APT	0.4225	0.2252	0.4187
NR 2D metrics	NIQE	0.2244	0.2421	0.1360
	BLiindS2	0.2225	0.2422	0.1329
	BIQI	0.4348	0.2237	0.4328

The scatter plot of each IQA metric (not shown here) shows that all methods are incapable of predicting worse qualities (bigger DMOS value indicates worse quality), which is however consistent with the results shown in Table 3.6 where no metric has a PLCC value higher than 0.7. While some of them do sometimes succeed in their prediction of high qualities (consistent with their PLCC values bigger than 0.5). Be

similar to the results in Table 3.6, the NR metrics NIQSV+, NIQSV, NIQE and BliindS2 show little correction with the subjective results, there is large empty regions in the corresponding scatter plots (consistent with their PLCC values smaller than 0.3).

3.6 Conclusions and Perspectives

This chapitre presents four proposed image quality assessment metrics for DIBR-synthesized views (two FR and two NR), followed by our IETR database and a relatively complete benchmarking of the state-of-the-art metrics using this database. All the data of this presented database, including images, the ground truth depth maps and their associated DMOS, is publicly accessible (https://vaader-data.insa-rennes.fr/data/stian/ieeetom/IETR_DIBR_Database.zip), for the improvement of the QoE of DIBR related applications.

The test DIBR-synthesized view dedicated quality metrics (including FR and NR) perform much better on IVC database than on the proposed IETR database. Among which, the performance of NR metrics decrease the most, which is consistent with the results in Tab. 3.6. One reason of this cross datasets performance decrease could be that the DIBR NR metrics NIQSV, NIQSV+ and APT tried to optimize their performances on the IRCCyN/IVC DIBR database where “old-fashioned” artifacts exist. On the new proposed IETR database, they cannot get a good performance when the “old-fashioned” are excluded. This indicates that further work has to be done to exploit deeply the characteristics of these specific distortions, for new objective metrics with a better correlation with subjective scores.

In the current database, only the MPEG MVD source images are included. In the future work, more source images, such as the images from Middlebury database [98], should be considered to make the experiment results more reliable.

3.7 Contributions in this field

In general, our work contributes four new dedicated metrics for DIBR-synthesized view quality assessment, and a new DIBR-synthesized image database for benchmarks and further development of dedicated metrics for this field.

The publications issued from these works are listed below:

[1] S. Tian, **L. Zhang**, L. Morin, O. Déforges. “SC-IQA: Shift compensation based image quality assessment for DIBR-synthesized views”. Visual Communications and Image Processing (VCIP), December 2018, Taichung, Taiwan.

[2] S. Tian, **L. Zhang**, L. Morin, O. Déforges. “A benchmark of DIBR Synthesized View Quality Assessment Metrics on a new database for Immersive Media Applications”. IEEE Transactions on Multimedia; October 2018; 21(5): 1235 - 1247. [IF=3.977]

[3] S. Tian, **L. Zhang**, L. Morin, O. Déforges. “Performance comparison of objective metrics on free-viewpoint videos with different depth coding algorithms”. SPIE Optical Engineering + Applications, August 2018, San Diego, California, USA.

[4] S. Tian, **L. Zhang**, L. Morin, O. Déforges. “A full-reference Image Quality Assessment metric for 3D Synthesized Views”. Image Quality and System Performance Conference, at IS&T Electronic Imaging 2018, 28 January - 1 February 2018, Burlingame, California, USA.

[5] S. Tian, **L. Zhang**, L. Morin, O. Déforges. “NIQSV+: A No Reference Synthesized View Quality Assessment Metric”. IEEE Transactions on Image Processing; December 2017; 27(4):1652-1664. [IF=5.071]

[6] S. Tian, **L. Zhang**, L. Morin, O. Déforges. “NIQSV: A No Reference Image Quality Assessment Metric for 3D Synthesized Views”. ICASSP, March 2017, New Orleans, USA.

PART II

Saliency detection

CHAPTER 4

SALIENCY AND SALIENT OBJECT DETECTION BASICS

4.1 Definitions

Visual saliency is the distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention. Eye-tracking is a well-known technique for analyzing the visual fixations, based on which the saliency map can be generated. The output of the **saliency detection** is a saliency map, a grayscale image that shows the probability distribution of each pixel being under attention. The heat map is a simple colored representation of the continuous saliency map. The **salient object detection** (or salient object segmentation) aims at finding the most conspicuous objects in an image that highly catches the user's attention, for which the ground truth is often a binary map where the white region corresponds to the salient object(s). The difference between the salient object detection ground truth and the saliency detection ground truth can be seen in Figure 4.1.

The works presented in this report only focus on the objective models which are designed to predict where we look on an image or a video, by using the popular datasets constituting a ground truth in the literature. We do not study how the fixation map is generated from the raw data obtained the eye-tracking system, nor do we study how the saliency/heat map is generated from the fixation map. We only use the ground-truth provided in the literature here.

Figure 4.1: From left to right, these are original image, salient object detection ground truth, and the saliency detection ground truth.

4.2 Performance evaluation metrics for salient object detection

For the Salient Object Detection (SOD), various metrics are used to measure the similarity between the generated saliency map (SM) and the ground truth (GT). The more commonly used metrics are:

- Mean Absolute Error (MAE): computed as the average absolute difference between all pixels in SM and Ground truth (GT). A smaller MAE value means a higher similarity and a better performance.

$$\mathrm{MAE} = \frac{1}{h1 \times w1} \sum_{i=1}^{h1 \times w1} |\mathrm{GT}(i) - \mathrm{SM}(i)| \tag{4.1}$$

  where $h1$ is the frame height, $w1$ is the frame width.

- Precision-Recall (P-R) curve [99]: SM is normalized to $[0, 255]$ and converted to a *binary mask (BM)* via a threshold that varies from 0 to 255. For each threshold, a pair of (Precision, Recall) values are computed which are used for plotting P-R curve. The curve closest to the upper right corner (1.0, 1.0) corresponds to the best performance.

$$\mathrm{Precision} = \frac{|\mathrm{BM} \bigcap \mathrm{GT}|}{|\mathrm{BM}|}, \quad \mathrm{Recall} = \frac{|\mathrm{BM} \bigcap \mathrm{GT}|}{|\mathrm{GT}|} \tag{4.2}$$

- F-measure: an adaptive threshold T is used to binarize SM to a BM, and then the

pair of (Precision, Recall) values are fused to evaluate the global performance:

$$\mathrm{F-measure} = \frac{(1+\beta^2) \times (\mathrm{Precision} \times \mathrm{Recall})}{(\beta^2 \times \mathrm{Precision} + \mathrm{Recall})} \tag{4.3}$$

β^2 is set to 0.3, and T is set to be the minimum value between T'_α and T_α as in the method [100].

$$T'_\alpha = \max(\mathrm{SM}(i)) \quad 1 \leq i \leq h1 \times w1, \quad T_\alpha = \frac{2}{h1 \times w1} \sum_{i=1}^{h1 \times w1} \mathrm{SM}(i) \tag{4.4}$$

A higher F-measure, Precision and Recall values mean a better performance.

For video SOD evaluation, the metrics values are firstly computed over each video, and secondly computed the mean values over all videos in each dataset.

4.3 Performance evaluation metrics for saliency detection

Four popular metrics adopted in several benchmarking are used in this paper for evaluating the saliency detection performance.

The Kullback-Leibler Divergence (KLD) measures the dissimilarity under the loss of information between predicted saliency distribution and ground-truth distribution. It is defined as:

$$L_{KLD}(P, Q^D) = \sum_i Q_i^D \log(\frac{Q_i^D}{P_i + \epsilon} + \epsilon), \tag{4.5}$$

where P and Q^D indicate the predicted saliency map and the ground-truth density distribution, respectively. i represents the ith pixel and ϵ is a regularization constant. The lower value of KLD means the higher similarity of two distribution.

The Pearson's Correlation Coefficient (CC) symmetrically calculates the linear relationship between two distributions. It penalizes false positives and false negatives equally. It is defined as

$$L_{CC}(P, Q^D) = \frac{\sigma(P, Q^D)}{\sigma(P) \cdot \sigma(Q^D)}, \tag{4.6}$$

where $\sigma(P, Q^D)$ is the covariance of P and Q^D, $\sigma(P)$ and $\sigma(Q^D)$ are the standard deviations of P and Q^D, respectively. The value of CC ranges from -1 to $+1$, where $+1$ indicates a perfect correlation, and -1 indicates a perfect correlation in opposite direction, and 0 indicates no correlation.

The Normalized Scanpath Saliency (NSS) measures the correspondence between predicted saliency map and ground-truth binary fixation map via computing the average of normalized predicted saliency map at fixation locations. NSS is defined as

$$L_{NSS}(P, Q^B) = \frac{1}{N} \sum_i \frac{P_i - \mu(P)}{\sigma(P)} \cdot Q_i^D, \tag{4.7}$$

where Q^B indicates the ground-truth binary fixation map, i indicates the i^{th} pixel, and N is the total number of fixated points. NSS of value 0 represents chance and positive value represents the correspondence above chance and negative value represents anti-correspondence.

The Area Under the receiver operating characteristic Curve (AUC) computes the area under the curve of true positive rate versus false positive rate for various thresholds referred to the ground-truth fixation map. In this report, we use the AUC-judd [101], where true positive/negative values is the summation of saliency value above threshold at fixated/unfixated pixels.

CHAPTER 5

SALIENT OBJECT DETECTION IN 2D NATURAL VIDEOS

5.1 Introduction

Salient object detection in images/videos plays an important role as a pre-processing step in many image processing applications such as autonomous driving [102], video re-targeting, surveillance and monitoring, person re-identification, ROI (region-of-interest) based compression and visual tracking and quality assessment. For example, in image quality assessment, the sensitivity of the human visual system to various visual signals is important. As salient object detection and image quality assessment are both related to how human vision system perceives an image, researchers incorporate saliency information to image quality assessment models aiming at improving their performance. One usual way is to adopt salient object detection as a weighting function to reflect the importance region in an image, like what we did in the SC-IQA (cf. section 3.2.2) for the synthesized view quality assessment.

In [99], traditional methods for SOD are categorized in two different ways depending on the types operation or attributes they exploit: 1) pixel/patch-based vs. superpixel/region-based ; 2) intrinsic cues (from the input image itself) vs. extrinsic cues (e.g. user annotations, depth map, or statistical information of similar images).

Recently, with the renaissance of deep learning techniques, there is a trend in the SOD domain to use deep-learning based methods because its significant improvement of performances. Since the first introduction in 2015, these algorithms have soon shown superior performance over traditional methods, and kept residing the top of various benchmarking leaderboards.

To better understand the state-of-the-art, we conducted a survey of deep-learning based methods for video SOD. In addition, two SOD algorithms were proposed: one traditional method - Virtual Border and Guided Filter-based (VBGF) algorithm, and one

deep-learning based method - an extension of the VBGF.

5.2 Comparative study of deep-learning based methods for video SOD

Table 5.1 lists the most relevant works, from which we can see that former works mainly focus on traditional methods. Among the recent works related to deep-learning methods, the survey presented in [103] is only for images; and the benchmark [104] only compares deep-learning methods proposed for images with traditional methods proposed for videos. The survey of existing deep-learning methods for SOD in videos is less explored. In this section, we taxonomically review the existing deep SOD methods (before 2018) and assess their performance generality on the most popular large-scale datasets :

- VOS [104] : a recently published large dataset for video SOD, which is based on human eye fixation. These videos are grouped into two subsets: 1) VOS-E contains easy videos which usually contain obvious foreground objects with many different types of slow camera motion. 2) VOS-N contains normal videos which contain complex or highly dynamic foreground objects, and dynamic or cluttered background.
- FBMS [105, 106] : a dataset originally designed for moving object segmentation. Moving objects attract large attention and thus can be regarded as salient objects in videos.
- DAVIS 2016-val [107] : a popular video dataset for video foreground segmentation. It is widely used for video SOD, because of the foreground properties (most of the objects in the video sequences have distinct colors, which can be regarded as salient objects).
- DAVIS-2017-val [108]) : an extension of DAVIS-2016 dataset.

5.2.1 Taxonomy of deep video SOD methods

According to whether the used neural network has to be trained, existing methods can be classified into two categories: 1) **without-training models** and 2) **with-training**

Table 5.1: Comparison of the existing survey/benchmark for SOD

	Year	Benchmark	Survey	Traditional	Deep-learning	Video	Image
[109]	2014	×	✓	✓	×	×	✓
[99]	2014	×	✓	✓	×	×	✓
[110]	2015	✓	×	✓	×	×	✓
[103]	2018	×	✓	×	✓	×	✓
[104]	2018	✓	×	✓	✓	✓	✓

models. The used deep representations of the first category are directly extracted from existing deep networks. If the network is trained with a large-scale datasets, its extracted deep features are supposed to have a good generality performance with the methods of the first category. Thus, this is a simple way to directly use these deep representations for further researches [111, 112]. In the second category, methods usually get more efficient deep representations through their own training phase, where the inputs-outputs relationship is learned by deep architectures. According to their utilization degree of the labeled datasets, the with-training models can be further divided into supervised and weakly-supervised models.

Supervised models need training datasets with pixel-wise annotations. According to the domain of the learned deep representation, supervised methods [100, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123] can be classified into 1) spatial [100, 120, 122, 123]; 2) temporal [121]; 3) or spatio-temporal [118, 119, 113, 116, 117, 114, 115]. Due to the fact that current datasets have limited manually labeled ground truth, some methods, e.g. [113], propose to generate simulated video data using synthesizing methods. Different from supervised methods, **weakly-supervised models** train the network without requiring all training datasets to have corresponding pixel-level annotations. Some models learn to detect the salient object from spatial domain with image-level annotations, based on the assumption that image-level tags can provide the classes of the dominant objects which can be regarded as the salient foregrounds, e.g. [124]. Sometimes, a small number of manually labeled data and a huge amount of weakly labeled data are used together. For example, in [125], one seventh of the frames in a video is manually labeled data and the rest is weakly labeled. Three existing SOD methods are used to generate the weakly labeled data, and their proposed network is trained using both manually and weakly labeled data. Then the weakly labeled data is updated using their proposed network, as well as the three existing SOD methods. Fig. 5.1 shows the classification of the deep-learning based SOD methods.

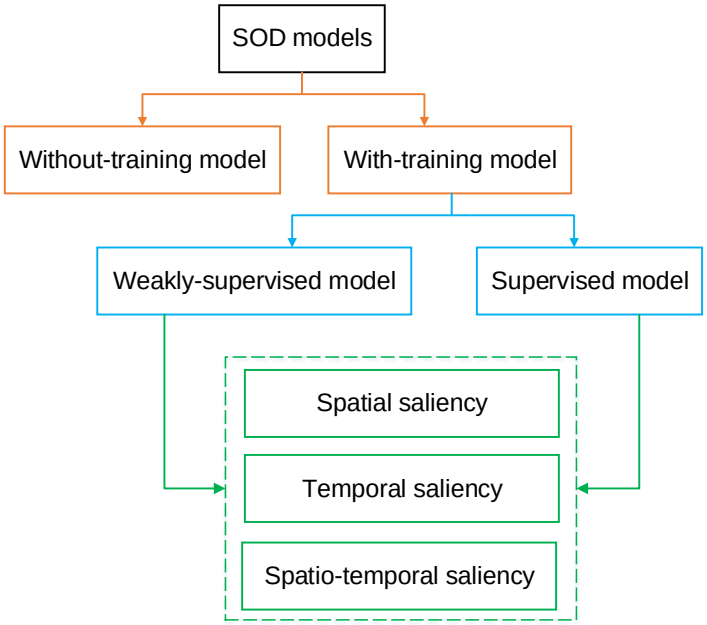


Figure 5.1: Methods classification according to the deep representations generation

5.2.2 Deep video SOD frameworks

This section gives detailed introduction of 11 representative methods, which the source codes or saliency results are provided by the authors. Among them, Chen *et al.* [111] propose a without-training model, and methods in [100, 121, 118, 119, 113, 120, 123, 124, 122, 125] are with-training models. Methods in [100, 121, 118, 119, 113, 120, 123, 122] are supervised models and with those in [124, 125] are weakly-supervised models.

Firstly, the global framework for each method is described and then the deep network designed in each method is analyzed.

Analysis of the frameworks of representative methods

As a matter of convenience, 11 methods are denoted as SCOMd [111], NRF [100], DHSNet [122], OSVOS [120], NLDF [123], LMP [121], SFCN [113], SegFlow [119], LVO [118], WSS [124], SCNN [125].

According to the involved tasks, these 11 frameworks can be divided into two categories: multi-task [124, 119] and single-task [111, 100, 121, 118, 113, 120, 123, 122,

125].

The **multi-task framework** not only predicts the salient objects, but also evaluates other tasks. It exploits the connections between the SOD task and other highly related tasks (such as classification and optical flow), and then improves the SOD performance by making use of the deep representation from these tasks. Specifically, Wang *et al.* [124] propose a weakly-supervised network which has two subnetworks: one is designed for classification and the other is designed for SOD. Firstly, using image-level tags as the ground truth, detection stream is jointly trained with the classification subnetwork for classification prediction. Secondly, the saliency prediction of the detection subnetwork is used as the ground truth for fine-tuning the detection subnetwork. An iterative Conditional Random Field (CRF) is then built for a further refinement. Both subnetworks share convolutional layers firstly and then are separated on the top of the shared layers, as shown in Fig. 5.2 (a), Cheng *et al.* [119] propose a supervised network which also consists of two subnetworks: the segmentation subnetwork and the flow subnetwork. A bi-directional feature propagation is built between these two networks as shown in Fig. 5.2 (b), and an iterative training is used for optimizing the segmentation task.

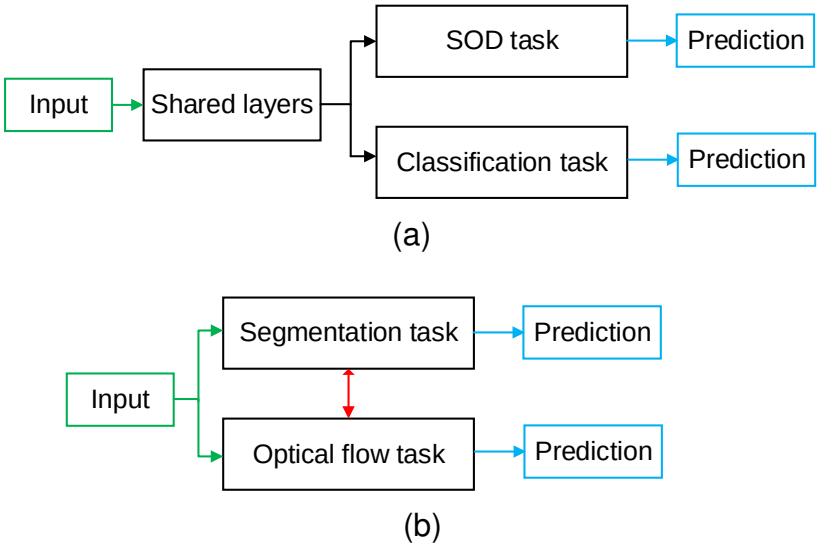


Figure 5.2: Multi-tasks models: (a) WSS and (b) SegFlow.

The **single-task framework** is designed just for the SOD task. Among them, SFCN, SCNN and OSVOS propose two Fully Convolutional Networks (FCN) with the same architecture in their frameworks. From Fig. 5.3 (a), Wang *et al.* [113] use the first FCN

for spatial saliency detection with the input of each frame, and use the other FCN for spatio-temporal saliency detection with the input of adjacent frame pairs and the detected spatial saliency results. The detected spatial saliency results is denoted as SFCNs.

From Fig. 5.3 (b), Tang *et al.* [125] firstly employ one FCN to get a spatial prior map, secondly generate temporal prior map from optical flow fields, thirdly combine these two prior maps to be a spatio-temporal prior map which guides the second FCN to generate the spatio-temporal saliency map. At last, the output saliency map is optimized by a CRF model. From Fig. 5.3 (c), Caelles *et al.* [120] use the first FCN as a foreground branch and use the second FCN as a contour branch. The output of the first FCN is combined with that from the second FCN to get the final foreground prediction. Note that Cheng *et al.* [119] and Caelles *et al.* [120] apply online training step, which improves the accuracy by using the ground truth of the first frame, but this is not considered in this chapter.

The SCOMd, NRF, LMP, DHSNet and NLDF models only adopt one network in their single-task frameworks. In SCOMd, the authors use the deep spatial features from the network proposed in [126, 127], instead of the traditional features, to define a new motion energy for SOD in video. In NRF, the authors firstly obtain the initial salient object and background estimation with their complementary convolutional neural network, and then construct a neighborhood reversible flow to propagate salient object and background along the most reliable inter-frame correspondences. In LMP, the authors detect motion patterns in videos with designed motion pattern network, and then use the spatial objectness cue and a CRF model to refine the results. The SCOMd, NRF and LMP are summarized in Fig. 5.4 (a). DHSNet and NLDF, as in Fig. 5.4 (b), are end-to-end training networks without any other processing. While, In LVO, the authors firstly use the network proposed in [128] to extract deep spatial features in the appearance stream, and then adopt the network proposed in [121] to detect motion patterns in the motion stream, and thirdly build a visual memory module which inputs the concatenation of appearance and motion streams to get the prediction. At last, the CRF model is applied to the network output. The LVO is shown in Fig. 5.4 (c).

In Table 5.2, above methods are summarized considering the detailed techniques (deep-learning or traditional) used in each domain (spatial, temporal or spatio-temporal fusion) for saliency detection.

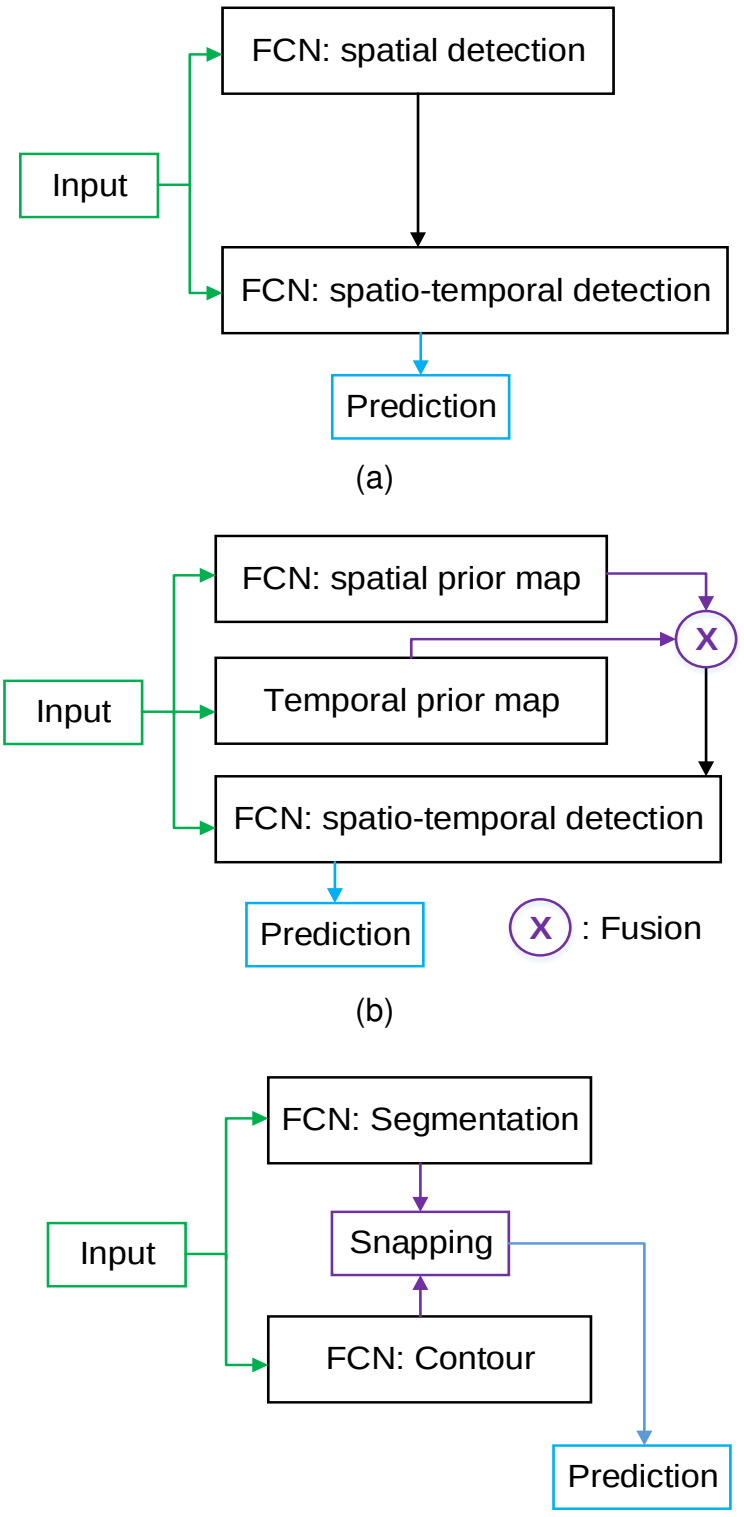


Figure 5.3: Single-task models: (a) SFCN, (b) SCNN and (c) OSVOS

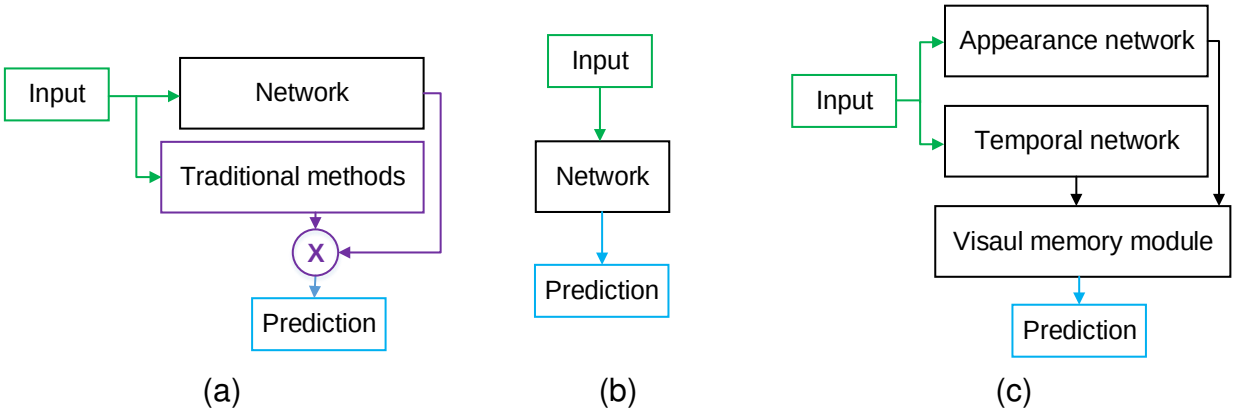


Figure 5.4: Single-task models: (a) SCOMd, NRF, LMP, (b) DHSNet, (c) NLDF, LVO.

Table 5.2: Techniques used in each domain for saliency detection("x" indicates that the method is not based on corresponding technique).

Methods	Spatial		Temporal		Fused spatio-temporal	
	deep	traditional	deep	traditional	deep	traditional
DHSNet[122]	✓	x	x	x	x	x
NLDF[123]	✓	x	x	x	x	x
WSS[124]	✓	x	x	x	x	x
OSVOS[120]	✓	x	x	x	x	x
SCOMd[111]	✓	x	x	✓	x	✓
SFCN[113]	✓	x	x	x	✓	x
NRF[100]	✓	x	x	✓	x	✓
LMP[121]	x	✓	✓	x	✓	x
SCNN[125]	✓	x	✓	✓	✓	✓
SegFlow[119]	✓	x	x	x	✓	x
LVO[118]	✓	x	✓	x	✓	✓

Analysis of the deep networks of representative methods

In this part, we analyze the networks designed in the representative methods.

A typical network for SOD is usually an encoder-decoder network, and hierarchical features are generated layer by layer. According to the feature scales used for saliency prediction, above networks can be divided into **single-scale network** and **multi-scale network**. The former [113, 125] only employs the top-level feature maps for saliency prediction as shown in Fig. 5.5.

The latter, e.g. [100, 119, 123, 120, 121, 124, 122], use skip connections [121, 119, 124, 120, 123, 122] or "À Trous" Pyramid Pooling (ASPP) [100] to employ multi-

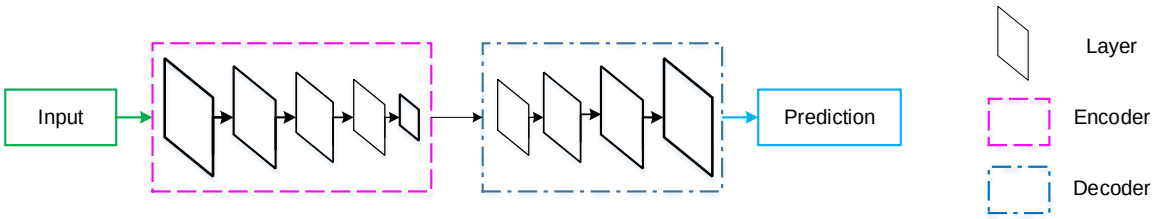


Figure 5.5: Single-scale network

scale feature maps for prediction. These multi-scale networks are illustrated in Fig. 5.6. Specifically, Tokmakov *et al.* [121] add skip connections from the encoder features to the mirror decoder features, which benefits the decoder features with finer details. Cheng *et al.* [119] and Wang *et al.* [124] mainly use feature maps from 3rd to 5th layers for predicting the final output. Luo *et al.* [123] add multiple skip connections to fully employ the deep information. Liu *et al.* [122] add skip connections between mirror layers, but with multiple predictions. Four predictions in Fig. 5.6 (d) are used in the training step. And only the last one is used to generate the final saliency result. Caelles *et al.* [120] add skip connections from the low-level layer to the high-level layer. Feature maps obtained from each layer are fused into a single output. Li *et al.* [100] use three parallel modules with ASPP to capture the multi-scale information. The outputs (Prediction1 and Prediction2 in Fig. 5.6 (f)) are both used to generate the saliency result.

Table 5.3 summarizes the used backbone and training datasets for each mentioned representative method.

Table 5.3: Backbone and Training datasets ("x" indicates that the method is not based on any backbone or the method is without-training)

Methods	Backbone	Training datasets
SCOMd[111]	VGG16	x
NRF[100]	VGG16	HKU-IS,MSRA10K,CSSD,DUT-OMRON
DHSNet[122]	VGG16	MSRA10K,DUT-OMRON
OSVOS[120]	VGG16	DAVIS 2016-train,PASCAL-Context
NLDF[123]	VGG16	MSRA-B
LMP[121]	x	FlyingThings3D
SFCN[113]	VGG16	MSRA10K,SegTrackV2,DUT-OMRON,FBMS-training
SegFlow[119]	ResNet101,FlowNetS	DAVIS 2016-train,MPI Sintel[129],KITTI,Scene Flow
LVO[118]	VGG16	DAVIS 2016-train
WSS[124]	VGG16	DUTS
SCNN[125]	VGG16	MSRA10K,SegTrackV2,FBMS-training

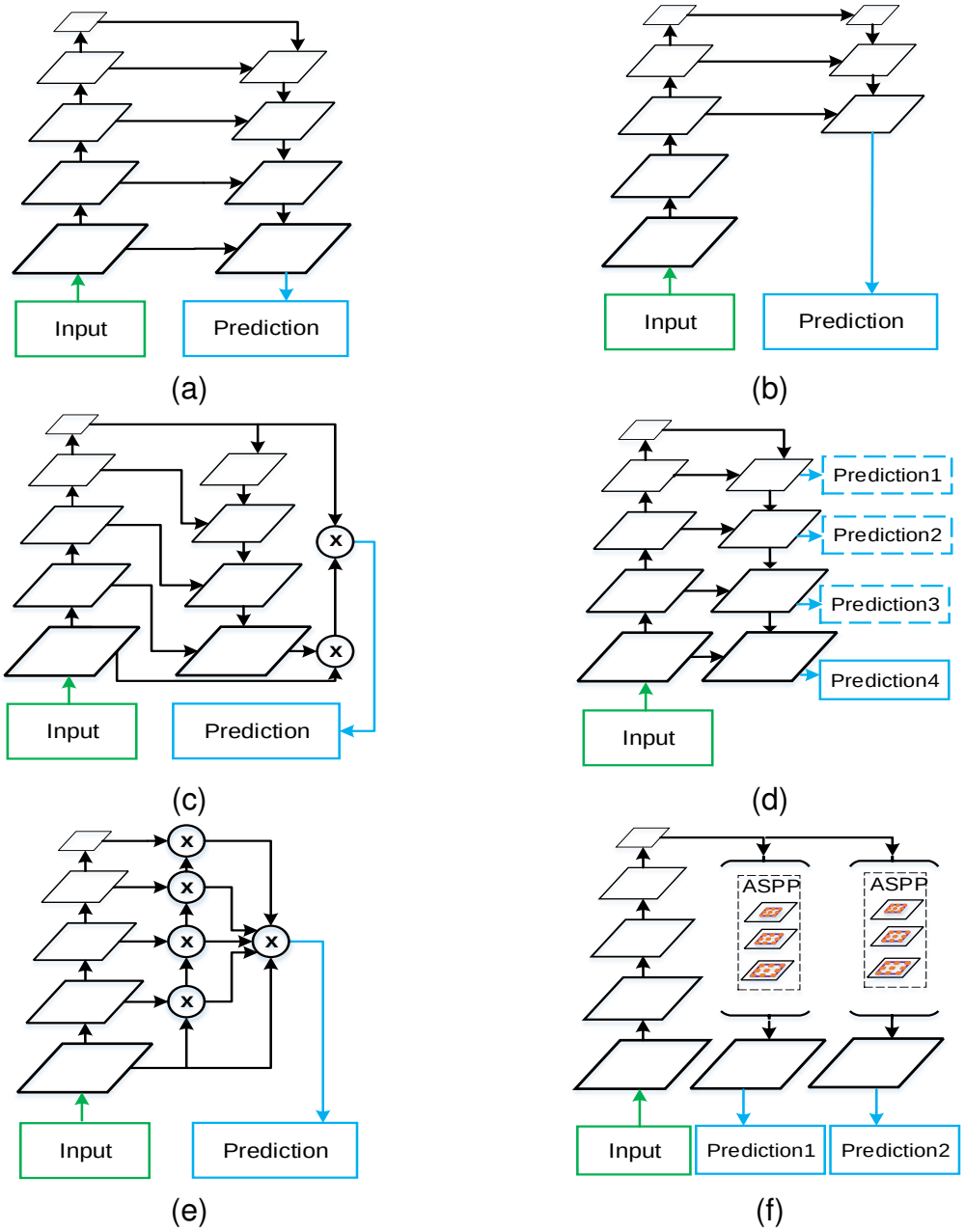


Figure 5.6: Multi-scale networks: (a) [121], (b) [119, 124], (c) [123], (d) [122], (e) [120], (f) [100].

Networks for SOD often built the encoder network based on a backbone (i.e. an existing trained model with published weights). Image classification networks (e.g. VGG [130] and ResNet [131]) are commonly used as backbones. These networks [130, 131] are trained on large-scale image datasets and have a strong ability to learn both low-level and high-level features. Note that various networks are proposed based on VGGNet or ResNet for dense prediction. FlowNetS [132] is only used for estimating the optical flow and the baseline in [119] to obtain the temporal feature.

Various training datasets are used for networks to learn deep representations: Image SOD datasets (e.g. MSRA-B, MSRA10K [133], DUT-OMRON [134], HKU-IS [135] and CSSD [136]) are used in most methods, e.g. [122, 123, 100, 113, 125]; image object segmentation datasets (e.g. DUTS [124]) are used in [124]; video object segmentation datasets (e.g. SegTrackV2 [137], DAVIS 2016-train) are used in methods [120, 118, 113, 119, 125]; contour datasets (e.g. PASCAL-Context [138]) are used in [120]; moving object segmentation datasets (e.g. FBMS-training [105]) are used in methods[125, 113]; optical flow datasets (FlyingThings3D [139]) are used in [121]; and datasets (MPI Sintel [129], KITTI [140], Scene Flow [106]) are used in [119]. Besides, some methods generate new datasets from existing datasets: Wang *et al.* [113] create synthesized video dataset due to the limitation of video SOD datasets, and Tokmakov *et al.* [118] create training sequences which simulate cases where the object stops moving.

During the training phase, a network learns all the parameters via minimizing errors between the result and the ground truth. A loss function is used to compute this error. The "cross entropy" is commonly used for SOD [122, 119, 118, 100, 123]. Given the generated SM and GT, the cross entropy loss *P* is given by Eq (5.1).

$$\mathrm{P} = -\sum_{i=1}^{h1\times w1} (g_i \log s_i + (1-g_i)\log(1-s_i)) \tag{5.1}$$

where h_1 is the frame height, w_1 is the frame width, $g_i \in \mathrm{GT}$ and $s_i \in \mathrm{SM}$. Since the numbers of salient and non-salient pixels are not balanced, the "weighted cross entropy", given by Eq (5.2), is more commonly used for SOD [125, 113, 120].

$$\mathrm{P} = -\sum_{i=1}^{h1\times w1} ((1-R)\times g_i \log s_i + R\times(1-g_i)\log(1-s_i)) \tag{5.2}$$

where R is the ratio of the number of salient pixels in GT over that of all pixels in GT.

Besides, motivated by the successful application of boundary Intersection over Union (IOU) loss in medical image segmentation [141], Luo *et al.* [123] add a boundary IOU loss, given by Eq (5.3), for SOD.

$$\mathrm{IOUloss} = 1 - \frac{2\left|C_{\mathrm{GT}} \bigcap C_{\mathrm{SM}}\right|}{\left|C_{\mathrm{GT}}\right| + \left|C_{\mathrm{SM}}\right|} \tag{5.3}$$

where C_{GT} and C_{SM} are contours pixels of GT and SM respectively, which are obtained using the magnitude of Sobel operator followed by a tanh activation. In order to prevent learning high responses at all locations, Wang *et al.* [124] apply sparse regularization on the generated saliency map ($\|\mathrm{SM}\|_1$) to reduce background noise during pre-training phases.

5.3 Proposed methods for video SOD: VBGF and its extension VBGFd

5.3.1 VBGF

In 2016, we firstly proposed a traditional method based on Virtual Border and Guided Filter (VBGF). It belongs to pixel-based model with intrinsic cues, the main shortcoming of which is that they cannot preserve well the contours of objects. This proposed method tackled this problem.

The block-diagram of the proposed Virtual Border and Guided Filter-based salient object detection for videos (VBGF) method is shown in Fig.5.7.

1) Spatial saliency detection (SD) part

In this part, Spatial saliency detection (SD), the virtual border-based distance transform in spatial domain, is designed.

Virtual border building Instead of using the frame border pixels as the seed set, we propose to add virtual borders around the original frame to obtain with-virtual-border frame. The virtual border, calculated using original frame border pixel values, is used to get the new seed set. Specifically, the virtual border is built in four steps (as shown

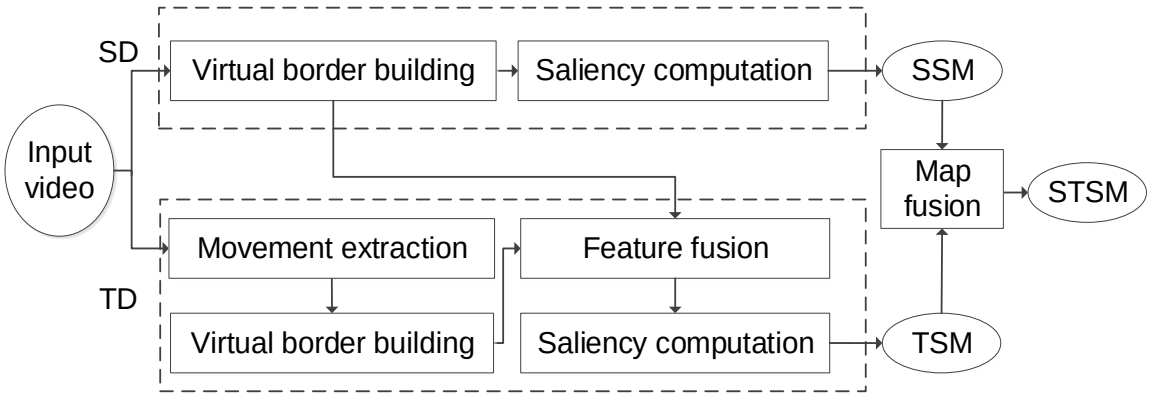


Figure 5.7: The proposed block-diagram. SD: Spatial saliency detection; SSM: Spatial saliency map; TD: Temporal saliency detection; TSM: Temporal saliency map; STSM: Spatio-temporal saliency map.

in Fig.5.8): Frame Border Selection, Frame Border Division, Representative Pixel Selection and Virtual Border Padding.

a) Frame Border Selection: it may suppose that the salient object could be connected with two or more borders. However, from the existing video datasets we observe that: in usual cases if the salient object appears in the frame border, it is often connected with only one border. Here for the sake of simplicity of the presentation, we select one original frame border to build the virtual border by two steps:

- Fast iterative Minimum barrier distance transform algorithm (FastMBD) [142] is applied to frame α to obtain the map M as Eq (5.4).

$$M = \frac{1}{3}({M_1}' + {M_2}' + {M_3}'). \tag{5.4}$$

  where ${M_1}'$, ${M_2}'$ and ${M_3}'$ are obtained respectively from three color channels of frame α in the CIELab color space. For each color channel I with the size of $h_1 \times w_1$, M' is generated as follows: If the pixel $x \in r_1$ (r_1 being the border of the frame α), its value in M' is 0. If pixel $x \in r_2$ (r_2 being the non-border of the frame), its value in M' is initialized as ∞. Let the 4-adjacent pixels around a pixel x in the region r_2 be x_{up} (up pixel), x_{left} (left pixel), x_{down} (down pixel) and x_{right} (right pixel). Then M' is updated in three steps, by using two auxiliary maps τ and ψ which are initialized by the pixel values in each channel of the original image.

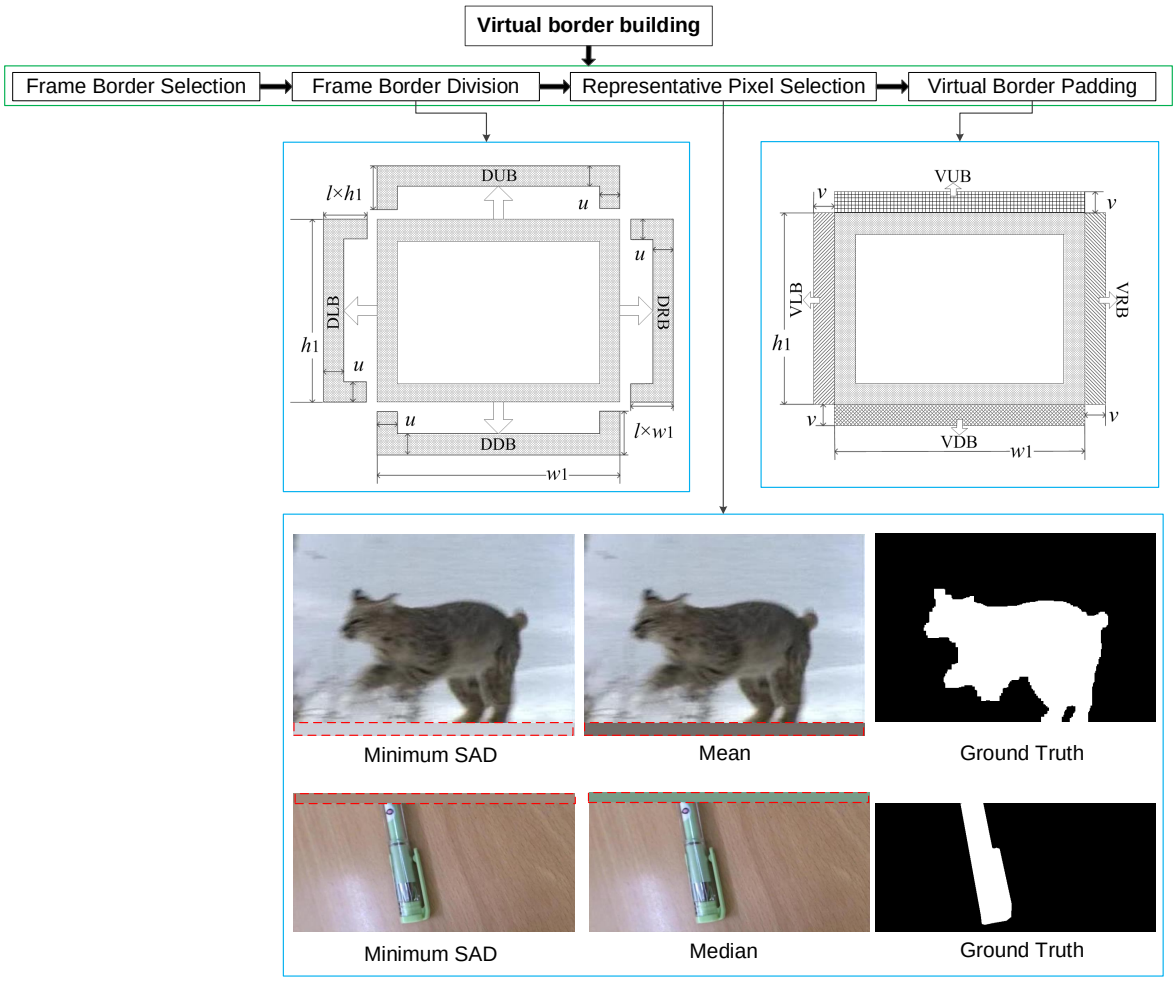


Figure 5.8: Virtual border building: (1) the bottom left one: generating the divided border from the highlighted frame border (with width u), h_1 is the frame height, w_1 is the frame width and l is the ratio of the corresponding border length, four divided borders: the DUB, the DDB, the DLB and the DRB are shown; (2) the bottom middle one: two examples of the representative pixel selection, the red dotted line denotes the virtual border padded with the selected representative pixel; (3) the bottom right one: building and padding the virtual border (with size v) with representative pixel value, four virtual borders: VUB, the VDB, the VLB and the VRB, are shown in four different textures.

Firstly, M' and the auxiliary maps are updated in raster scan order using Eq (5.5). Secondly, M' and the auxiliary maps are updated in inverse raster scan order like in Eq (5.5) but $y \in \{x_{\text{down}}, x_{\text{right}}\}$. Thirdly, M' and the auxiliary maps are updated in raster scan order again.

$$\text{if} \quad M'(x) > O_y(x) \begin{cases} M'(x) = O_y(x) \\ \tau_y = \max\{\tau_y, I(x)\} \\ \psi_y = \min\{\psi_y, I(x)\} \end{cases} \qquad (5.5)$$
$$y \in \{x_{\text{up}}, x_{\text{left}}\}$$

where $O_y(x) = \max\{\tau_y, I(x)\} - \min\{\psi_y, I(x)\}$.

- The frame border nearest to the non-zero region in the map M is selected to build the virtual border. Here, the threshold δ is used to determine the non-zero region.

b) Frame Border Division: after one original border selected, the corresponding divided border is obtained from the original frame border (with width u). The DUB, the DDB, the DLB and the DRB are shown in the bottom left part in Fig.5.8. The reason lying behind this division is that: the region in the frame corner is often connected with two borders and its feature is also related to these two borders. Thus, the irregular shape connecting three borders is used to calculate the virtual border. The parameters u and l are selected empirically. In this chapter, u is set to 5 and l is set to 18%. Preliminary experiments showed that these values make the algorithm robust to various background complexities.

c) Representative Pixel Selection: for the generated divided border, sum of absolute differences (SAD) is computed for each pixel by summing all the absolute differences between this pixel and other pixels in the divided border:

$$\text{SAD}(x) = \sum_{x' \in \text{DB}} |I(x) - I(x')| \qquad (5.6)$$

where $\text{DB} \in \{\text{DUB, DDB, DLB and DRB}\}$, I is the feature channel. The pixel having the minimum SAD is selected to represent the divided border. For color images, the SAD is computed by summing the three color channels:

$$\text{colorSAD}(x) = \sum_{x' \in \text{DB}} \sum_{i \in \{r,g,b\}} \left| I^i(x) - I^i(x') \right| \qquad (5.7)$$

We have also considered using the mean or median value of the border's intensities as the representative pixel value. Various experiments conducted on different frames have shown that the minimum SAD choice performs better than the mean and the median values in most of the cases (cf. the 1st example image in Fig.5.8 where the representative pixel is chosen from the salient object instead of the background when using the mean value of the border's intensities). The same way, choosing the median value of the border's intensities as the representative pixel value fails, as can be seen on the 2nd example image in Fig.5.8. As the minimum SAD performs better in most cases and in order to be more robust in all situations, we adopt the minimum SAD in the proposed method.

d) Virtual Border Padding: around the selected original frame border, we build the corresponding virtual border with the above representative pixel to get the with-virtual-border frame D. The VUB, the VDB, the VLB and the VRB are shown in the bottom right part in Fig.5.8. Existing methods usually regard the border (with width 1) to be background and seed sizes are set to be 1. Here we set the virtual border size v to 5, which helps the proposed "virtual border building" to be applied to other distance transform based saliency detection methods.

Saliency computation After the "virtual border building", the spatial saliency map SSM is obtained by apply the FastMBD [142] to the with-virtual-border frame D and then remove the virtual border from the resulted map to obtain the spatial saliency map SSM. One example is given to show the process of spatial saliency detection in Fig.5.9.

2) Temporal saliency detection (TD) part

In temporal saliency detection (TD), given an input video sequence, the movement information is extracted from the whole video and then the salient object is detected from this movement information. This part is related to the method we called TGFV and published in TGFV17 [143].

Movement extraction The optical flow vectors between pairs of successive frames are obtained using a fast optical flow method [144]. Then the optical flow vector is mapped to Munsell color system [145] to produce the color optical flow map E.

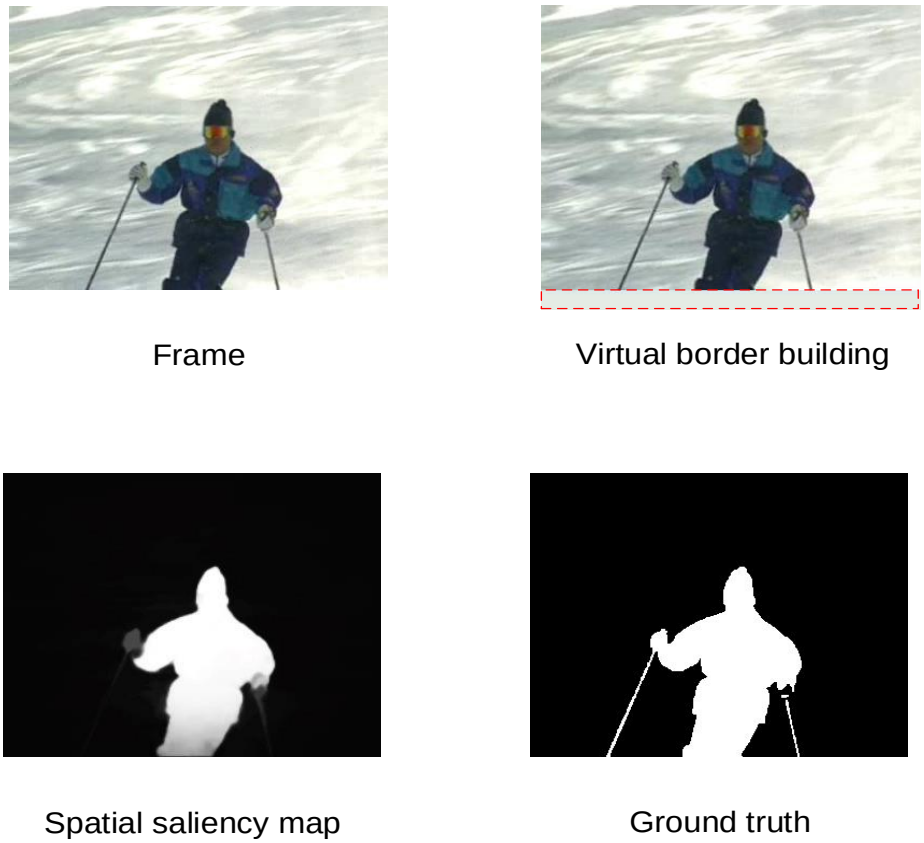


Figure 5.9: (Better viewed in color) An example of the spatial saliency detection. The red dotted line denotes the virtual border.

Virtual border building Based on the backgroundness cue, the global motion is usually connected to E borders. The global motion is mainly generated by the background and camera motion. Camera motion appears in the whole color optical flow map and background motion has a high probability to be connected with E borders. Thus, E borders can reflect the global motion caused by both the background motion and the camera motion. The distance of each pixel to the border pixels of E calculated by the FastMBD [142] can indicate its temporal saliency. The larger the distance, the higher the temporal saliency value. As the same problem in the spatial saliency detection, when the salient object touches frame borders, its movement information also touches E borders. If we directly apply the FastMBD [142] on E, the salient object movement part connected to E borders is hard to be detected. Thus, we add virtual borders on E using the same procedure as described in the SD part to obtain the with-virtual-border

color optical flow map F.

Feature fusion In our spatial saliency detection, only color and luminance features are used to detect the saliency, while edges are inherent features of the image and intrinsically salient for visual perception. Though some researches detect the salient object by considering edges, their results may be still inaccurate. Thus we propose a new Feature fusion way that fuses the spatial edge with the temporal information, considering that: 1) the salient object movement is often bigger than the background movement, thus the background and the salient object are often shown in different colors in the color optical flow map; 2) if the movements within the salient object are different, the salient object cannot be detected completely. If the spatial edges are added onto F, the salient object edges will be enhanced. The pixel's distance in blur edges will be increased if the pixel belongs to the salient object or decreased if the pixel belongs to the background. Thus we performed the guided filtering. The guided filter [146] is a linear filtering process, which involves a guidance image C^1, an input image C^2 and an output image C^3. The C^3 at a pixel i is computed using the filter kernel K which is a function of C^1 but independent of C^2.

$$C^3{}_i = \sum_j K_{ij}(C^1)C^2{}_j, \tag{5.8}$$

where i and j are pixel indexes, and

$$K_{ij}(C^1) = (|\omega_k|)^{-2} \sum_{(i,j)\in\omega_k} (1 + (C^1{}_i - \mu_k)(C^1{}_j - \mu_k)({\sigma_k}^2 + \epsilon)^{-1}), \tag{5.9}$$

where ω_k is the square window centered at the pixel k in C^1, $|\omega_k|$ is the number of pixels in ω_k, ϵ is a regularization parameter, and μ_k and ${\sigma_k}^2$ are the mean and the variance of C^1 in ω_k. The main assumption of the guided filter is a local linear model between C^1 and C^3. Thus, C^3 has an edge if C^1 has an edge.

The proposed method use with-virtual-border frame D as the guidance image and with-virtual-border color optical flow map F as the input image to get the filtered image G as Eq (5.10),

$$G_i = \sum_j |\omega_k|^{-2} \sum_{(i,j)\in\omega_k} (1 + (D_i - \mu_k)(D_j - \mu_k)({\sigma_k}^2 + \epsilon)^{-1})F_j, \tag{5.10}$$

where i and j are pixel indexes, ω_k is the square window centered at the pixel k in D_i, μ_k and ${\sigma_k}^2$ are the mean and the variance of D_i in ω_k. ϵ is set to be 10^{-6}. $|\omega_k|$ is decided by the frame size. Large frame size needs large $|\omega_k|$. We use 20×20 for Fukuchi and FBMS datasets, and use 60×60 for VOS dataset since VOS has larger average frame size than that of Fukuchi and FBMS [104, 147].

Saliency computation The FastMBD [142] is applied on the filtered image G and then the virtual border region is removed to obtain the temporal saliency map TSM. One example is given to show the process of the temporal saliency detection in Fig.5.10.

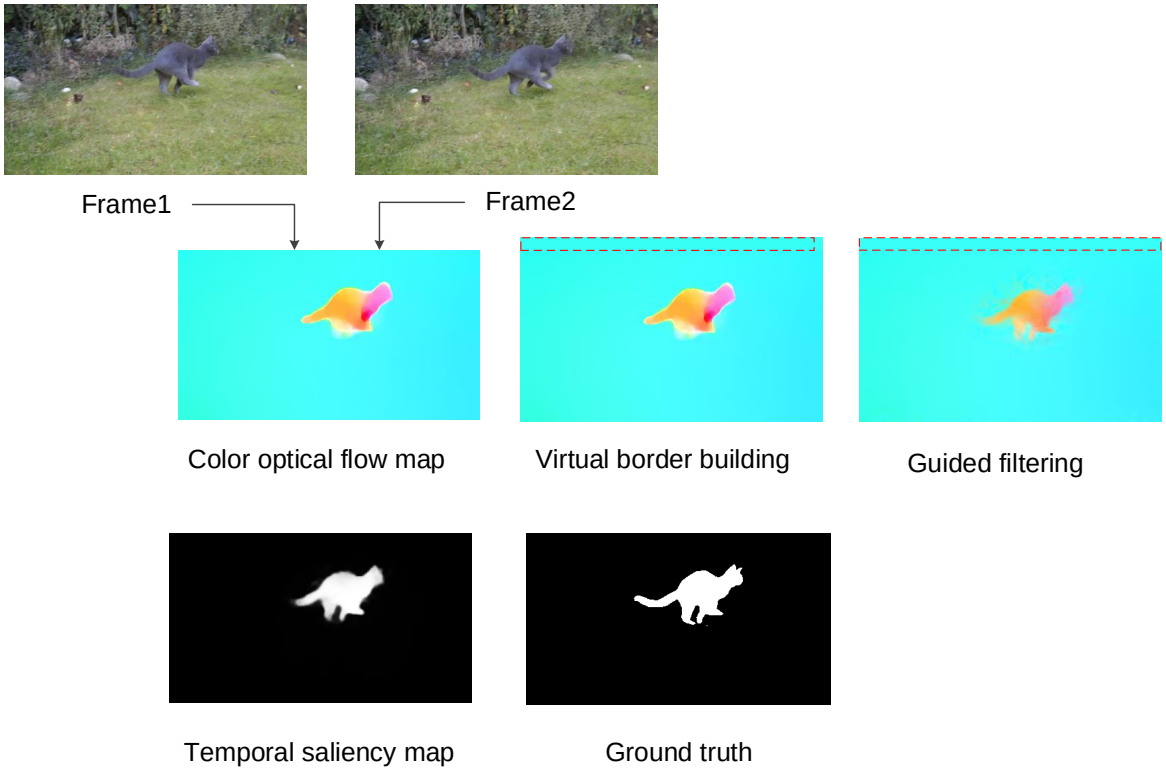


Figure 5.10: An example of the temporal saliency detection. The red dotted line denotes the virtual border.

3) Spatial and temporal saliency maps fusion

Given the spatial saliency map SSM and the temporal saliency map TSM, the fusion is made to obtain STSM by four steps:

- SSM and TSM are firstly fused as Eq (5.11), where $\mathrm{ratio}_1 = \mathrm{mu}_T/(\mathrm{mu}_S + \mathrm{mu}_T)$, $\mathrm{ratio}_2 = 1 - \mathrm{ratio}_1$.

$$\mathrm{STSM} = \mathrm{ratio}_1 \times \mathrm{SSM} + \mathrm{ratio}_2 \times \mathrm{TSM} \tag{5.11}$$

  where mu_S and mu_T are respectively the mean entropies of all the spatial saliency maps and all the temporal saliency maps for a video sequence (with $\varkappa$ the number of frames) as Eq (5.12).

$$\begin{aligned} \mathrm{mu}_S &= \sum_{j=1}^{\varkappa}(-\sum_{j'=1}^{255}(\mathrm{Prob}_{j'}^{S^j} \times \log(\mathrm{Prob}_{j'}^{S^j})))/\varkappa \\ \mathrm{mu}_T &= \sum_{j=1}^{\varkappa}(-\sum_{j'=1}^{255}(\mathrm{Prob}_{j'}^{T^j} \times \log(\mathrm{Prob}_{j'}^{T^j})))/\varkappa \end{aligned} \tag{5.12}$$

  where $\mathrm{Prob}_{j'}^{S^j}$ and $\mathrm{Prob}_{j'}^{T^j}$ are respectively the normalized histogram of j^{th} spatial saliency map and j^{th} temporal saliency map: $\mathrm{Prob}_{j'} = \mathrm{num}_{j'}/(h_1 \times w_1)$, $\mathrm{num}_{j'}$ is the number of pixel (equal to j') in saliency map. Here, the idea is that mu_i ($i = S, T$) are used to decide the confidence of SSM and TSM. The disorder degree of saliency map reflects the difficulty degree to detect the salient objects. If mu_i ($i \in \{S, T\}$) is larger, the saliency detection in this domain is worser.

- STSM is optimized using Eq (5.13)

$$\mathrm{STSM} = \mathrm{SSM} \qquad if \quad \mathrm{mu}_S < \mathrm{mu}_T \tag{5.13}$$

  The frame is often more complex than the color optical flow map, which results in that the disorder degree of SSM is usually larger than that of TSM. If mu_S is smaller than mu_T, it means it is difficult to detect the salient object in TSM. Thus, SSM has a high confidence.

- STSM is optimized using Eq (5.14)

$$\mathrm{STSM} = \mathrm{SSM} \qquad if \quad \sigma_S > \sigma_T \tag{5.14}$$

  σ_S and σ_T are respectively the standard deviations of non-zero regions in two grayscale images H_S and H_T, which are generated by the following steps: firstly, converting frame α from RGB to HSI color space, then eliminating the hue and

saturation information while retaining the luminance to get the grayscale images α'; secondly, using a threshold δ to neglect the pixels with low saliency value from the images SSM and TSM as in Eq (5.15)

$$H_{S_{ij}} = \begin{cases} 0 & if \quad \mathrm{SSM}_{ij} < \delta \\ \alpha'_{ij} & otherwise \end{cases} \quad H_{T_{ij}} = \begin{cases} 0 & if \quad \mathrm{TSM}_{ij} < \delta \\ \alpha'_{ij} & otherwise \end{cases} \tag{5.15}$$

where i and j are pixel indexes in the images. The appearance of the wrongly detected background is mostly different from the salient object in the grayscale image, which results in that H_i ($i \in \{\mathrm{S,T}\}$) contains more luminance values and thus σ_i ($i \in \{\mathrm{S,T}\}$) is smaller. If σ_S is bigger than σ_T, it means SSM has a high confidence.

- Low saliency value (lower than δ) in SSM is decreased to 0.1 times.

  The pixels with low saliency value in saliency map are unimportant for visual saliency but have a large influence in computing the detection confidence. Thus, δ is used to decrease their affection and set to 70 here.

5.3.2 VBGFd

With the trend of the usage of deep networks for the video SOD, the VBGF has been extended to VBGFd in 2017, of which the block-diagram is shown in Fig.5.11. It is performed on an NVIDIA 1080 GPU, and is implemented in Python.

Compared with Fig.5.7, the “Virtual border building” in both “SD” and “TD” blocks is removed. The “Saliency computation” in VBGF is a traditional methods, while the “Saliency computation” in extension of the VBGF (VBGFd) is a deep-salient detection method proposed in [122] - DHSNet (because of its top performances in our comparative study). In VBGFd, the first two steps in the “Map fusion” part use the ratio between the entropies for each frame in Eq.5.11.

5.3.3 Performance evaluation of VBGF and VBGFd

In this section, the large-scale video SOD dataset VOS and its two subsets VOS-E, VOS-N are used to show the performance of VBGF and VBGFd.

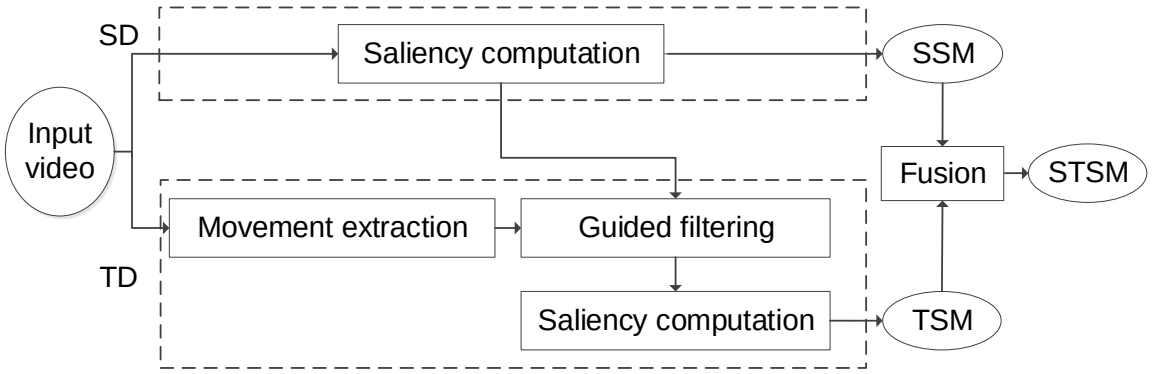


Figure 5.11: The proposed block-diagram. SD: Spatial saliency detection; SSM: Spatial saliency map; TD: Temporal saliency detection; TSM: Temporal saliency map; STSM: Spatio-temporal saliency map.

Ablation study for VBGFd

In Table 5.4, we list the performances of VBGFd with different components. The 3th, 5th and 6th columns show the results of the spatial saliency map, temporal saliency map and spatio-temporal saliency map. The 4th column shows the result of the temporal saliency detection without guided filtering. By comparing the 4th and 5th columns in Table 5.4, the performance is better for all performance evaluation metrics with the "guided filtering". By comparing the 3rd, 5th and 6th columns in Table 5.4, the performance is better for most evaluation metrics when the spatial saliency map and the temporal saliency map are fused together.

Performance of VBGF and VBGFd on VOS dataset

The VBGFd and VBGF were compared with state-of-the-art models on the VOS dataset and its two sub-datasets.

In Table 5.5, Table 5.6 and Table 5.7, we inserted the performance of our proposed models into the the benchmarking table (cf. Table III in the paper [104]) provided with the VOS dataset. Note that here we only list 13 state-of-the-art models (image-based deep-learning and video-based traditional models) reported in [104]. 13 state-of-the-art models are LEGS[148], MCDL[149], MDF[135], ELD[150], DCL[151], RFCN[152], DHSNet[122], SIV[153], FST[154], NLC[155], SAG[156], GF[157] and SSA[104]. These models are categorized into two parts: [I+D] for deep-learning and image-based, [V+U]

Table 5.4: Comparison of the proposed VBGFd componets' performance on dataset VOS, VOS-E, VOS-N. proSSM: proposed spatial saliency map; proTSM: proposed temporal saliency map; proSTSM: proposed spatio-temporal saliency map. The Bold number indicates the best result in each line.

Dataset	Metrics	Proposed VBGFd components			
		proSSM	proTSM without guided filtering	proTSM with guided filtering	proSTSM
VOS-E	Precision↑	0.863	0.398	0.528	**0.881**
	Recall↑	**0.905**	0.380	0.480	0.877
	F-measure↑	0.872	0.394	0.516	**0.880**
	MAE↓	0.049	0.189	0.154	**0.046**
VOS-N	Precision↑	0.649	0.407	0.407	**0.690**
	Recall↑	**0.851**	0.389	0.392	0.806
	F-measure↑	0.686	0.403	0.403	**0.714**
	MAE↓	0.055	0.136	0.132	**0.059**
VOS	Precision↑	0.753	0.403	0.466	**0.783**
	Recall↑	**0.877**	0.385	0.435	0.840
	F-measure↑	0.778	0.399	0.458	**0.795**
	MAE↓	**0.052**	0.162	0.143	0.053

Table 5.5: Performance benchmarking of VBGFd, and VBGF, and 13 state-of-the-art models on the dataset VOS-E. The best three scores in each column are marked in red, green and blue, respectively.

	Models	VOS-E			
		Precision↑	Recall↑	F-measure↑	MAE↓
[I+D]	LEGS	0.820	0.685	0.784	0.193
	MCDL	0.831	0.787	0.821	0.081
	MDF	0.740	0.848	0.762	0.100
	ELD	0.790	0.884	0.810	0.060
	DCL	0.864	0.735	0.830	0.084
	RFCN	0.834	0.820	0.831	0.075
	DHSNet	0.863	0.905	0.872	0.049
[V+U]	SIV	0.693	0.543	0.651	0.204
	FST	0.781	0.903	0.806	0.076
	NLC	0.439	0.421	0.435	0.204
	SAG	0.709	0.814	0.731	0.129
	GF	0.712	0.798	0.730	0.153
	SSA	0.875	0.776	0.850	0.062
	VBGF	0.797	0.773	0.791	0.085
	VBGFd	0.881	0.877	0.880	0.046

Table 5.6: Performance benchmarking of VBGFd, and VBGF, and 13 state-of-the-art models on the dataset VOS-N. The best three scores in each column are marked in red, green and blue, respectively.

	Models	VOS-N			
		Precision↑	Recall↑	F-measure↑	MAE↓
[I+D]	LEGS	0.556	0.593	0.564	0.215
	MCDL	0.570	0.645	0.586	0.085
	MDF	0.527	0.742	0.565	0.098
	ELD	0.569	0.838	0.615	0.081
	DCL	0.583	0.809	0.624	0.079
	RFCN	0.614	0.783	0.646	0.080
	DHSNet	0.649	0.851	0.686	0.055
[V+U]	SIV	0.451	0.523	0.466	0.201
	FST	0.619	0.691	0.634	0.117
	NLC	0.561	0.610	0.572	0.123
	SAG	0.354	0.742	0.402	0.150
	GF	0.346	0.738	0.394	0.331
	SSA	0.660	0.682	0.665	0.103
	VBGF	0.558	0.688	0.583	0.130
	VBGFd	0.690	0.806	0.714	0.059

Table 5.7: Performance benchmarking of VBGFd, and VBGF, and 13 state-of-the-art models on the dataset VOS. The best three scores in each column are marked in red, green and blue, respectively.

	Models	VOS			
		Precision↑	Recall↑	F-measure↑	MAE↓
[I+D]	LEGS	0.684	0.638	0.673	0.204
	MCDL	0.697	0.714	0.701	0.083
	MDF	0.630	0.793	0.661	0.099
	ELD	0.676	0.861	0.712	0.071
	DCL	0.719	0.773	0.731	0.081
	RFCN	0.721	0.801	0.738	0.078
	DHSNet	0.753	0.877	0.778	0.052
[V+U]	SIV	0.568	0.533	0.560	0.203
	FST	0.697	0.794	0.718	0.097
	NLC	0.502	0.518	0.505	0.162
	SAG	0.526	0.777	0.568	0.140
	GF	0.523	0.767	0.565	0.244
	SSA	0.764	0.728	0.755	0.083
	VBGF	0.674	0.729	0.686	0.108
	VBGFd	0.783	0.840	0.795	0.053

for video-based and traditional. From these three tables, we can see that among the tested 15 models, the VBGFd has the best score for 7 times, while the best benchmarked model DHSNet has the best score for 5 times. In general, VBGFd performs the best among the tested models.

5.4 Conclusions and Perspectives

This chapitre presents our work on the video SOD that aims at separating salient objects from background in each frame of a video sequence. We have done an overview of deep-learning methods for this domain, where four popular datasets and five commonly used evaluation metrics were used and our results shows that the methods DHSNet and NRF have the best performances over all the tested databases, before 2018. We also proposed a traditional method (VBGF), as well as its extension integrating the deep-learning technique (VBGFd). The VBGFd achieves the largest number of highest scores for 4 metrics on the recent benchmarking dataset VOS at the moment when it was proposed.

Some future works can be derived from the previous analyses:

- Improve the simple map fusion method in the proposed VBGF and VBGFd using deep learning techniques.

- Employ more video saliency cues: it is valuable to investigate for other deep representations that can improve the quality of video saliency detection. The image object-level cue used in the VBGFd is the most popular choice. Human visual attention usually pays more attention on certain categories, thus the object classification cue can be considered as another choice to detect the video SOD.

- Explore more temporal saliency features and spatio-temporal saliency features: from our experiments, deep-learning technique performs well for detecting the salient object from the spatial domain. Most of the existing video SOD mainly rely on the spatial saliency detection and based on the backbone network. However the goal of video SOD is to detect the object which is salient in the whole video sequence. Further exploration for the direct usage of spatio-temporal networks need to be explored.

- Try weakly-supervised networks [125]: fully supervised models improve detection performance but rely on large training dataset with provided ground truth. Weakly-supervised models that do not rely on large pixel-wise labels attract much attention in recent years. However, its accuracy is still far from satisfactory, and further accuracy

improving is one topic to investigate in the future.

5.5 Contributions in this field

Following the trend in the video SOD field, we proposed a deep-learning based method after a traditional method. An overview of the existing deep-learning based methods has also been done, which helps to better understand the state-of-the-art and may pave the way for future deep models.

Publications issued from the studies in this field:

[1] Q. Wang, **L. Zhang**, K. Kpalma. “VBGF: Virtual Border and Guided Filter-based salient object detection for Videos”. Pattern Recognition, minor review.

[2] Q. Wang, **L. Zhang**, K. Kpalma. “An overview on deep-learning based methods for salient object detection in videos”. Pattern Recognition, submitted.

[3] Q. Wang, **L. Zhang**, K. Kpalma. “Fast filtering-based temporal saliency detection using minimum barrier distance”. ICME2017W, July 2017, Hong Kong, China.

CHAPTER 6

SALIENCY PREDICTION FOR OMNIDIRECTIONAL IMAGES

6.1 Introduction

Virtual Reality (VR), which is one of the fastest growing multimedia technology in the entertainment industry, attracts many attentions due to its capability of providing users the immersive experience in surrounding visual and audio environments. By wearing Head-Mounted Displays (HMDs), users can freely explore the scene in all directions simply by rotating their heads to different point of views when they watch the panorama (or omnidirectional) images or videos which capture all the information in 360° longitude and 180° latitude as a sphere. This interactive property enables users to feel like being in a virtual world. At the same time, it gives rise to various new challenges, such as transmission [158], compression [159], quality assessments [160, 161, 162], etc. Those new challenges are dissimilar to the cases in 2D traditional media since users can actively select the content they would like to watch with the HMD, while they are only allowed to passively receive the given content in 2D traditional videos. Therefore, the saliency (where users pay more attention) prediction in 360° content becomes essential for user behavior analysis and could benefit 360° VR applications [163, 164].

Visual fixation prediction in 360° images can be separated into head movement prediction and head+eye movement prediction [165]. The former predicts the center point in every viewports [166] when users move their heads while watching 360° images. The latter predicts viewer's eye gaze [167]. Although the hypothesis that the center of viewports are observer's eye fixation is followed by [168, 169], Rai *et al.* [170] discovered that the fixation distribution is like a doughnut shape distribution which has probability peaks far away from center by 14 degrees. In our work, we focus on visual fixation prediction based on head+eye movement of which the output is a gray scale saliency map presenting the probabilities of every pixel being seen by viewers with no

specific intention.

Compared to visual attention models for 360° images, those for 2D traditional images have been well developed in recent years [171, 172, 173]. Seminal methods were proposed based on low-level or high-level semantic feature extraction from handcrafted filters [174, 175] or Deep Convolutional Neuron Networks (DCNN) [176, 177, 178, 179] thanks to the establishment of several large scale datasets [180, 181, 182]. Unfortunately, these models are not immediately usable for 360° images because of the severe geometric distortion on the top and bottom areas in equirectangular projection. Furthermore, it is not practical to adjust 2D models by training and testing on 360° images because of 1) the lack of a sufficient large 360° image saliency dataset, and 2) the inherent high resolution problem in 360° images whose optimal resolution is at least 3600×1800 pixels recommended by MPEG-4 3DAV group [183] to provide favorable quality. This image resolution excesses the computational limitation of 2D models based on DCNN.

In this Chapter, we proposed two new saliency prediction models for omnidirectional images based on Generative Adversarial Network (GAN): SalGAN360 and its extension MV-SalGAN360 using multi-resolutional Field of View (FoV) and adaptive weighting losses.

6.2 SalGAN360: Saliency Prediction for omnidirectional images with Generative Adversarial Network

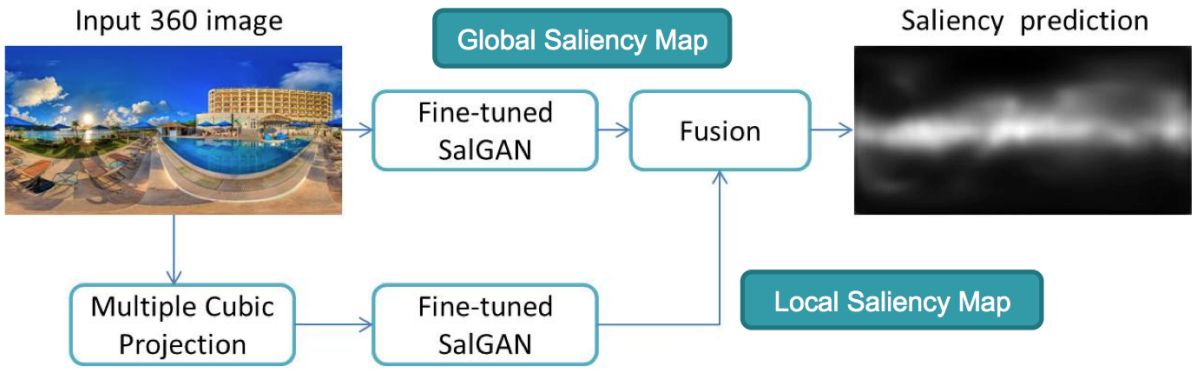


Figure 6.1: Diagram of SalGAN360.

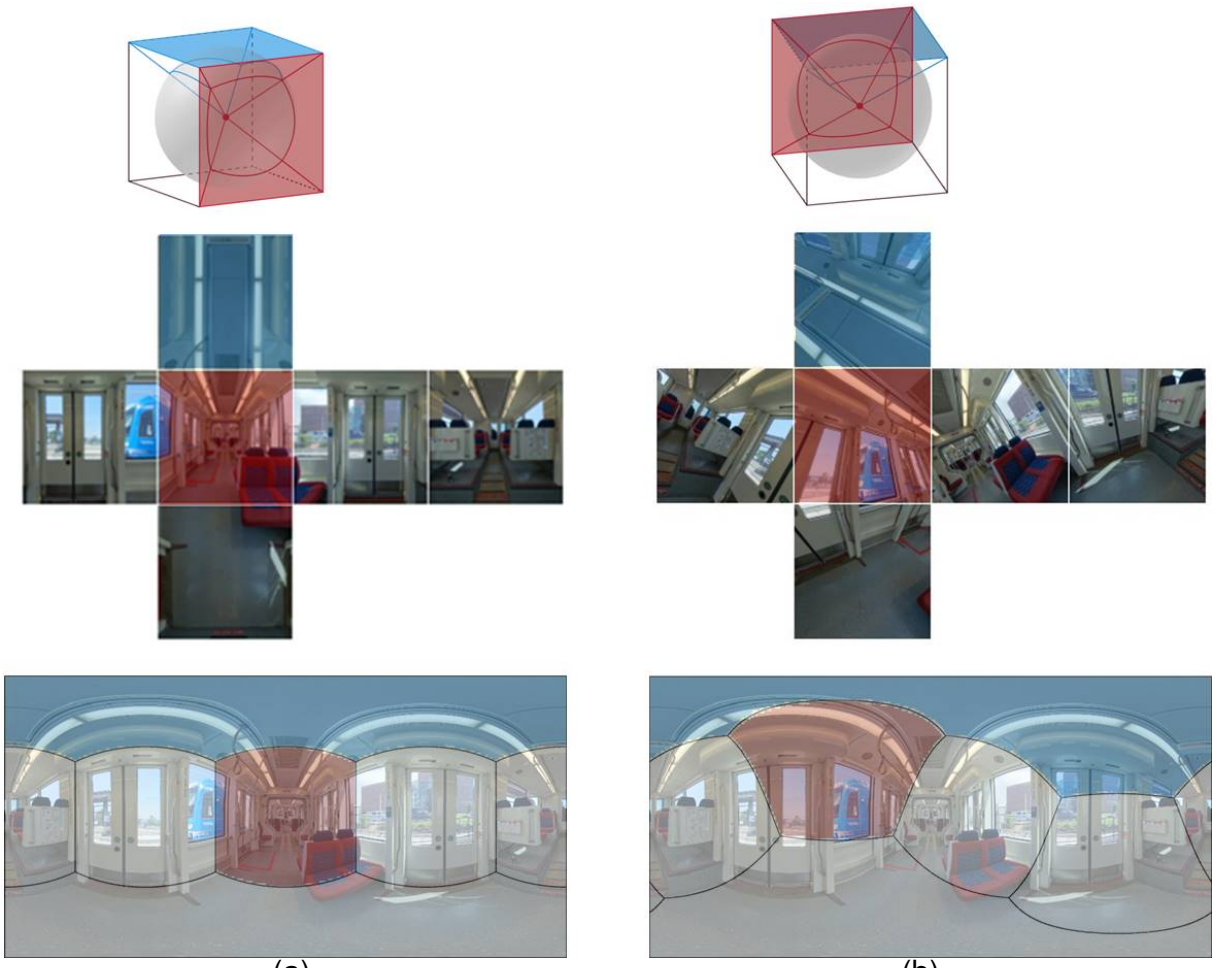
The overall diagram of the SalGAN360 is shown in Figure 6.1. In its 1st part illustrated on the top of Figure 6.1, a fine-tuned SalGAN takes an entire 360° image as the input to detect the global visual attention in all directions. The 2nd part (on the bottom), divides a 360° image with Multiple Cubic Projection (MCP) method into several rectilinear images from different viewports. The rectilinear image is given as input to the fine-tuned SalGAN for the local visual attention detection. Finally, the outputs of all the rectilinear images are integrated into a 360° saliency map with global saliency from the 1st part.

6.2.1 Multiple Cubic Projection

The most common projection of 360° images is equirectangular projection, which induces distortion along with the elevation. This characteristic makes it inappropriate to compute saliency probability directly, since it is far away from what observers actually see. Another popular projection is cubic mapping, which preforms rectilinear projection on 6 cube with 90° FoV each. In each cube face, the distortion is not as obvious as in the equirectangular image, but there are still distortions close to the frontier caused by the discontinuity between the cube faces. To simulate what observers actually see with the Head-Mounted Displays (HMD), we transfer an equirectangular image into multiple cubic maps by rotating the center of cube to multiple horizontal and vertical angles. Figure 6.2 shows the projection from sphere to cube, then to equirectangular format. From the expanded view of cube on the second row, we can see that the distortion of each cube face is slighter than that in equirectangular image on the third row. However, the frontier of cube faces is not continuous with each other. We then rotate the cube direction horizontally and vertically to render other rectilinear images cross the frontier (as shown in Figure 6.2(b)). Each rectilinear image is provided as an input to the saliency prediction model independently to estimate local saliency maps.

6.2.2 Fine tuning of SalGAN

The central element of SalGAN360 is extended from the SalGAN, a Generative Adversarial Network (GAN) composed of two Deep Convolutional Neuron Network (DCNN) (namely generator and discriminator) to predict visual saliency map on traditional 2D images. In order to solve the problem of the lack of a sufficient large 360° image



(a) (b)

Figure 6.2: Illustration of Multiple Cubic Projection. The cube in (a) is centered the same as equirectangular image. The cube in (b) is rotated 30° to the top and 60° to the left.

Table 6.1: Our Training Method on SalGAN

Generator		
Block	Layer	Our Training Method
Conv1	2 conv, max-pool	Fix
Conv2	2 conv, max-pool	
Conv3	3 conv, max-pool	
Conv4	3 conv, max-pool	
Conv5	2 conv, max-poo l	
Uconv5	2 uconv, upscale	Fine Tune
Uconv4	3 uconv, upscale	
Uconv3	2 uconv, upscale	
Uconv2	2 uconv, upscale	
Uconv1	3 uconv, sigmoid	Randomly Initialize
Discriminator		
Conv1	2 conv, max-pool	Fix
Conv2	2 conv, max-pool	
Conv3	3 conv, max-pool	Fine Tune
Fc4	1 fc	
Fc5	1 fc	
Prob	1 fc	

dataset, we fine tune the network by retraining SalGAN initialized with pretrained weights. Table 6.1 details our training method. In generator, we fix the weights of encoder part and fine tune the weights of decoder except the last deconvolutional layers which are trained from random initialization to give more freedom to generate saliency map of 360° image patches. In discriminator, the lower two layers extracting basic features are fixed while decision layers are fine tuned.

Saliency predictions are usually evaluated through different metrics to capture different quality factors. We propose a new loss function of generator given by a combination of three evaluation metrics to improve the performance on different factors. The overall loss function is defined as follows:

$$L = \mu_{BCE} + \sigma_{BCE}(L') \tag{6.1}$$

$$L' = L^{KLdiv}_{normal}(\hat{S}, S^{den}) - L^{CC}_{normal}(\hat{S}, S^{den}) - L^{NSS}_{normal}(\hat{S}, S^{fix}) \tag{6.2}$$

where $\hat{S}$, S^{den} and S^{fix} are respectively the predicted saliency map, the ground truth

density distribution and the ground truth binary fixation map. L' combines three evaluation metrics - KLD , CC and NSS - normalized as follows:

$$L_{normal}(\hat{S}, S^{den}) = \frac{L(\hat{S}, S^{den}) - \mu}{\sigma} \tag{6.3}$$

where μ and σ are mean and standard deviation computed from the scores of evaluation metrics on the saliency maps predicted from the SalGAN. During fine tuning, eq.(1) is used to set the range of L' the same as that of binary cross entropy (Binary Cross Entropy (BCE)), which is defined as:

$$L_{BCE} = -\frac{1}{N}\sum_{j=1}^{N} S_j^{den} \log \hat{S}_j + (1 - S_j^{den}) \log (1 - \hat{S}_j) \tag{6.4}$$

As in [177], the final loss function for the generator during adversarial training can be expressed as:

$$L_{GAN} = \alpha L - \log D(I, \hat{S}) \tag{6.5}$$

where $D(I, \hat{S})$ is the probability of fooling the discriminator. We use the hyperparameter $\alpha = 0.05$, the same as in SalGAN.

6.2.3 Fusion method

The proposed fusion method firstly re-project every 6 local saliency maps in the same cube into equirectangular format. It should be noted that the cube is rotated back to the same direction as that of the input 360° image. We overlap all the equirectangular saliency maps from each cube by simply using the mean value (we assume that observer pays attention to all the contents from different viewport in local saliency map). Local saliency map is then combined linearly with global saliency map from the 1st part of the SalGAN360:

$$\hat{S}_{360} \quad = \quad 0.5\hat{S}_{Global} + 0.5\hat{S}_{Local} \tag{6.6}$$

where $\hat{S}_{360}$ is the final output of the SalGAN360, $\hat{S}_{Global}$ and $\hat{S}_{Local}$ are the predicted global and local saliency maps.

6.3 MV-SalGAN360: A multi-resolutional FoV extension of SalGAN360 with adaptive weighting losses

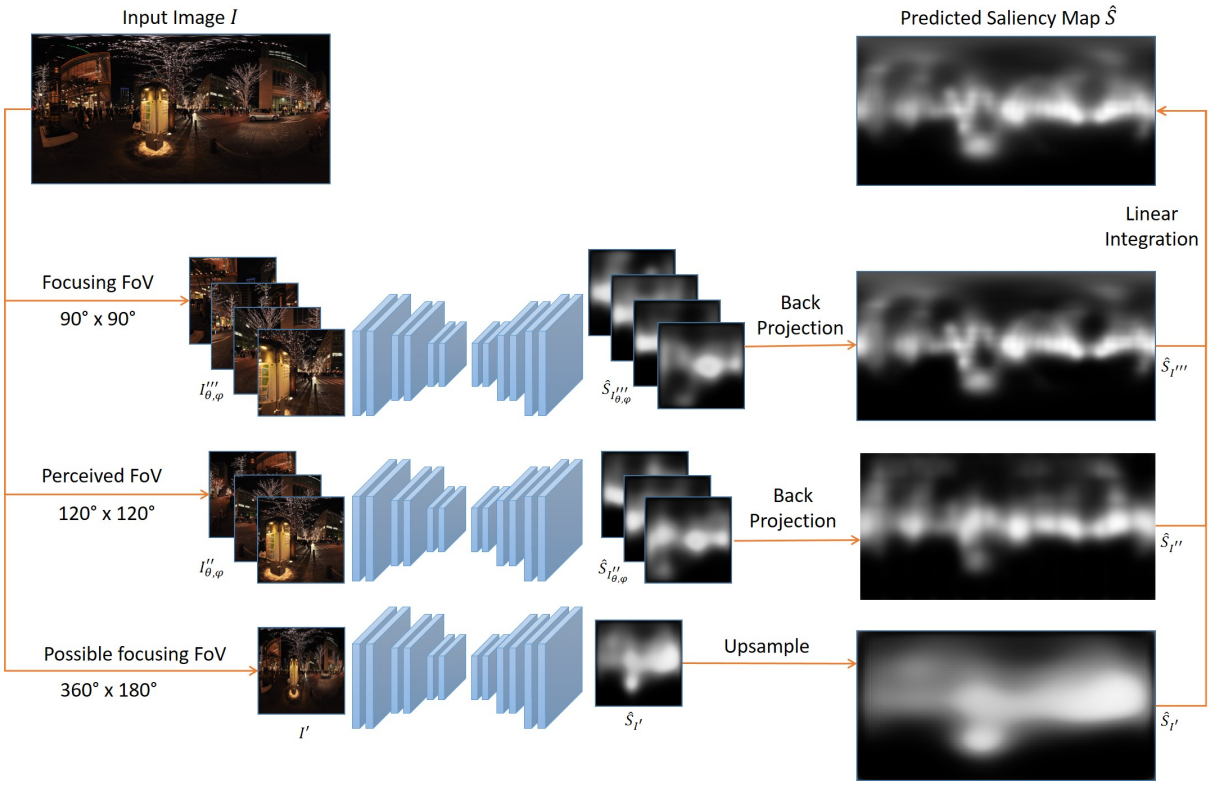


Figure 6.3: The overall diagram of our model. The architecture contains a 2D saliency model fine-tuned with our proposed adaptive weighting loss function. It predicts saliency maps in three FoVs and its diverse viewport images, respectively. The output saliency maps $\hat{S}$ are linearly integrated to combine the saliency maps in different FoVs.

Human visual saliency is highly related to the image scale. People tend to look at fine details when the image is zoomed in and look at coarse details when the image is zoomed out. When observers wear HMD, they do not see the entire 360° image in a glance but only the content inside her/his current viewport. It is similar to the condition that she/he takes a close look to a large image and rotates head to look at other parts of this image. Hence, user visual attention is guided by the salient region not only within the current viewport but also within the overall content in 360° image. According to human visual physiology and the designation of the most common HMD, *i.e.* HTC Vive [184] and Oculus Rift [185], on the market, we propose a tailor-made model taking

advantages of three different FoV in low, middle, and high resolutions as input. 360° image is projected into three different FoVs and down-sampled into the same size. Each image is processed by a 2D saliency model, and the estimated saliency map of all the FoV images are linearly integrated to yield an 360° saliency map.

To alleviate the issue of size limitation of existing 360° image datasets, the 2D saliency model used here is pretrained in a large scale 2D image saliency dataset Salicon [181] first, then adjusted in a relatively small 360° image dataset via fine tuning. Previous 2D saliency models [178, 179] used an uniform or manually weighting sum of losses to take into account several evaluation factors including KLD, CC and NSS simultaneously. It is time-consuming to tune an appropriate weights and the effectiveness of those weights have not been validated yet. To avoid the questionable pre-defined weights in the loss function, as we defined in Eq. (6.1), we propose thus an adaptive weighting loss function which updates the weights iteratively during the fine tuning.

Fig. 6.3 demonstrates the overall architecture of our model.

6.3.1 Multi-resolutional FoV basis

Fixation prediction in 360° image can be regarded as eye movement in a single viewport and head movement in an entire 360° image. Following eye movement in the current viewport, the user's head may rotate to neighboring viewport to look at different contents. Fig. 6.4[1] presents the region of human visual FoV and the FoV provided by HTC Vive and Oculus Rift. It shows that although the largest human visual FoV in horizontal and vertical range is about 180° and 120°, the FoV provided by HMD is only about 120° horizontally and vertically. According to the theory of Peripheral Vision [186], which explains the vision occurs outside the fixation point, we define three FoVs as:

1. Focusing FoV: Humans have the highest visual acuity in the region inside 60° in diameter [186]. Since eye balls can move freely in HMD, we extend the region of Focusing FoV to 90° horizontally and vertically from the center gaze point.

2. Perceived FoV: It is the FoV that observers perceive instantly in HMD before any movement of eyes and rotation of heads. Hence it ranges 120° horizontally and

1. https://www.reddit.com/r/Vive/comments/4ceskb/fov_comparison/

vertically as the designation of HMD.

3. Possible Focusing FoV: Observers are allowed to rotate their head to change viewport. Therefore, all the possible FoV they can see is the entire FoV of 360° images which have 360° and 180° in horizon and vertical.

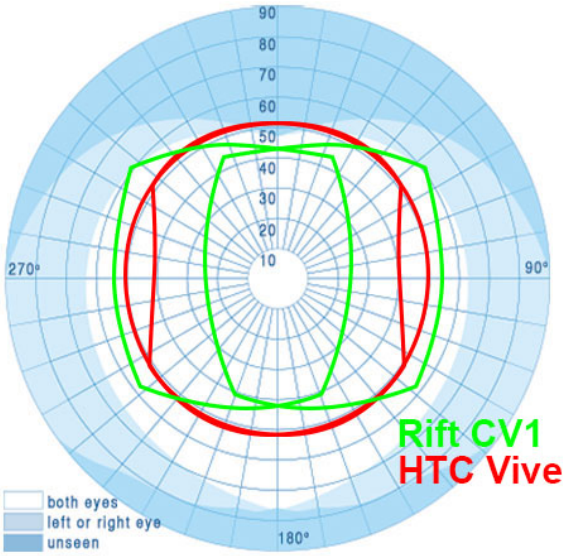


Figure 6.4: Region of human Field of View (FoV) and the FoV provided by HMDs. Red circle presents FoV of 2 eyes in HTC Vive, where two strainght lines on the left and right side are the edges of FoV of right eye and left eye, respectively. Two green rectangle represent FoV of 2 eyes in Oculus Rift. The largest human FoV reaches 120° vertically and 180° horizontally, but HMDs only provide 120° vertically and horizontally. Thus, we define three FoV to describe the viewing condition when observers wearing HMD.

To enumerate all the possible points of views that users may look at, we transform a 360° images from equirectangular format to rectilinear images with respect to diverse viewports in three FoVs, *i.e.* $90° \times 90°$, $120° \times 120°$, and $360° \times 180°$. All these viewport images are down-sampled to the same rectangular size, and served as the inputs to a 2D saliency model to capacitate the model to extract both fine and course features.

In Fig. 6.3, the input image is denoted as I. The viewport images are denoted as I' for $360° \times 180°$ FoV, $I''_{\theta,\phi}$ for $120° \times 120°$ FoV, and $I'''_{\theta,\phi}$ for $90° \times 90°$ FoV, where θ and ϕ represent the normal angles of viewport images. For I''', we project the viewport plane images in every 30° in longitude and latitude. All these viewport images are used to

fine-tune the 2D saliency model pretrained using the SALICON dataset. Then the predicted saliency maps from all the viewport images are back-projected and averaged into a equirectangular saliency map $\hat{S}_{I'''}$. For I'', it is also projected from input image in every 30° in longitude and latitude, but not used for training. The saliency maps predicted by fine-tuned 2D model are back-projected and averaged into a equirectangular saliency map $\hat{S}_{I''}$. On the other side, the image I is downsampled to the image I' which has the same size as that of the I'' and I''' viewport image. The saliency map of the downsampled 360° image is also estimated by the fine-tuned 2D model. Finally, the equirectangular saliency map predicted from three FoVs are linearly integrated to generate the final saliency map.

6.3.2 Adaptive Weighting

A plenty of evaluation metrics are available to score the predicted saliency map according to the definition of the saliency and the representation of the ground-truth map [101]. In the databases used in this paper, the ground-truth for each 360° image includes a binary fixation map recording user's gaze positions and a continuous saliency map which presents the probability distribution post-processed by convoluting each fixation location with a Gaussian filter with its standard deviation equal to human visual angle.

For the purpose of accomplishing the best performance in most evaluation factors, we combine three evaluation metrics (KLD, CC and NSS) together and propose an adaptive weighting method to balance the influence of each component. Under the hypothesis that when a metric's scores are spread out over a wider range of values, this metric should be less considered, our loss function is defined as:

$$\begin{aligned} L = \tfrac{1}{\sigma_{KLD}} L_{KLD}(P, Q^D) - \tfrac{1}{\sigma_{CC}} L_{CC}(P, Q^D) \\ - \tfrac{1}{\sigma_{NSS}} L_{NSS}(P, Q^B) \end{aligned} \tag{6.7}$$

where σ can be seen as the relative weighting of each component. Large σ decreases the impact while small σ increases the impact of its corresponding evaluation score.

In the process of fine-tuning, parameters initialized from pretrained model can be learned with back-propagation according to the loss between output and ground-truth. Using Equation (6.7) as loss function, predicted saliency map gradually approaches to the ground-truth in every iteration. We measure the standard deviations of KLD, CC

and NSS in each epoch and update them in the next epoch.

Along with more iterations are updated, the value of standard deviation σ decreases and the value of loss function increases. An adaptive learning rate is used to prevent excessive gradient decent. The decay rate of the learning rate is defined as $(1-iter/max\_iter)^{0.9}$, where $iter$ denotes the current number of iteration and max_iter denotes the estimated total number of iterations. It becomes smaller and smaller during training.

6.4 Performance evaluation of two proposed methods

Our model is implemented within the SalGAN [177] framework. It detects the saliency map with GAN including a generator to predict saliency map and a discriminator to distinguish the authenticity of predicted map. In this section, we describe the experimental settings, datasets and metrics used for evaluation, and analyze each component in our architecture. Our method is compared with several state-of-the-art methods.

6.4.1 Experimental Setup

Datasets

To ensure a comprehensive comparison, we use 4 datasets with different image contents, different acquisition equipments and saliency maps generated in different ways to evaluate our method. We list the descriptions of these datasets in the following:

- **Salient360! 2017** [187]: This dataset released 60 omnidirectional images to the public for free-use, and 25 omnidirectional images for evaluating the saliency models in ICME2017 challenge. In order to equally compare our model with others, we follow the rules of this challenge to train our model with free-use 60 images and evaluate with 25 images used in the challenge. All the images are in equirectangular format with resolutions ranged from 5376×2688 pixels to 18332×9166 pixels. There are 20 images used for head movement and 40 images used for head+eye movement. Fixation locations and head positions of each image are collected from at least 40 observers wearing HMD Oculus-DK2 and watching each image for 25 seconds. The starting position is set in the center of images at the beginning of each visualization. A small eye-tracking camera

is embedded in HMD to record fixation of dominant eye at 60 Hz. A Gaussian of 3.34° visual angle is applied to blur all the fixation points within the viewport plane, then back-projected to the final equirectangular saliency map.

- **Salient360! 2018** [188]: It was built similar to Salient360! 2017 but the authors improved some aspects of the processing of raw data and generation of saliency maps (e.g. using information for the two eyes, and some more). That is why the provided saliency maps are very different from the Salient360! 2017 dataset. There are 101 equirectangular omnidirectional images and their saliency map and fixation maps in this dataset. The ground-truth of 85 images was released to public for the training and the validation purpose, while 26 images was kept secretly for the test and the benchmark [189]. We give our method to the authors to get its performance on the test images and compare it with other state-of-the-art methods with known performance (on the benchmark website) but without paper to be referred to.
- **Stanford** [190]: 22 panoramas including indoor and outdoor scenes are used to record 122 users' eye fixation under three different viewing conditions: viewed with HMD in a standing or seating position in a non-swivel chair, and seated in front of a desktop monitor. Users are more willing to move and rotate their heads in the standing position. All the panoramas were viewed in 30 seconds began in the different starting points. Fixations were recorded with a pupil-labs1 stereoscopic eye tracker installed in Oculus DK2 HMD at 120 Hz. Fixation maps were convolved by a Gaussian with standard deviation of 1° visual angle to yield continuous saliency maps. Panoramas viewed with HMD at the same start point standing and seating are used in our comparison.

Evaluation Metrics

We execute 4 evaluation metrics which are KLD, CC, NSS and AUC-judd [101] are the same as the demonstration in Section 4.3.

Note that it is incorrect to compare two saliency maps in equirectangular format since it oversamples the points close to the north pole and south pole. Therefore, we abide by the comparison method used in the Challenge Salient360! 2017 [187] and Salient360! 2018 [188] which only compares predicted saliency map and ground-truth map with the sampled points uniformly distributing on a sphere.

Training and Testing

SALICON dataset is used to pretrain our method in SalGAN framework. The hyper-parameters follow the suggestions from [177]. Our models are then fine-tuned on the dataset Salient360! 2017 via transfer learning. 30 images in this dataset are used for training, 10 images for validation and 25 images for evaluation. In validation and test process, rectilinear images of Focusing FoV are projected in every 10° along with longitude and latitude to enumerate all the possible viewport that observers may see. Then predicted viewport saliency maps are back-projected and averaged into an equirectangular map. To save computational cost, rectilinear images of Perceived FoV are only projected in every 30°, and back-projected and averaged into an equirectangular map as Focusing FoV. A Gaussian filter is used here to slightly blur the prediction maps. Two equirectangular saliency maps predicted from Focusing FoV and Perceived FoV are linearly integrated with the saliency map estimated from the Possible Perceived FoV into a final equirectangular map.

6.4.2 Ablation study for the MV-SalGAN360

Compared to the SalGAN360, the MV-SalGAN360 proposes two more components. Thus in this section we firstly analyze each component in the architecture of MV-SalGAN360 to show its contribution.

Multi-resolutional Field of View

Fig. 6.5 shows the comparison between models using Focusing FoV, Perceived FoV, Possible Perceived FoV, and their integrations with and without fine-tuning. We use the saliency map estimated from SalGAN directly for the models without fine-tuning. Results show that, compare to Focusing FoV with fine-tuning, the integration of Focusing FoV and Perceived FoV with fine-tuning improves the performance in KLD, NSS and AUC-Judd but slightly worsens the performance of CC. Moreover, the integration of three FoVs with fine-tuning enhances the performance in CC, NSS and AUC-Judd, but slightly worsens the performance of KLD. Taking account all the four evaluation metrics, we conclude that the integration of three FoVs outperforms the others in Fig. 6.5.

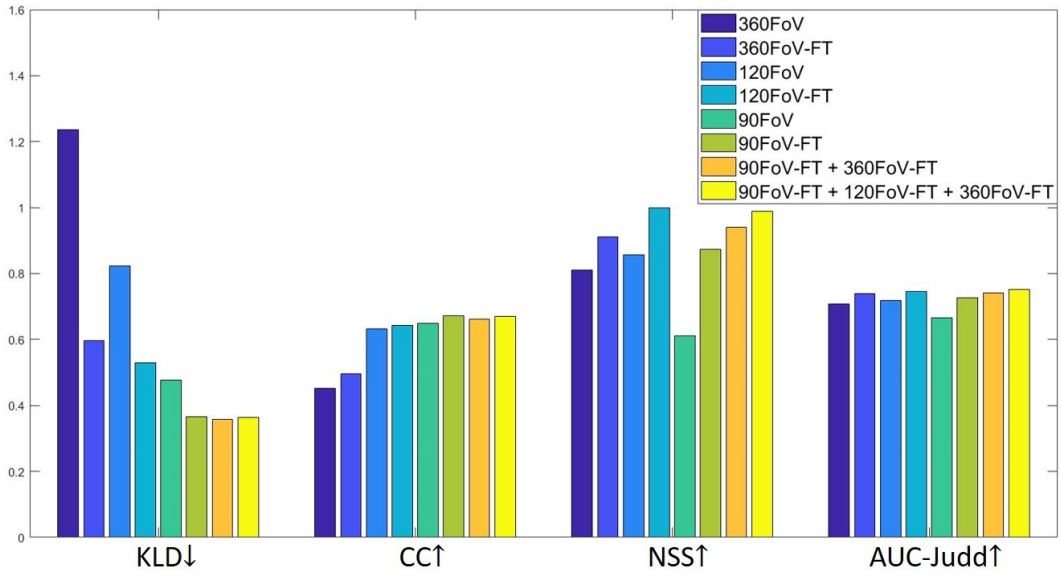


Figure 6.5: Comparison of three FoVs models with and without fine-tuning (FT). 90FoV, 120FoV, and 360FoV represents Focusing FoV, Perceived FoV, and Possible Perceived FoV, respectively. Lower KLD value indicates a better performance, and a higher score of other metrics means a better performance.

Fine-tuning with Adaptive Weighting

Table 6.2: Results of fine-tuning on viewport plane image

Method	KLD↓	CC↑	NSS↑	AUC-Judd↑
without FT	2.722	0.424	0.476	0.616
FT with fixed weighting	1.863	**0.538**	**0.563**	0.633
FT with adaptive weighting	**1.410**	0.533	0.548	**0.644**

In Table 6.2, we evaluate the results of the models with/without fine-tuning with adaptive weighting loss function. The model fine-tuned with fixed weighting, which parameters stay equal in training procedure, is also demonstrated. We can see that fine-tuning, no matter with adaptive weighting or fixed weighting, enhances the performance of the 4 metrics in a large margin. Comparing the adaptive weighting with the fixed weighting, the former has a evident improvement on KLD and AUC-Judd but small

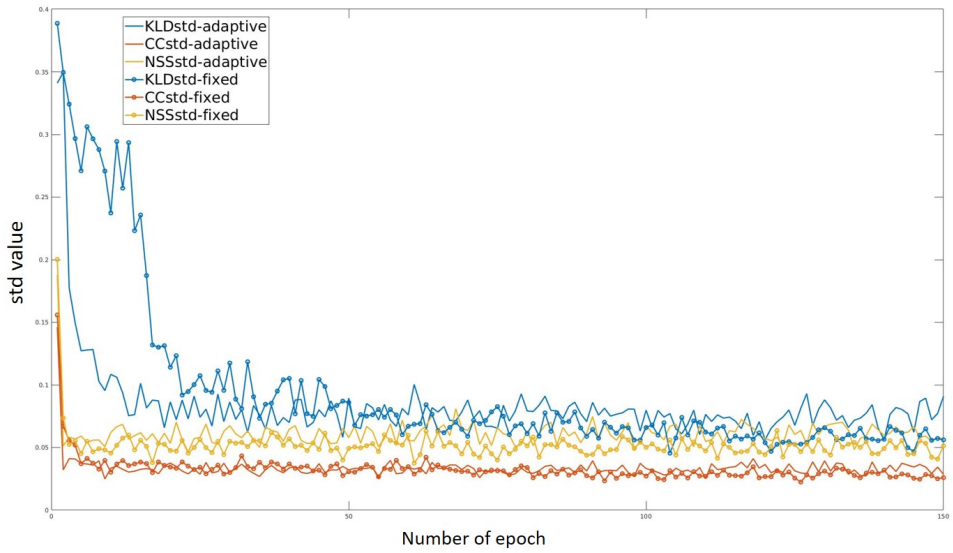


Figure 6.6: Training plots showing convergence of the σ value of three evaluation metrics considered in our loss function with adaptive weighting and fixed weighting. The σ of KLD converges more rapidly in the adaptive weighing case while the convergence of the other two metrics has no evident difference.

decrease on CC and NSS. Fig. 6.6 illustrates the standard deviation values of 3 evaluation metrics which are considered in our loss function. We can see that the standard deviation of KLD in adaptive weighting model decreases more drastically than the one in fixed weighting model and achieve lower value in the end of training. However, the standard deviation tendencies of CC and NSS in adaptive weighting model and fixed weighting model are quite similar. This might be the reason why the adaptive model achieves a better result in a large margin on KLD but no better on CC and NSS. The saliency maps predicted from the model without fine-tuning are more concentrated, while the maps predicted from adaptive weighting model are more wide-spread in the middle area than that from the fixed weighting model. This result is consistent with the findings [170] that observers look at more on equator area in 360° images.

6.4.3 Comparison with state-of-the-art

We compare our two proposed models with the state-of-the-art methods on three datasets (Salient360! 2017, Salient360! 2018 and Stanford).

Table 6.3 compares our models with 11 saliency prediction models in Salient360!

2017 dataset. SalGAN is compared here to present the performance of 2D model used in 360° images without any modification. The other 7 models listed in Table 6.3, which are Maugey *et al.* [191], SalNet360[192], GBVS360 [193], BMS360 [193], Startsev *et al.* [194], Ling *et al.* [195] and Zhu *et al.* [196], are the participants of the Grand Challenge Salient360! ICME2017. Their performance are validated by the organizers of the challenge. Our models outperform the others on all the 4 evaluation scores, especially on KLD and NSS. The MV-SalGAN360 has even better performances than the SalGAN360.

Table 6.3: Comparison results on dataset [187] (the best and the second-best scores are highlighted in bold style and blue color)

Method	KLD↓	CC↑	NSS↑	AUC-Judd↑
Maugey *et al.* [191]	0.585	0.448	0.506	0.644
Xu *et al.*[197]	-	0.409	0.699	0.659
SalNet360[192]	0.458	0.548	0.755	0.701
SalGAN[177]	1.236	0.452	0.810	0.708
Startsev *et al.* [194]	0.42	0.62	0.81	0.72
GBVS360 [193]	0.698	0.527	0.851	0.714
BMS360 [193]	0.599	0.554	0.936	0.736
SalGAN&FSM [169]	0.896	0.512	0.910	0.723
Zhu *et al.* [196]	0.481	0.532	0.918	0.734
Ling *et al.* [195]	0.477	0.550	0.939	0.736
SalGAN360	0.431	0.659	0.971	0.746
MV-SalGAN360	**0.363**	**0.671**	**0.988**	**0.751**

In the Salient360!2018 dataset, Table 6.4 compares our models with the other three models participated ICME2018 Grand Challenge. We submitted our model to the benchmark built by the challenge organizers. Thus, all the performance is validated by them with their private test dataset. MV-SalGAN360 achieves the best results on all the 5 indexes and has remarkably higher scores on NSS and AUC-Judd.

Table 6.5 presents the performances of our model and other 5 state-of-the-arts in Stanford dataset. All the models listed here used the same training dataset and parameters as those listed in Table 6.3. Our models have remarkably higher scores than other tested models.

Fig. 6.7 and Fig. 6.8 illustrate the qualitative results obtained by the MF-SalNet360 and other state of the art models (SalGAN&FSM, Startsev et al. and SalNet360) on Salient360! 2017 evaluation datasets and Stanford dataset. We can see that our model

Table 6.4: Results of Benchmark [189] (the best and the second-best scores are highlighted in bold style and blue color)

Method	KLD↓	CC↑	SIM↑	NSS↑	AUC-Judd↑
SJTU model	1.238	0.520	0.573	1.397	0.820
Wuhan University	0.899	0.607	0.612	1.617	0.822
SalGAN360	0.739	0.642	0.635	1.585	0.820
MV-SalGAN360	**0.704**	**0.643**	**0.637**	**1.625**	**0.830**

Table 6.5: Comparison results on dataset [190] (the best and the second-best scores are highlighted in bold style and blue color)

Method	KLD↓	CC↑	NSS↑	AUC-Judd↑
Startsev *et al.* [194]	5.666	0.431	1.148	0.754
SalNet360 [192]	5.849	0.390	1.200	0.772
SalGAN [177]	5.280	0.361	1.236	0.783
SalGAN&FSM [169]	5.333	0.375	1.286	0.794
SalGAN360	4.659	**0.488**	1.530	0.829
MV-SalGAN360	**4.642**	**0.488**	**1.548**	**0.834**

is capable of detecting high saliency regions on people, animals, and objects. The predicted saliency maps from our model are also more concentrated on the salient region. For the images which have no strong saliency contents, our model successfully predicts equator area where humans tend to put more attention to.

6.5 Conclusions and Perspectives

We have firstly proposed the SalGAN360, a new model predicting the saliency map for 360° images. Then considering that 360° images are usually in high resolution and observers only see a part of content in current viewport in HMD, we thought about utilizing multi-resolutional FoV to improve the performance of SalGAN360, via the integration of salient features extracted from diverse viewport plane image in small ($90° \times 90°$), middle ($120° \times 120°$), and large ($360° \times 180°$) Field of View (FoV). We show that the two models have better performance than the tested state-of-the-art models on several datasets.

One perspective is to extend our model to 360° videos, thus we also need to explore the usage of deep networks in the temporal domain and how to effectively fuse them

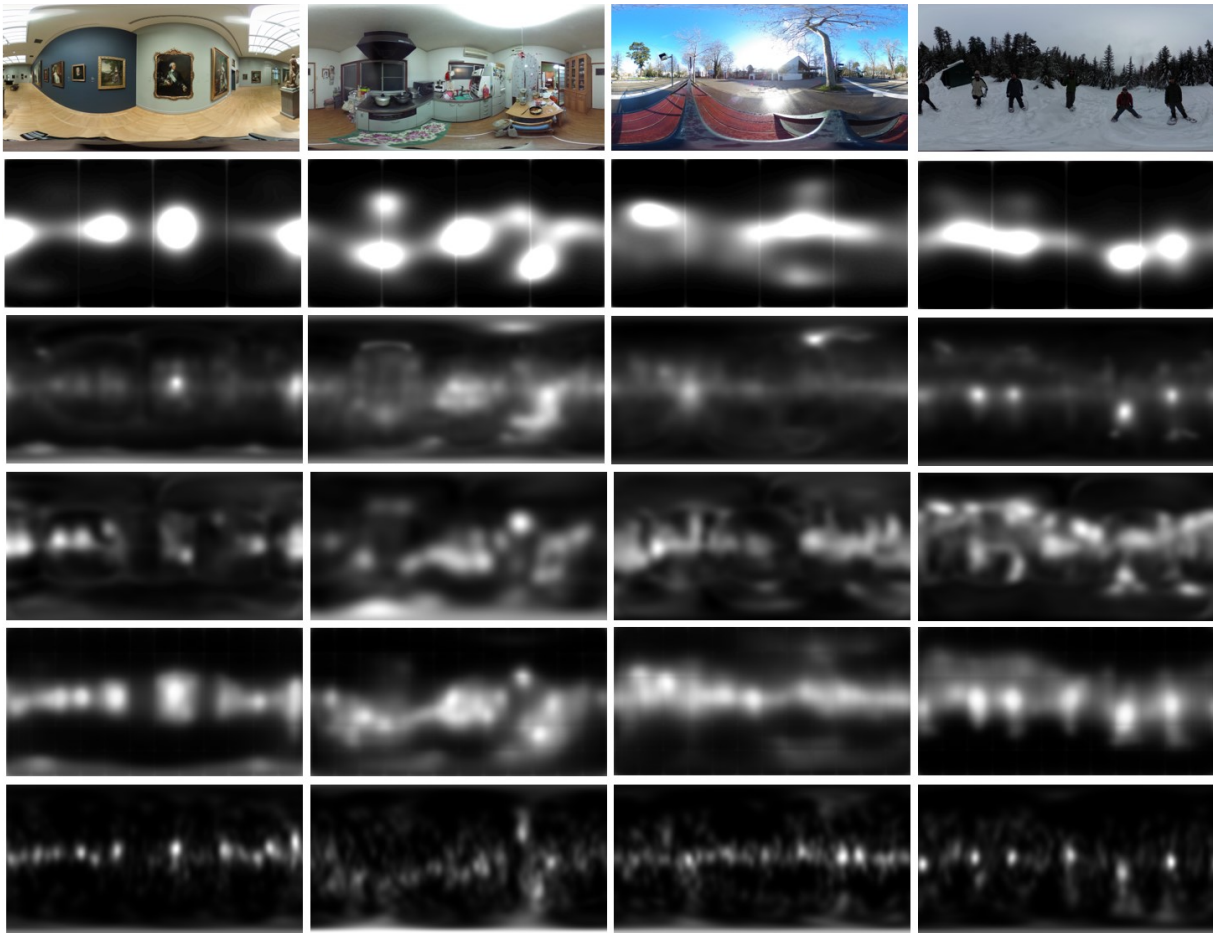


Figure 6.7: Qualitative results and comparison with other state of the art models on Salinet360! 2017 [187] test set (From top to down, each line corresponds to the original image, the saliency maps from SalGAN&FSM [169], Startsev et al. [194], SalNet360 [192]and MF-SalNet360, and the Ground Truth).

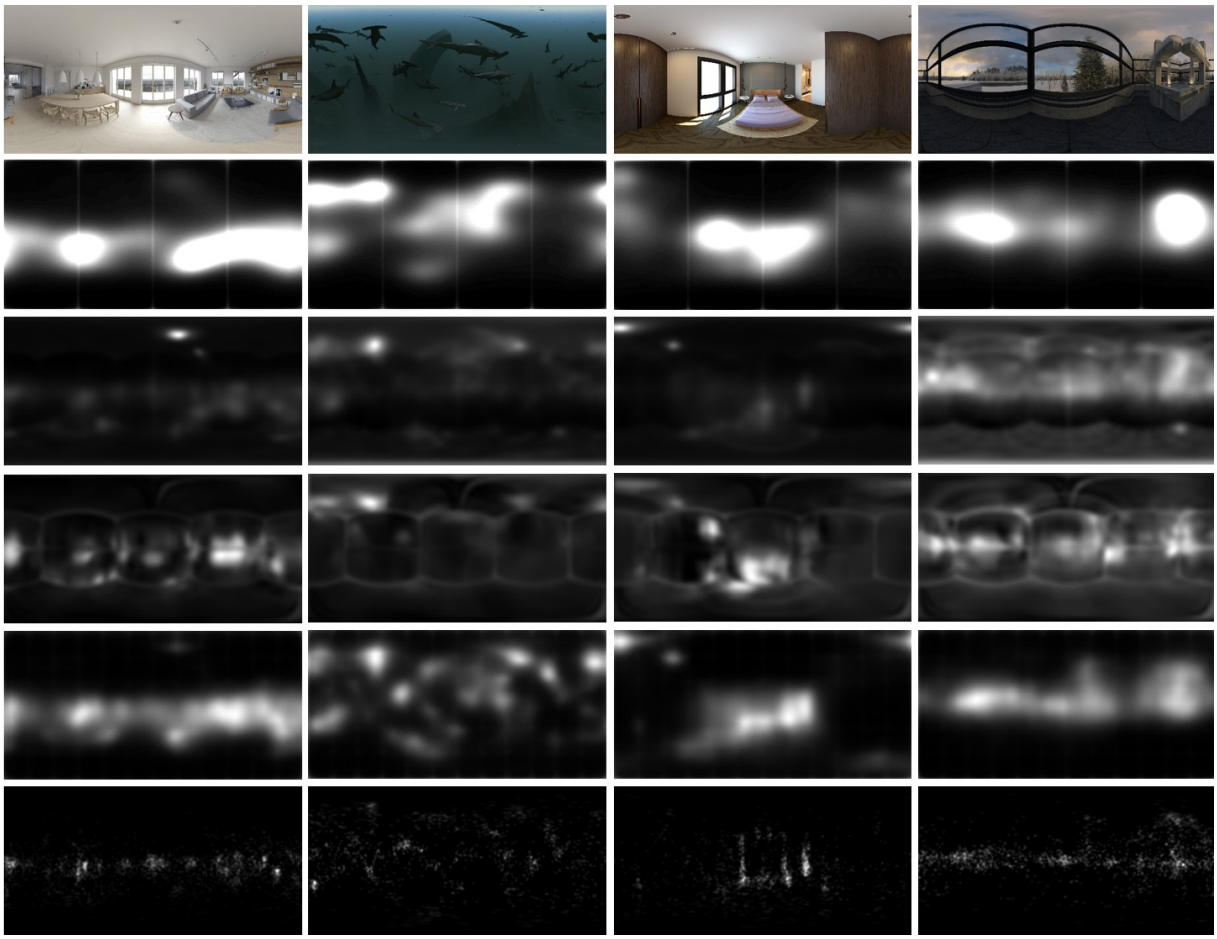


Figure 6.8: Qualitative results and comparison with other state of the art models on Stanford dataset [190]. The view angle of Gaussian blur is set to 1° in this dataset, so that the saliency regions in the ground-truth are much smaller than that in Salinet360! 2017 [187] dataset (view angle is 3.34°). From top to down, each line corresponds to the original image, the saliency maps from SalGAN&FSM [169], Startsev et al. [194], SalNet360 [192]and MF-SalNet360, and the Ground Truth.

together. In reality, observers' visual attention could be attracted or changed by the sound when they wear HMDs. Thus it will also be interesting to study the saliency model when visual-audio signal is provided.

6.6 Contributions in this field

The first model we proposed, i.e. the SalGAN360, got the 1st place in ICME Grand Challenge « Prediction of Head+Eye Saliency for 360 Images » in 2018. Two publications related to this topic are :

[1] F. Chao, **L. Zhang**, W. Hammidouche, O. Déforges. "SalGAN360: Visual Saliency Prediction on 360 Degree Images with Generative Adversarial Networks". ICME2018, July 2018, San Diego, California, USA.

[2] F. Chao, **L. Zhang**, W. Hammidouche, O. Déforges. "A Multi-Field of View Viewport-based Visual Fixation Model Using Adaptive Weighting Losses for 360° Images". IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING (J-STSP), special issue "Perception-Driven 360-Degree Video Processing", submitted.

This relatively new topic is actually an on-going work, and I believe that we will contribute more to this field in the short term.

PART III

Perspectives of research

CHAPTER 7

RESEARCH PROJECT

I will continue to work on “Image Quality Assessment and Saliency Detection: Human Perception Modelling and Applications” for the next five years.

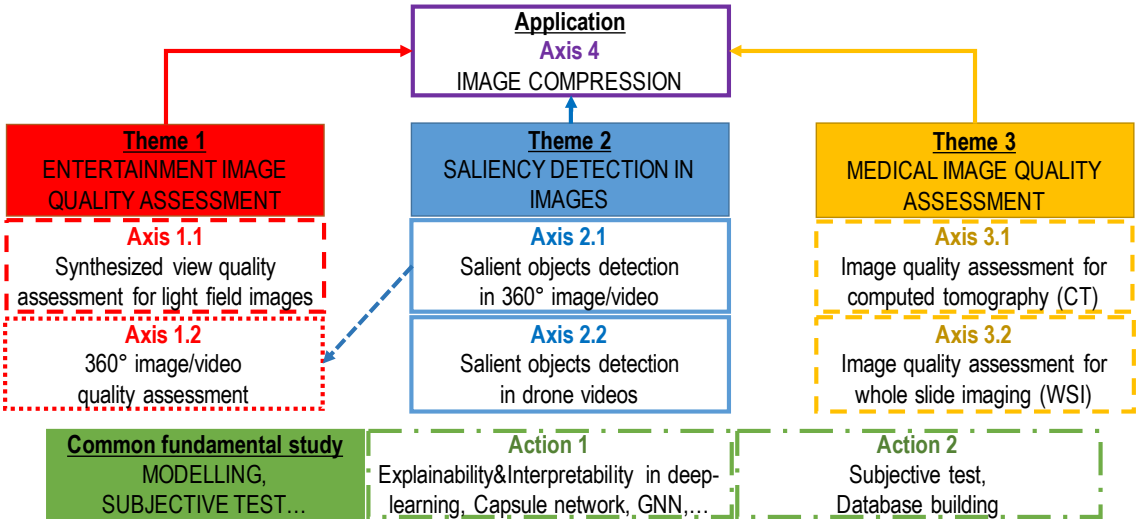


Figure 7.1: Illustration of my research project (the axes with solid borders denote that I have or will have a PhD funding for the topics, the axes with dashed borders denote that I have a collaboration for the topics, the axes with dotted borders denote that no funding/collaboration yet; the actions will be explored in each axis and may be used by other axes; the results from Axis 2.2 can be used in Axis 2.1).

The research project described here focuses naturally on my three main research interests, as illustrated in Figure 7.1:

- Theme 1: entertainment image quality assessment
- Theme 2: saliency detection in images
- Theme 3: medical image quality assessment

In each theme, different image types (compared to previous works) and new challenges will be tackled.

The fruits from the three themes can further be applied on the image compression, which is the most renowned research activity (for now) of the VAADER team, for improvements. Note that nowadays the image compression community has realized that the simple metric Peak Signal to Noise Ratio (PSNR) does not correlate with the perceived image quality of end-users. I have reasons to believe that the trend will be the development of compression algorithms evaluated and optimized by perceptual quality metrics (like what Netfilx is doing). This is indeed the Axis 4 in my research project.

Except the fundamental study specific to each axis, there are some common points that could be useful as tools for all the axes, e.g. the new deep networks studies and the subjective test protocols. For example: 1) In the last decade, the deep-learning has exploded with interesting and promising results. With major achievements in image recognition, speech recognition and highly complex games, deep-learning continues to disrupt society. As researchers, I think that we should be careful to use the black-box deep-learning algorithms. It would be better if we can provide explanations of their decisions in some level of detail, before applying or adapting them for our own applications. Recently proposed deep networks (e.g. Capsule network, Graph neural network, etc.) will also be studied and considered whether/how to be used in our own applications. 2) On the one hand, the subjective test using human observers is the ultimate test for the validation of our algorithms, since the three themes are all related to human perception. Different test protocols should be studied and chosen for different applications. On the other hand, for certain applications, maybe there is no database in the literature or the number of images in the existing databases is not enough. In that case, we need to build our own databases. Since we have already set up a standardized psycho-visual room in our team and we have bought several acquisition and display devices, we will have the necessary materials for the subjective tests.

7.1 Theme 1 Entertainment image quality assessment

With coming era of immersive media (estimated 2020-2023), the image quality assessment metrics (IQMs) appear as crucial for improving the quality of the users' immersive experience and reducing the side effects during the users' observation, whereas they have been little exploited until now. A good objective IQM correlated with human MOS

can directly guide the development of compression, representation and other processing methods for immersive medias. The FR (or RR) metrics normally are easier and have better performances (because of richer information), but their major limitation is that they always need a reference which is unavailable in many circumstances. Thus, I feel it's urgent to work on objective IQMs (especially NR metrics) for two new imaging modalities: light field images (LFI) and 360° images, where come from Axis 1.1 and 1.2.

7.1.1 Axis 1.1 Synthesized view quality assessment for light field images

State of the art

Light field imaging has emerged as a technology allowing to capture richer visual information from our world. As opposed to traditional photography, which captures a 2D projection of the light in the scene integrating the angular domain, light fields collect radiance from rays in all directions, demultiplexing the angular information lost in conventional photography. On the one hand, this higher dimensional representation of visual data offers powerful capabilities for scene understanding, and substantially improves the performance of traditional computer vision problems such as depth sensing, post-capture refocusing, segmentation, video stabilization, material classification, etc. On the other hand, the high-dimensionality of light fields also brings up new challenges in terms of data processing chain (capture, compression, content editing, and display). On the one hand, the quality of LFIs can be distorted at each stage of the processing chain. On the other hand, super multi-view light-field displays require input views acquired at high angular resolution, covering a large field of view. Acquiring content with a large number of real cameras is often economically and technologically prohibitive, so interpolating intermediate views from those captured using sparse camera arrays has been suggested to achieve the required high view density. However, such synthesized views also suffer from several artifacts, which can severely affect the perceived quality.

Tamboli et al. [198] investigated limited distortions (Gaussian blur, additive noise and JPEG) and proposed a IQM combining spatial information from each constituent image and angular information (depth cues) from consecutive images. Later they extended this IQM to videos by integrating optical flow values. Another work of these authors [2] focused on the quality of synthesized views rendered by VSRS, a depth-

image-based rendering (DIBR) algorithm, and evaluated 2D and synthesized views dedicated IQMs in this context. They mentioned the importance of the accuracy of the depth map used for view synthesis. As a 1st step in this domain, the big limitation is that they only used 3 scenes with very simple background (uniform wall) in [198] and only 2 scenes in [199]. Kiran Adhikarla et al. [200] tried to include more scenes (9 synthetic and 5 real-world scenes) and more types of distortion (6 quantization steps for 3D-HEVC encoder, 3 view synthesis methods and modeling the crosstalk between adjacent views) in their database and evaluated several existing 2D IQMs. Using the database [200], Fang et al. [201] then proposed a IQM by extracting local features from LFIs and global features from epipolar plane images (EPIs) which contains both spatial and angular information of LFIs, and showed that it has a better performance than existing 2D IQMs. However, the view synthesis methods in [200] were too simple (without any state-of-the-art method). Viola and Touradj [202] proposed a VALID database focusing on the compression distortions, using 5 real-world scenes compressed by HEVC and VP9 encoders at various bitrates. A more recent work [203] also focused on the compression and noise distortions, and proposed an IQM based on depth map feature and tested on 7 synthetic scenes with ground-truth depth maps.

Scientific challenges and Objectives

To the best of my knowledge, there are only 3 objective IQM [198, 201, 203] dedicated to LFIs in the literature at the moment. All of them focused on the distortions which also exist in the traditional 2D images (e.g. compression, noisy). However, the LFI specific distortions (e.g. the synthesized view quality and the depth map quality) have not been considered in these studies. The related databases didn't include the newest view synthesis or depth estimation algorithms. That's why two international standardization groups, JPEG Pleno and MPEG-I Visual (working on the standard framework for the LFI representation and compression, respectively) have mentioned the lack but the importance of quality metrics for LFIs at the 2019 CLIM workshop. Jung from MPEG-I Visual also mentioned that the problem of LFI is not a compression only problem, since a light-field is needed to be rendered as dense as possible (without occlusions/holes under sparse capture and transmission/compression constraints), which involves view synthesis, which involves (often) depth estimation (http://clim.inria.fr/workshop.htm).

IQMs in [198, 201] are FR and IQM in [203] are RR, which need the original views with the same position as the target synthesized view or the features from these original

views, which are not available in real applications. These metrics are useful for the benchmarking, but a NR IQM is in great need for many other applications.

For the subjective test, Light Field Display (LFD) is only used in [198, 200]. However, since this technology is still far to be a consumer level product, in the literature 2D displays are commonly used, where participants were shown a LFI of a certain viewpoint or the LF contents as pre-recorded animations navigating between the perspective views in a serpentine order to mimic the parallax effect. Studies comparing the synthesized view quality on 2D display and LFD will also be interesting for this topic.

The objectives of this axis are: 1) Conduct subjective tests where human observers will assess different combinations of state-of-the-art depth estimation and view synthesis algorithms (including their ground-truth as the reference). Depth map is an important feature for LFIs [6], the quality of depth maps influences a lot the quality of the synthesized view. Thus, the 1st thing to be explored is the relation between the depth maps quality and the synthesized view quality. 2) Propose NR synthesized view quality metrics by exploring how to use the original views adjacent to the target synthesized view (available in real applications) and their depth maps (can be estimated in real applications). As a first step, ground-truth depth maps could be used, thus only artifacts introduced by view synthesis algorithms will be evaluated; then a good depth estimation algorithm would be selected and used in our metric for real-world scenes.

Fundamental studies

Fundamental study 1: Synthesized view quality assessment metrics - In the context of S. Tian's PhD thesis, we have proposed 2 FR and 2 NR metrics (4 published papers) for the evaluation of depth-image-based rendering (DIBR) algorithms for free viewpoint television (FTV) applications (where the baselines are large). Note that the view synthesis is easier for dense light fields with narrow baselines, but more difficult for sparse light fields with large baselines, where the quality evaluation may make a bigger difference. Thus, the 4 metrics could be tested on sparse LFIs as a first attempt. Our FR metrics have already showed good performance; but the NR metrics performed well for known and tested DIBR algorithms, but not very well for new view synthesis algorithms (since they are distortion-specific, i.e. designed for known distortions). The information we used in the 2 NR metrics was also very limited (only the target synthesized view is used). Considering that there are more original views adjacent to the target synthesized view in LF than in FTV and LF has more powerful capability of depth estimation,

we may use more information (e.g. the original views adjacent to the target synthesized view, as well as their depth maps) to improve our NR metrics or propose novel metrics for LFIs.

Fundamental study 2: Database with depth maps and depth estimation - Recently, I collaborated with ULB and proposed together a deep end-to-end network for light field depth estimation, which is about 3 times smaller and 3 times faster than the current top-performing depth estimation method Epinet, which may be a candidate for the 2nd objective. Note that the ground-truth depth maps are only available for synthetic scenes. Both of the two existing synthetic databases HCI and CVIA-HCI imitate images captured by micro-lens array based light field cameras. While light field camera arrays can produce images with a much higher spatial resolution, no synthetic database simulating this type of images has been proposed before. To fill this gap, we constructed such a database including 30 synthetic scenes using Blender. A paper on this work has been submitted.

7.1.2 Axis 1.2 360° image/video quality assessment

State of the art

With the explosion of Virtual Reality technologies, the production and usage of omnidirectional images (a.k.a 360° images) is presenting new challenges in the domains of compression, transmission and rendering. The evaluation of the quality of images generated by these technologies is therefore paramount. Several subjective evaluations have been done in the literature for studying the influences of different compression levels, geometric projection methods, head mounted displays (HMD), stalling patterns during 360° video streaming. Existing objective IQMs can be divided into two categories: extension of traditional 2D IQMs and deep-learning based metrics. In 2015, Yu et al. proposed a sphere based PSNR (S-PSNR), which computes PSNR for the set of points uniformly distributed on a spherical surface instead of on the rectangular domain. In 2016, Sun et al. proposed the Weighted Spherical PSNR (WS-PSNR), of which the weight is determined by how much the sampled area is stretched in the representation; Zakharchenko et al. proposed the Craster Parabolic Projection PSNR (CPP-PSNR). In 2018, Chen et al. proposed the spherical structural similarity index (S-SSIM) for omnidirectional video quality evaluation, which calculates the luminance, contrast and structural similarities of each pixel in the spherical domain. In 2019, the

researchers in Facebook proposed SSIM360 and 360VQM to verify the performance of 360° video pipeline on encoding and streaming; Kim et. al split the omnidirectional image into a set of patches and estimated the local quality and weight of each patch through an adversarial network which are summed to get the final quality score; Kim et al. also proposed a deep generative model to predict the VR motion sickness score.

Scientific challenges and Objectives

Compared to LFI, there are more works on objective IQMs of 360° images, but still limited. The existing IQMs mainly consider compression artifacts and geometric distortions occurring in the projection. Little work considered the human attention which is especially important for 360° images, where human observers cannot look at the entire image at a glance using the HMD. In addition, all the existing methods are FR metrics. No NR metric has been proposed yet, while it may be a nice opportunity for deep-learning based approaches.

The objectives of this axis are: 1) Deep-learning technologies has promoted the development of IQM for 360° images and could actually be used to predict the subjective quality score without using the reference image (thus a NR metric). There are already several databases providing subjective quality scores in the literature (e.g. [204]) and I know that several labs are also creating databases with more images and more distortion types, so there will be no problem of training data. I want to explore the usage of different deep networks in this domain for proposing a NR IQM. 2) Human attention is highly related to the perceived quality of 360° images and has been used in [205] where the NCP-PSNR model weighted the distortion of pixels according to their locations in panoramic video and the CP-PSNR model assigns weights to pixel-wise distortion based on the viewing direction predicted with respect to the content of omnidirectional video. However, the existing work used the gaze fixations as the human attention guide, while humans naturally and intrinsically segment the images into objects in their human visual system. I think that it is more reasonable and effective to use salient objects as the human attention guide in an IQM. For example, it is found that the average location of gaze fixation from the center of the view-port varies between 14 and 20 visual degrees in [206], but it will be interesting to explore if the varied gaze fixations belong to the same object. However, no study focusing on salient objects for 360° images has been done, that's why I will begin this axis after the axis 2.1 of which the results will serve as a fundamental study for this axis.

7.2 Theme 2 Saliency detection in images

Visual saliency is the distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention. Eye-tracking is a well-known technique for analyzing the visual fixations, based on which the saliency map can be generated. The output of the saliency detection is a saliency map, a grayscale image that shows the probability distribution of each pixel being under attention. The heat map is a simple colored representation of the continuous saliency map. Automatic saliency prediction model for 2D images is of great research interest since a long time because of its wide range of applications. In 2010, there was a turning point in this domain for 2D images: the 1st salient object detection model has been proposed and attracted a lot of attention since then. The salient object detection aims at finding and segmenting the most conspicuous objects in an image that highly catches the user's attention, for which the ground truth is often a binary map where the white region corresponds to the salient objects (the ground truth can also be a grayscale image, but the gray level of one object is the same and brighter means the object is more salient). The main reason of this turning point is that the salient objects can be directly used in the applications. For example, saliency information can be used to guide and enhance the compression, where the salient parts will be given more bits (a higher quality consequently) and non-salient parts will be given less bits. Of course, it is more efficient to directly use salient objects, instead of the sparse saliency points (the pixel-wise map needs to be converted into block-wise).

For the same reason, I believe that there will be the same turning point (from saliency prediction to salient objects detection) for new types of images, e.g. 360° image, drone or unmanned aerial vehicle (UAV) images, etc. That's why I propose the axis 2.1 and 2.2 in theme 2. I've already got the funding for the two axes, to have one PhD student on each axis beginning from October 2019 (cf. section VI): one is funded by the China Scholarship Council (CSC); the other is funded by the Directorate General of Armaments of French Government Defense (DGA, 50%) and Bretagne Region (ARED, 50%).

7.2.1 Axis 2.1 Salient objects detection in 360° image/video

State of the art

There are some (not many) works exploring the object detection for 360° images since 2017, but they detect all the objects without differentiating salient from non-salient. One example is a recent work [207], in which a dataset consisting of 903 frames and 7199 annotated objects was proposed, two deep-learning based algorithms originally proposed for 2D images (R-CNN and YOLO), as well as their multi-projection variant of YOLO were compared on this dataset.

Scientific challenges and Objectives

There are several scientific challenges: 1) No annotated data with ground-truth salient objects on 360° images in the literature. 2) No salient objects detection model exists in the literature as a reference. 3) Compared to objects in 2D images, objects in 360° images often have severe geometric distortions (may be different for different projection methods).

The objectives of this axis are: 1) Construct a database of 360° images with annotated salient objects, not based on human manually selection, but based on gaze data got from eye trackers. As argumented in [208], the manual annotations without an eye-fixation guided methodology do not reveal real human attention behavior. The authors in [208] are the first to establish a visual-attention-consistent densely annotated database for salient objects detection in 2D videos. Inspired by their approaches, I also want to construct such a database for 360° images. But even as a first database in this domain, I will use the eye-fixation as the guide. There is no lack of 360° content, and there are more and more databases providing eye fixation data. In the case that the eye fixation data is not enough from the literature, we have bought an HTC Vive Pro Eye HMD with integrated eye-tracker in our psycho-visual room, thus we could also conduct subjective tests in our lab to propose a large-scale database. How to reasonably divide this database into training, validation and test set will also be considered. 2) Our large-scale database with annotated salient objects will pave the way for exploring the application of deep networks in this domain. We may begin with the interpretability problem by comparing and analyzing different deep networks' performances on our database (from successful CNN-based methods for 2D images, our successful usage of GAN to new architectures like capsule network), determining which features

in a particular input vector contribute most strongly to a neural network's output and the success/failure of the model. Based on these studies, we will propose our own deep-learning based model for salient objects detection in 360° images/videos. 3) A specificity for 360° videos is the spatial audio in 360-degree, which allows the viewers experience a video's sound in all directions, just like real life. The spatial audio signal is considered as a powerful way of directing viewers' attention, e.g. viewers tend to look at the person who is speaking when there are several persons around. An interesting way is thus to investigate the influence of spatial sound in videos on eye movement and to propose an audio-visual model to predict salient objects in videos more accurately. We've already started to explore the sound effect in the context of the PhD thesis of F. Chao, and a subjective test is being conducted for providing saliency ground-truth with spatial audio direction (the first study on 360° images in the literature). In the context of this axis, we will continue to further explore this, for salient objects detection.

Fundamental studies

Fundamental study 1: Saliency prediction for 360° images - As a first step to understand the human attention mechanism when observers watching the 360° images, I firstly proposed to work on the saliency prediction (the PhD thesis of F. Chao), as what the researchers did on 2D images. In the context of this PhD work, we have already summarized a list of existing databases regarding 360° images/videos; and proposed two algorithms based on Generative Adversarial Networks (GAN) for predicting saliency for 360° images: one gained the 1st place in the ICME 2018 Grand Challenge on Head+Eye saliency prediction (published); the other has a better performance than the current top-performance model on the benchmark of "Prediction of Head+Eye Saliency for 360° Image" (https://salient360.ls2n.fr/un-salient360-benchmark/results-images/). Note that the majority of the existing works only consider head movements as proxy for gaze data (since it is not always easily accessible), despite the importance of eye movements in the exploration of omnidirectional content. We chose to work on Head+Eye saliency prediction which reflect more human attention.

Fundamental study 2: Salient objects detection for 2D videos - In the context of the PhD work of Q. Wang, we have also worked on traditional methods, as well as deep-learning based methods for salient objects detection for 2D videos (1 published paper). Though the priors commonly used for 2D videos (e.g. central-bias prior) are not appropriate for 360° videos, our experiences on how to use the temporal information may

give us a clue for 360° videos model design. We also compared and analyzed different architectures of deep networks and their influences on the salient objects detection performances on 2D videos (1 submitted paper).

7.2.2 Axis 2.2 Salient objects detection in drone videos

State of the art

Despite the fact that the first attempts of drone development were connected to military purposes, nowadays drones (one type of unmanned aerial vehicles (UAVs)) are used in several applications. More specifically, drones can be used in a huge variety of domains such as pilot training, disaster management, environmental protection, delivering services, etc. Among the existing applications, these connected with surveillance tasks can be considered as the most perspective ones. Considering that the majority of surveillance systems' abilities aim to simulate human visual behavior, understanding how UAV videos are perceived by human vision and proposing an automatic salient objects detection model could deliver critical information towards such systems' improvement. The ANR ASTRID project that I lead ("Saliency Detection from Operators' Point of View and Intelligent Compression of Drone videos") was proposed indeed for surveillance task. This axis is closely related to this project, but focuses on salient objects which can be more directly used. In the literature, the 1st work goes back to 2010 [209] which uses an image contrast map derived from the combination of seminal work in this area, multi-scale mean-shift segmentation with additional histogram enhancement and additional multi-channel edge information. But the data was too small to more realistically validate this algorithm. Nowadays, there are more databases with the development of deep-learning based methods, but the deep networks were used for saliency prediction or objects localization and tracking for drone or UAV videos, not yet for salient objects detection.

Scientific challenges and Objectives

In ground-level platforms, many saliency models have been developed to perceive the visual world as the human does. However, they may not fit a drone that can look from many abnormal viewpoints. Indeed, this new-born image type is distinct from traditional 2D image type in many aspects, including the bird-point-of-view which modifies the se-

mantic and size of objects, the loss of pictorial depth cues, i.e. the lack of horizontal line, and the presence of camera movements. How to propose a salient objects detection model considering these specificities is the biggest challenge for this axis. In addition, the environmental background is an important cue for salient objects detection on drone videos, how to use it as a priori has not been explored.

The objectives of this axis are: 1) Construct a database of UAV videos with annotated salient objects, not based on human manually selection, but based on our two saliency databases with gaze data got from eye trackers. 2) From the fundamental study 2, we found that deep learning models trained on traditional 2D contents show the most promising results. This outcome is highly encouraging, so we can try firstly state-of-the-art deep models for salient objects detection in 2D videos through fine-tuning or training on our own UAV database planned for the Objective 1). 3) From the fundamental study 2, we also found that the results are quite content-dependent. I consider to classify the environmental background into different types of scenes as a pre-processing, then used it as a priori to guide or refine the salient objects detection.

Fundamental studies

Fundamental study 1: Saliency databases of UAV videos - In the context of the ANR ASTRID project, we have constructed 2 databases with ground-truth saliency, as well as the raw data (fixation positions and durations), using an eye tracker. One is without task (already published on ftp://ftp.ivc.polytech.univ-nantes.fr/EyeTrackUAV) and the other is with task (detection of new appeared objects in the scene). The database with task is much more similar to the surveillance task.

Fundamental study 2: Existing deep saliency prediction models comparison - In the context of the ANR ASTRID project, we did a benchmark of state-of-the-art models (originally proposed for traditional 2D videos) for saliency prediction. This benchmark studies comprehensively two challenging aspects, namely the peculiar characteristics of UAV contents and the temporal dimension of videos, and enables to identify the strengths and weaknesses of current static, dynamic, supervised and unsupervised models for drone videos. We also identified several strategies for the development of visual attention in UAV videos.

7.3 Theme 3 Medical image quality assessment

7.3.1 State of the art

Medical image quality assessment was my PhD thesis topic, and I'm always interested by it because it is important for the improvement of health technology. The most reasonable method in this domain is the task-based approach, where the image that enables medical experts to gain a better task performance or to spend less time for interpretation with the same diagnostic reliability is said to have a better quality. The objective of clinically relevant numerical observers is to approach the diagnostic task performance of medical experts for the objective quality assessment of medical images. The numerical observers have already been widely accepted in this domain. For example, when medical companies want to release a new technology in the US, they are often required by the Food and Drug Administration (FDA) to compare their new technology with others using a numerical observer. The key problem of designing a numerical observer lies in modeling the diagnostic process of radiologists, for which one commonly accepted approach is to divide the diagnostic process into three tasks: detection, localization and characterization. The detection task requires simply a confidence rating concerning the presence of an abnormality, e.g. a lesion or a nodule. The localization task consists in indicating the locations of abnormalities. The characterization task, related to assessing the different elements of the abnormality appearance, normally involves a linguistic response describing distinctive characteristics or essential features of abnormalities. Different numerical observers have been proposed for detection and localization tasks, but no one has been proposed for the characterization task. Another thing that should be noticed is that deep learning methods have constantly updated state-of-the-art performance results across different application aspects in medical imaging domain, but they have not been used for medical image quality assessment.

7.3.2 Scientific challenges and Possible solutions

Compared to entertainment image data, the medical data was much more difficult to obtain just several years ago. We could not even get the most recent medical data without collaborations with hospitals. However, things are changing rapidly in recent years: with the waves of deep-learning and grand challenges, there are more and more publicly accessible medical data and there is a website making efforts gathering them (cf.

https://grand-challenge.org/challenges/). Though most of them are for segmentation, the original data is also provided.

After getting the data, we need annotations from medical experts, which is even more difficult. Through the collaboration with Prof. Chen of Southeast University, who has a close relationship with several hospitals in Nanjing, there is no problem to find medical experts for subjective tests now. We have already conducted one test with radiologists (who compared our denoising method with other methods on ultrasound images) in the affiliated Hospital of Nanjing Medical University for the PhD thesis of M. Outtas, who begins to work permanently in our team as a contracted teacher-researcher since September 2019.

With the data boom, the available medical experts, and one more permanent person working on medical images with me, I find it is a perfect time to restart to explore in the medical image quality assessment theme. Before 2012, proposing a numerical observer that can perform the characterization task may be too complex to be solved. But today, this may be solved with the deep learning tools for which medical domain may be a good application.

7.3.3 Studied modalities

There are different imaging modalities in the medical imaging domain, which will be favorable for the diagnosis of different pathologies and present different specific physical and physiological phenomena. Thus, the studied modality should be chosen firstly, then the studied pathology can be chosen by considering both the studied task and the studied modality. In this theme, for the next 5 years, I plan firstly to focus on two modalities: computed tomography (CT, Axis 3.1) and whole slide imaging (WSI, Axis 3.2).

Use of CT has increased dramatically over the last two decades in many countries (including China), but the radiation used in CT scans can damage body cells, including DNA molecules, which can lead to radiation-induced cancer. How to reduce radiation exposure while maintaining image quality becomes thus a key point in CT imaging. I collaborated with Prof. Chen of Southeast University for the 1st time because he wanted to extend the numerical observer PCJO I originally proposed for MRI during my PhD thesis to CT. This work was done by a master student we co-supervised in 2017, which paves the way for our project creation on Low Dose CT Imaging in 2018, funded

by the National Natural Science Foundation of China (2019-2022). In this project, deep-learning methods will be explored for improving low dose CT imaging, as well as for CT image quality metric design.

WSI is a modality I identified as a future work during my PhD defense, because the automated quality control is extremely important for the adoption of digital pathology workflows for clinical use. Clinical pathology is witnessing a paradigm shift by transferring from glass tissue slides (observed by optical microscopy) to digital slides scanned with whole slide imaging (WSI) systems. In clinical pathology departments, hundreds to thousands of digital slides (each with very high resolution ? about 80000 x 60000 pixels) are scanned each day and the number of cases has been steadily increasing, creating a challenging workload for digital pathology. Routine diagnosis of pathology slides requires high quality high-throughput images, which are directly affected by the dynamic environment of the physical optics and sensor electronics of the scanner. Many WSI scanners need a manual quality inspection of digital slides to determine whether an image needs to be re-scanned. In high-throughput scanning systems, which contain hundreds of slides for processing, it is impractical to perform a manual check for each individual slide. A robust and highly reliable automated solution is thus in great need. At CVPR 2019, I saw a work proposing a digital pathology database which comprises of 17668 patch images extracted from 100 slides annotated with up to 57 hierarchical histological tissue types (HTTs) [210]. Their data is generalized to different tissue types across different organs and aims to provide training data for supervised multi-label learning of patch-level HTT in a digitized WSI. I think that we can a similar database for the characterization task modeling, by constructing a hierarchical taxonomy from medical experts' characterization reports. Then a numerical observer involving a state-of-the-art deep network trained on this characterization database may be able to mimic medical experts' characterization process, and be further used for WSI image quality assessment and control. I plan to apply for a thesis funding on IQM for WSI in 2020/2021.

7.4 Application - Image compression

A common application of image quality assessment metrics (IQMs) and salient objects detection models is the image compression, the axis 4 in my research project.

My on-going project is about end-to-end trainable models for image compression,

the PhD thesis topic of T. Ladune (to the end of 2021). Recent machine learning methods for lossy image compression have generated significant interest in both the machine learning and image processing communities since the work in [211], on which all the current end-to-end networks are based. In these approaches, image compression is achieved by first mapping pixel data into a quantized latent representation and then losslessly compressing the latents. Within the deep learning research community, the transforms typically take the form of CNNs, which learn nonlinear functions with the potential to map pixels into a more compressible latent space than the linear transforms used by traditional image codecs. To improve compression performance, recent methods have focused on better encoder/decoder transforms and on more sophisticated entropy models. Our objective in this project is to attain a better quality with a certain compression ratio. There are two possible ways we will explore to improve the existing models: 1) The current model is trained to minimize the PSNR, that's why the performance is not really related to subjective test results. We can try to minimize a perceptual IQM, instead of the PSNR, to see if the perceptual quality can be improved at the end. Another way is to integrate another CNN trained to learn the perceptual quality in the compression scheme, thus the transform coding can be directly guided by the perceptual quality. 2) At the workshop and challenge on Learned Image Compression at CVPR 2019, I noticed that existing deep-learning based compression algorithms perform actually very bad on salient regions for human perception (e.g. human face). Hence, the perceptual quality has a great chance to be improved if we take into account the visual attention mechanism in the current quantization step. As faces play an important role in entertainment images, a mixed saliency map model including face detection will be a good choice.

With the development of IQMs and salient objects detection models in the 3 themes, the fruits can also be applied on compression of the corresponding image types in the same ways, e.g. light field images, 360° images, drone videos and medical images. The works concerning light field images and 360° images would even be able to be disseminated via contributions to two international standardization groups (JPEG Pleno and MPEG-I Visual).

BIBLIOGRAPHY

[1] ITU-T Recommendation BT-500, “Methodology for the Subjective Assessment of the Quality of Television Pictures”, *in*: (2012).

[2] ITU-T Recommendation P.910, “Subjective video quality assessment methods for multimedia applications”, *in*: (2008).

[3] ITU-T Recommendation BT.1788, “Methodology for the subjective assessment of video quality in multimedia applications”, *in*: (2007).

[4] Recommendation ITU-T P.911, “Subjective audiovisual quality assessment methods for multimedia applications”, *in*: *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)* (1998).

[5] Recommendation ITU-T P.800, “Methods for subjective determination of transmission quality”, *in*: *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)* (1996).

[6] Alexandre Benoit et al., “Quality assessment of stereoscopic images”, *in*: *EURASIP journal on image and video processing* 2008.*1* (2009), p. 659024.

[7] ITU-T Recommendation G.1011, “Reference guide to quality of experience assessment methodologies”, *in*: (2016).

[8] Akira Takahashi, David Hands, and Vincent Barriac, “Standardization activities in the ITU for a QoE assessment of IPTV”, *in*: *IEEE Communications Magazine* 46.*2* (2008).

[9] ITU-T Rec. J.247, “Objective perceptual multimedia video quality measurement in the presence of a full reference”, *in*: (2008).

[10] ITU-T Rec. P.1201, “Parametric non-intrusive assessment of audiovisual media streaming quality”, *in*: (2012).

[11] ITU-T Rec. G.1070, “Opinion model for video-telephony applications”, *in*: (2012).

[12] ITU-T Rec. G.1071, “Opinion model for network planning of video and audio streaming applications”, *in*: (2016).

[13] ITU-T Rec. P.1202, “Parametric non-intrusive bitstream assessment of video media streaming quality”, *in*: (2012).

[14] ITU-T Rec. P.1203, “Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport”, *in*: (2017).

[15] ITU-T Rec. J.343, “Hybrid perceptual bitstream models for objective video quality measurments”, *in*: (2014).

[16] Shyamprasad Chikkerur et al., “Objective video quality assessment methods: A classification, review, and performance comparison”, *in*: *IEEE transactions on broadcasting* 57.*2* (2011), pp. 165–182.

[17] VQEG, “Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, Phase I”, *in*: (2008).

[18] Cisco Visual Networking Index Cisco, *Global Mobile Data Traffic Forecast Update, 2015–2020 White Paper, 2016*.

[19] *Skype usrs*, https://mspoweruser.com/skype-300-million-monthly-active-users/.

[20] *Skype calls*, https://news.microsoft.com/bythenumbers/skype-calls.

[21] Francesca De Simone et al., “Subjective quality assessment of H. 264/AVC video streaming with packet losses”, *in*: *EURASIP Journal on Image and Video Processing* 2011.*1* (2011), p. 190431.

[22] Recommendation ITU-T H.323, “Packet-based multimedia communications systems”, *in*: *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)* (2009).

[23] *Netdisturb software*, https://www.ffmpeg.org/.

[24] Benjamin Belmudez et al., “Audio and video channel impact on perceived audio-visual quality in different interactive contexts”, *in*: *Multimedia Signal Processing, 2009. MMSP’09. IEEE International Workshop on*, IEEE, 2009, pp. 1–5.

[25] Benjamin Belmudez and Sebastian Möller, “Audiovisual quality integration for interactive communications”, *in*: *EURASIP Journal on Audio, Speech, and Music Processing* 2013.*1* (2013), pp. 1–23.

[26] Takanori Hayashi et al., “Multimedia quality integration function for videophone services”, *in*: *Global Telecommunications Conference, 2007. GLOBECOM'07. IEEE*, IEEE, 2007, pp. 2735–2739.

[27] Recommendation ITU-T P.920, “Interactive test methods for audiovisual communications”, *in*: *International Telecommunication Union - Telecommunication Standardisation Sector (ITU-T)* (2000).

[28] Anna Hart, “Mann-Whitney test is not just a test of medians: differences in spread can be important”, *in*: *British Medical Journal* 323.*7309* (2001), p. 391.

[29] Jari Korhonen, Ulrich Reiter, and Eugene Myakotnykh, “On the relative importance of audio and video in the presence of packet losses”, *in*: *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, IEEE, 2010, pp. 64–69.

[30] Ralf Steinmetz, “Human perception of jitter and media synchronization”, *in*: *Selected Areas in Communications, IEEE Journal on* 14.*1* (1996), pp. 61–72.

[31] Zhou Wang et al., “Image quality assessment: from error visibility to structural similarity”, *in*: *IEEE transactions on image processing* 13.*4* (2004), pp. 600–612.

[32] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, “Multiscale structural similarity for image quality assessment”, *in*: *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2, IEEE, 2003, pp. 1398–1402.

[33] Margaret H Pinson and Stephen Wolf, “A new standardized method for objectively measuring video quality”, *in*: *IEEE Transactions on broadcasting* 50.*3* (2004), pp. 312–322.

[34] Kristian Skarseth et al., “OpenVQ: A Video Quality Assessment Toolkit”, *in*: *Proceedings of the 2016 ACM on Multimedia Conference*, ACM, 2016, pp. 1197–1200.

[35] Kalpana Seshadrinathan et al., “A subjective study to evaluate video quality assessment algorithms”, *in*: *IS&T/SPIE Electronic Imaging*, International Society for Optics and Photonics, 2010, 75270H–75270H.

[36] Phong V Vu and Damon M Chandler, “ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices”, *in*: *Journal of Electronic Imaging* 23.*1* (2014), pp. 013016–013016.

[37] Abdul Rehman, Kai Zeng, and Zhou Wang, “Display device-adapted video quality-of-experience assessment”, *in*: *SPIE/IS&T Electronic Imaging*, International Society for Optics and Photonics, 2015, pp. 939406–939406.

[38] *Netflix techblog*, http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html, [Online; accessed 17-January-2017].

[39] *MSU Software*, http://www.compression.ru/video/quality_measure/video_measurement_tool.html.

[40] Zhou Wang, Ligang Lu, and Alan C Bovik, “Video quality assessment based on structural distortion measurement”, *in*: *Signal processing: Image communication* 19.*2* (2004), pp. 121–132.

[41] *Video Quality Metric Software*, https://www.its.bldrdoc.gov/resources/video-quality-research/software.aspx.

[42] *MOtion-based Video Integrity Evaluation (MOVIE) Index*, http://live.ece.utexas.edu/research/quality/movie.html.

[43] *ViS3 source code*, http://vision.eng.shizuoka.ac.jp/vis3/.

[44] *SSIMplus software*, http://www.ssimwave.com/.

[45] *Perceptual video quality assessment based on multi-method fusion*, https://github.com/Netflix/vmaf.

[46] *OpenVQ Toolkit*, https://bitbucket.org/mpg_code/openvq.

[47] Anush Krishna Moorthy et al., “Video quality assessment on mobile devices: Subjective, behavioral and objective studies”, *in*: *IEEE Journal of Selected Topics in Signal Processing* 6.*6* (2012), pp. 652–671.

[48] Anush Krishna Moorthy et al., “Mobile Video Quality Assessment Database”, *in*: *IEEE ICC Workshop on Realizing Advanced Video Optimized Wireless Networks* (2012).

[49] Anush K Moorthy et al., “Subjective analysis of video quality on mobile devices”, *in*: *Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), Scottsdale, Arizona*, Citeseer, 2012.

[50] Francesca De Simone et al., “Subjective assessment of H. 264/AVC video sequences transmitted over a noisy channel”, *in*: *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, IEEE, 2009, pp. 204–209.

[51] Francesca De Simone et al., “A H. 264/AVC video database for the evaluation of quality metrics”, *in*: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010, pp. 2430–2433.

[52] Ines Saidi et al., “Audiovisual Quality Study For Videotelephony On Ip Networks”, *in*: *Proc. MMSP* (2016).

[53] Stephen Wolf and Margaret Pinson, “Video quality measurement techniques”, *in*: *2002.* (2002).

[54] Christoph Fehn, “Depth-Image-Based Rendering (DIBR), compression, and transmission for a new approach on 3D-TV”, *in*: *Electronic Imaging 2004*, International Society for Optics and Photonics, 2004, pp. 93–104.

[55] Howard Rheingold, *Virtual reality: exploring the brave new technologies*, Simon & Schuster Adult Publishing Group, 1991.

[56] Ronald T Azuma, “A survey of augmented reality”, *in*: *Presence: Teleoperators and virtual environments* 6.*4* (1997), pp. 355–385.

[57] Neus Sabater et al., “Dataset and Pipeline for Multi-view Light-Field Video”, *in*: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, IEEE, 2017, pp. 1743–1753.

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

[59] *The Future of Video: Enabling Immersion*, `https://developer.att.com/blog/shape-future-of-video`.

[60] *Industrial demonstration*, `https://igrv2017.sciencesconf.org/resource/page/id/12`.

[61] Michael Yuen and HR Wu, “A survey of hybrid MC/DPCM/DCT video coding distortions”, *in*: *Signal processing* 70.*3* (1998), pp. 247–278.

[62] Zhiqiang Zhou, Bo Wang, and Jinlei Ma, “Scale-aware edge-preserving image filtering via iterative global optimization”, *in*: *IEEE Transactions on Multimedia* (2017).

[63] Zhou Wang, Alan C Bovik, and BL Evan, “Blind measurement of blocking artifacts in images”, *in*: *International Conference on Image Processing, 2000.* Vol. 3, IEEE, 2000, pp. 981–984.

[64] Xiaojun Feng and Jan P Allebach, “Measurement of ringing artifacts in JPEG images”, *in*: *Proceedings of SPIE*, vol. 6076, 2006, pp. 74–83.

[65] E. Bosc et al., “An edge-based structural distortion indicator for the quality assessment of 3D synthesized views”, *in*: *2012 Picture Coding Symposium*, 2012, pp. 249–252.

[66] Philippe Hanhart and Touradj Ebrahimi, “Quality assessment of a stereo pair formed from decoded and synthesized views using objective metrics”, *in*: *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2012*, IEEE, 2012, pp. 1–4.

[67] Emilie Bosc et al., “Towards a new quality metric for 3-D synthesized view assessment”, *in*: *IEEE Journal of Selected Topics in Signal Processing* 5.*7* (2011), pp. 1332–1343.

[68] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, “Surf: Speeded up robust features”, *in*: *European conference on computer vision*, Springer, 2006, pp. 404–417.

[69] Martin A Fischler and Robert C Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”, *in*: *Readings in computer vision*, Elsevier, 1987, pp. 726–740.

[70] Huaizu Jiang et al., “Salient object detection: A discriminative regional feature integration approach”, *in*: *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, IEEE, 2013, pp. 2083–2090.

[71] Anush Krishna Moorthy and Alan Conrad Bovik, “Visual importance pooling for image quality assessment”, *in*: *IEEE journal of selected topics in signal processing* 3.*2* (2009), pp. 193–201.

[72] Xiangkai Liu et al., “Subjective and objective video quality assessment of 3D synthesized views with texture/depth compression distortion”, *in*: *IEEE Transactions on Image Processing* 24.*12* (2015), pp. 4847–4861.

[73] IVC-IRCCyN lab, *IRCCyN/IVC DIBR image database*, `http://ivc.univ-nantes.fr/en/databases/DIBR_Images/`, last accessed Aug. 30th 2017, [Online].

[74] Alexandru Telea, “An image inpainting technique based on the fast marching method”, *in*: *Journal of graphics tools* 9.*1* (2004), pp. 23–34.

[75] Yuji Mori et al., “View generation with 3D warping using depth information for FTV”, *in*: *Signal Processing: Image Communication* 24.*1* (2009), pp. 65–72.

[76] Karsten Mueller et al., “View synthesis for advanced 3D video systems”, *in*: *EURASIP Journal on Image and Video Processing* 2008.*1* (2009), pp. 1–11.

[77] Patrick Ndjiki-Nya et al., “Depth image based rendering with advanced texture synthesis”, *in*: *2010 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2010, pp. 424–429.

[78] Martin Köppel et al., “Temporally consistent handling of disocclusions with texture synthesis for depth-image-based rendering”, *in*: *2010 IEEE International Conference on Image Processing*, IEEE, 2010, pp. 1809–1812.

[79] Ke Gu et al., “Model-based referenceless quality metric of 3D synthesized images using local image description”, *in*: *IEEE Transactions on Image Processing* (2017).

[80] Federica Battisti et al., “Objective image quality assessment of 3D synthesized views”, *in*: *Signal Processing: Image Communication* 30 (2015), pp. 78–88.

[81] Dragana Sandić-Stanković, Dragan Kukolj, and Patrick Le Callet, “Multi–Scale Synthesized View Assessment Based on Morphological Pyramids”, *in*: *Journal of Electrical Engineering* 67.*1* (2016), pp. 3–11.

[82] D. Sandić-Stanković et al., “Free viewpoint video quality assessment based on morphological multiscale metrics”, *in*: *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6, DOI: `10.1109/QoMEX.2016.7498949`.

[83] Pierre-Henri Conze, Philippe Robert, and Luce Morin, “Objective view synthesis quality assessment”, *in*: *IS&T/SPIE Electronic Imaging*, International Society for Optics and Photonics, 2012, pp. 82881M–82881M.

[84] Zhou Wang et al., “Image quality assessment: from error visibility to structural similarity”, *in*: *IEEE Transactions on Image Processing* 13.*4* (2004), pp. 600–612, ISSN: 1057-7149, DOI: 10.1109/TIP.2003.819861.

[85] Zhou Wang, Eero P Simoncelli, and Alan C Bovik, “Multiscale structural similarity for image quality assessment”, *in*: *Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, 2004.* Vol. 2, IEEE, 2003, pp. 1398–1402.

[86] Zhou Wang and Qiang Li, “Information content weighting for perceptual image quality assessment”, *in*: *IEEE Transactions on Image Processing* 20.*5* (2011), pp. 1185–1198.

[87] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, “Making a completely blind image quality analyzer”, *in*: *IEEE Signal Processing Letters* 20.*3* (2013), pp. 209–212.

[88] Anush Krishna Moorthy and Alan Conrad Bovik, “A two-step framework for constructing blind image quality indices”, *in*: *IEEE Signal processing letters* 17.*5* (2010), pp. 513–516.

[89] Michele A Saad, Alan C Bovik, and Christophe Charrier, “Blind image quality assessment: A natural scene statistics approach in the DCT domain”, *in*: *IEEE Transactions on Image Processing* 21.*8* (2012), pp. 3339–3352.

[90] ITURBT ITU, “Methodology for the subjective assessment of video quality in multimedia applications”, *in*: *Rapport technique, International Telecommunication Union* (2007).

[91] Masayuki Tanimoto et al., “Reference softwares for depth estimation and view synthesis”, *in*: *ISO/IEC JTC1/SC29/WG11 MPEG* 20081 (2008), p. M15377.

[92] Ce Zhu and Shuai Li, “Depth image based view synthesis: New insights and perspectives on hole generation and filling”, *in*: *IEEE Transactions on Broadcasting* 62.*1* (2016), pp. 82–93.

[93] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama, “Region filling and object removal by exemplar-based image inpainting”, *in*: *IEEE Transactions on image processing* 13.*9* (2004), pp. 1200–1212.

[94] Guibo Luo et al., “A hole filling approach based on background reconstruction for view synthesis in 3D video”, *in*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1781–1789.

[95] Mashhour Solh and Ghassan AlRegib, “Hierarchical hole-filling for depth-based view synthesis in FTV and 3D video”, *in*: *IEEE Journal of Selected Topics in Signal Processing* 6.*5* (2012), pp. 495–504.

[96] Vincent Jantet, Christine Guillemot, and Luce Morin, “Object-based Layered Depth Images for improved virtual view synthesis in rate-constrained context”, *in*: *Image Processing (ICIP), 2011 18th IEEE International Conference on*, IEEE, 2011, pp. 125–128.

[97] Ilkoo Ahn and Changick Kim, “A novel depth-based virtual view synthesis method for free viewpoint video”, *in*: *IEEE Transactions on Broadcasting* 59.*4* (2013), pp. 614–626.

[98] Damiel Scharstein, R Szeliski, and C Pal, *Middlebury stereo datasets*, 2012.

[99] Ali Borji et al., “Salient object detection: A survey”, *in*: *arXiv preprint arXiv:1411.5878* (2014).

[100] Jia Li et al., “Primary Video Object Segmentation via Complementary CNNs and Neighborhood Reversible Flow”, *in*: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 1426–1434.

[101] Zoya Bylinskii et al., “What do different evaluation metrics tell us about saliency models?”, *in*: *IEEE transactions on pattern analysis and machine intelligence* 41.*3* (2019), pp. 740–757.

[102] Fanlei Yan, “Autonomous vehicle routing problem solution based on artificial potential field with parallel ant colony optimization (ACO) algorithm”, *in*: *Pattern Recognition Letters* 116 (2018), pp. 195–199.

[103] Junwei Han et al., “Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection: A Survey”, *in*: *IEEE Signal Process. Mag.* 35.*1* (2018), pp. 84–100.

[104] Jia Li, Changqun Xia, and Xiaowu Chen, “A Benchmark Dataset and Saliency-Guided Stacked Autoencoders for Video-Based Salient Object Detection”, *in*: *IEEE Trans. Image Processing* 27.*1* (2018), pp. 349–364.

[105] Thomas Brox and Jitendra Malik, “Object Segmentation by Long Term Analysis of Point Trajectories”, *in*: *Computer Vision - ECCV 2010 - 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V*, 2010, pp. 282–295.

[106] Peter Ochs, Jitendra Malik, and Thomas Brox, “Segmentation of Moving Objects by Long Term Video Analysis”, *in*: *IEEE Trans. Pattern Anal. Mach. Intell.* 36.*6* (2014), pp. 1187–1200.

[107] Federico Perazzi et al., “A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation”, *in*: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 724–732.

[108] Jordi Pont-Tuset et al., “The 2017 DAVIS Challenge on Video Object Segmentation”, *in*: *arXiv* abs/1704.00675 (2017).

[109] Houwen Peng et al., “RGBD Salient Object Detection: A Benchmark and Algorithms”, *in*: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III*, 2014, pp. 92–109.

[110] Ali Borji et al., “Salient Object Detection: A Benchmark”, *in*: *IEEE Trans. Image Processing* 24.*12* (2015), pp. 5706–5722.

[111] Yuhuan Chen et al., “SCOM: Spatiotemporal Constrained Optimization for Salient Object Detection”, *in*: *IEEE Trans. Image Processing* 27.*7* (2018), pp. 3345–3357.

[112] Trung-Nghia Le and Akihiro Sugimoto, “SpatioTemporal utilization of deep features for video saliency detection”, *in*: *2017 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops, Hong Kong, China, July 10-14, 2017*, 2017, pp. 465–470.

[113] Wenguan Wang, Jianbing Shen, and Ling Shao, “Video Salient Object Detection via Fully Convolutional Networks”, *in*: *IEEE Trans. Image Processing* 27.*1* (2018), pp. 38–49.

[114] Trung-Nghia Le and Akihiro Sugimoto, “Video Salient Object Detection Using Spatiotemporal Deep Features”, *in*: *IEEE Trans. Image Processing* 27.*10* (2018), pp. 5002–5015.

[115] Hongmei Song et al., “Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection”, *in*: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, 2018, pp. 744–760.

[116] Trung-Nghia Le and Akihiro Sugimoto, “Deeply Supervised 3D Recurrent FCN for Salient Object Detection in Videos”, *in*: *British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017*, 2017.

[117] Xiao Wei et al., “Two-stream recurrent convolutional neural networks for video saliency estimation”, *in*: *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB 2017, Cagliari, Italy, June 7-9, 2017*, 2017, pp. 1–5.

[118] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid, “Learning Video Object Segmentation with Visual Memory”, *in*: (2017), pp. 4491–4500.

[119] Jingchun Cheng et al., “SegFlow: Joint Learning for Video Object Segmentation and Optical Flow”, *in*: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017, pp. 686–695.

[120] Sergi Caelles et al., “One-Shot Video Object Segmentation”, *in*: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 5320–5329.

[121] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid, “Learning Motion Patterns in Videos”, *in*: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 531–539.

[122] Nian Liu and Junwei Han, “DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection”, *in*: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 678–686.

[123] Zhiming Luo et al., “Non-local Deep Features for Salient Object Detection”, *in*: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 6593–6601.

[124] Lijun Wang et al., “Learning to Detect Salient Objects with Image-Level Supervision”, *in*: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 3796–3805.

[125] Yi Tang et al., “Weakly Supervised Salient Object Detection with Spatiotemporal Cascade Neural Networks”, *in*: *IEEE Trans. Circuits Syst. Video Techn.* (2018).

[126] Qibin Hou et al., “Deeply Supervised Salient Object Detection with Short Connections”, *in*: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 5300–5309.

[127] Qibin Hou et al., “Deeply Supervised Salient Object Detection with Short Connections”, *in*: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 5300–5309.

[128] Liang-Chieh Chen et al., “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”, *in*: *IEEE Trans. Pattern Anal. Mach. Intell.* 40.*4* (2018), pp. 834–848.

[129] Thomas Brox and Jitendra Malik, “Large Displacement Optical Flow: Descriptor Matching in Variational Motion Estimation”, *in*: *IEEE Trans. Pattern Anal. Mach. Intell.* 33.*3* (2011), pp. 500–513.

[130] Karen Simonyan and Andrew Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition”, *in*: *arXiv* abs/1409.1556 (2014).

[131] Kaiming He et al., “Deep Residual Learning for Image Recognition”, *in*: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778.

[132] Alexey Dosovitskiy et al., “FlowNet: Learning Optical Flow with Convolutional Networks”, *in*: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 2758–2766.

[133] Tie Liu et al., “Learning to Detect a Salient Object”, *in*: *IEEE Trans. Pattern Anal. Mach. Intell.* 33.*2* (2011), pp. 353–367.

[134] Chuan Yang et al., “Saliency Detection via Graph-Based Manifold Ranking”, *in*: *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, 2013, pp. 3166–3173.

[135] Guanbin Li and Yizhou Yu, “Visual Saliency Detection Based on Multiscale Deep CNN Features”, *in*: *IEEE Trans. Image Processing* 25.*11* (2016), pp. 5012–5024.

[136] Qiong Yan et al., “Hierarchical Saliency Detection”, *in*: *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, 2013, pp. 1155–1162.

[137] Fuxin Li et al., “Video Segmentation by Tracking Many Figure-Ground Segments”, *in*: *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, 2013, pp. 2192–2199.

[138] Roozbeh Mottaghi et al., “The Role of Context for Object Detection and Semantic Segmentation in the Wild”, *in*: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, 2014, pp. 891–898.

[139] Nikolaus Mayer et al., “A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation”, *in*: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 4040–4048.

[140] Sergey Ioffe and Christian Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”, *in*: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2015, pp. 448–456.

[141] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation”, *in*: *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, 2016, pp. 565–571.

[142] Jianming Zhang et al., “Minimum Barrier Salient Object Detection at 80 FPS”, *in*: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015, pp. 1404–1412.

[143] Qiong Wang, Lu Zhang, and Kidiyo Kpalma, “Fast filtering-based temporal saliency detection using Minimum Barrier Distance”, *in*: *2017 IEEE International Conference on Multimedia & Expo Workshops, ICME Workshops, Hong Kong, China, July 10-14, 2017*, 2017, pp. 232–237.

[144] Yinlin Hu, Rui Song, and Yunsong Li, “Efficient Coarse-to-Fine Patch Match for Large Displacement Optical Flow”, *in*: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 5704–5712.

[145] Simon Baker et al., “A Database and Evaluation Methodology for Optical Flow”, *in*: *International Journal of Computer Vision* 92.*1* (2011), pp. 1–31.

[146] Kaiming He, Jian Sun, and Xiaoou Tang, “Guided Image Filtering”, *in*: *Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I*, 2010, pp. 1–14.

[147] Ken Fukuchi et al., “Saliency-based video segmentation with graph cuts and sequentially updated priors”, *in*: *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, ICME 2009, June 28 - July 2, 2009, New York City, NY, USA*, 2009, pp. 638–641.

[148] Lijun Wang et al., “Deep networks for saliency detection via local estimation and global search”, *in*: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 3183–3192.

[149] Rui Zhao et al., “Saliency detection by multi-context deep learning”, *in*: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 1265–1274.

[150] Gayoung Lee, Yu-Wing Tai, and Junmo Kim, “Deep Saliency with Encoded Low Level Distance Map and High Level Features”, *in*: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 660–668.

[151] Guanbin Li and Yizhou Yu, “Deep Contrast Learning for Salient Object Detection”, *in*: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 478–487.

[152] Linzhao Wang et al., “Saliency Detection with Recurrent Fully Convolutional Networks”, *in*: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, 2016, pp. 825–841.

[153] Esa Rahtu et al., “Segmenting Salient Objects from Images and Videos”, *in*: *Computer Vision - ECCV 2010 - 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V*, 2010, pp. 366–379.

[154] Anestis Papazoglou and Vittorio Ferrari, “Fast Object Segmentation in Unconstrained Video”, *in*: *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*, 2013, pp. 1777–1784.

[155] Alon Faktor and Michal Irani, “Video Segmentation by Non-Local Consensus voting”, *in*: *British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5, 2014*, 2014.

[156] Wenguan Wang, Jianbing Shen, and Fatih Porikli, “Saliency-aware geodesic video object segmentation”, *in*: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 3395–3402.

[157] Wenguan Wang, Jianbing Shen, and Ling Shao, “Consistent Video Saliency Using Local Gradient Flow Optimization and Global Refinement”, *in*: *IEEE Trans. Image Processing* 24.*11* (2015), pp. 4185–4196.

[158] Mary-Luc Champel and Síbastien Lasserre, “The Special Challenges of Offering High Quality Experience for VR Video”, *in*: *SMPTE 2016 Annual Technical Conference and Exhibition*, SMPTE, 2016, pp. 1–10.

[159] Yu-Chuan Su and Kristen Grauman, “Learning Compressible 360° Video Isomers”, *in*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7824–7833.

[160] Mai Xu et al., “Assessing visual quality of omnidirectional videos”, *in*: *IEEE Transactions on Circuits and Systems for Video Technology* (2018).

[161] Chen Li et al., “Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model”, *in*: *arXiv preprint arXiv:1807.10990* (2018).

[162] Suiyi Ling, Gene Cheung, and Patrick Le Callet, “No-Reference Quality Assessment for Stitched Panoramic Images Using Convolutional Sparse Coding and Compound Feature Selection”, *in*: *2018 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2018, pp. 1–6.

[163] Cagri Ozcinar, Julián Cabrera, and Aljosa Smolic, “Visual Attention-Aware Omnidirectional Video Streaming Using Optimal Tiles for Virtual Reality”, *in*: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* (2019).

[164] Cagri Ozcinar and Aljosa Smolic, “Visual attention in omnidirectional video for virtual reality applications”, *in*: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2018, pp. 1–6.

[165] Yashas Rai, Patrick Le Callet, and Philippe Guillotel, “Which saliency weighting for omni directional image quality assessment?”, *in*: *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2017, pp. 1–6.

[166] Mai Xu et al., “Predicting head movement in panoramic video: A deep reinforcement learning approach”, *in*: *IEEE transactions on pattern analysis and machine intelligence* (2018).

[167] Yanyu Xu et al., “Gaze prediction in dynamic 360 immersive videos”, *in*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5333–5342.

[168] Evgeniy Upenik, Martin Řeřábek, and Touradj Ebrahimi, “Testbed for subjective evaluation of omnidirectional visual content”, *in*: *2016 Picture Coding Symposium (PCS)*, IEEE, 2016, pp. 1–5.

[169] Ana De Abreu, Cagri Ozcinar, and Aljosa Smolic, “Look around you: Saliency maps for omnidirectional images in VR applications”, *in*: *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2017, pp. 1–6.

[170] Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet, “A dataset of head and eye movements for 360 degree images”, *in*: *Proceedings of the 8th ACM on Multimedia Systems Conference*, ACM, 2017, pp. 205–210.

[171] Ali Borji and Laurent Itti, “State-of-the-art in visual attention modeling”, *in*: *IEEE transactions on pattern analysis and machine intelligence* 35.*1* (2012), pp. 185–207.

[172] Wenguan Wang et al., “Revisiting video saliency: A large-scale benchmark and a new model”, *in*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4894–4903.

[173] Ali Borji, “Saliency prediction in the deep learning era: An empirical investigation”, *in*: *arXiv preprint arXiv:1810.03716* (2018).

[174] Jonathan Harel, Christof Koch, and Pietro Perona, “Graph-based visual saliency”, *in*: *Advances in neural information processing systems*, 2007, pp. 545–552.

[175] Jianming Zhang and Stan Sclaroff, “Saliency detection: A boolean map approach”, *in*: *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 153–160.

[176] Xun Huang et al., “Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks”, *in*: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 262–270.

[177] Junting Pan et al., “Salgan: Visual saliency prediction with generative adversarial networks”, *in*: *arXiv preprint arXiv:1701.01081* (2017).

[178] Marcella Cornia et al., “Predicting human eye fixations via an lstm-based saliency attentive model”, *in*: *IEEE Transactions on Image Processing* 27.*10* (2018), pp. 5142–5154.

[179] Sen Jia and Neil DB Bruce, “Eml-net: An expandable multi-layer network for saliency prediction”, *in*: *arXiv preprint arXiv:1805.01047* (2018).

[180] Tilke Judd et al., “Learning to predict where humans look”, *in*: *2009 IEEE 12th international conference on computer vision*, IEEE, 2009, pp. 2106–2113.

[181] Ming Jiang et al., “Salicon: Saliency in context”, *in*: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1072–1080.

[182] Yin Li et al., “The secrets of salient object segmentation”, *in*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287.

[183] ISO/IEC JTC1/SC29/WG11, “Applications and Requirements for 3DAV”, *in*: *Doc. N5877* (2003).

[184] *HTC VIVE Specs*, `https://www.vive.com/us/product/vive-virtual-reality-system`, [Online].

[185] *Oculus Rift*, `https://www.oculus.com/rift`, [Online].

[186] Hans Strasburger, Ingo Rentschler, and Martin Jüttner, “Peripheral vision and pattern recognition: A review”, *in*: *Journal of vision* 11.*5* (2011), pp. 13–13.

[187] Jesús Gutiérrez et al., “Introducing UN Salient360! Benchmark: A platform for evaluating visual attention models for 360° contents”, *in*: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2018, pp. 1–3.

[188] Jesús Gutiérrez et al., “Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360 still images”, *in*: *Signal Processing: Image Communication* 69 (2018), pp. 35–42.

[189] Jesús Gutiérrez et al., “Introducing UN Salient360! Benchmark: A platform for evaluating visual attention models for 360° contents”, *in*: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2018, pp. 1–3.

[190] Vincent Sitzmann et al., “Saliency in VR: How do people explore virtual environments?”, *in*: *IEEE transactions on visualization and computer graphics* 24.*4* (2018), pp. 1633–1642.

[191] Thomas Maugey, Olivier Le Meur, and Zhi Liu, “Saliency-based navigation in omnidirectional image”, *in*: *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2017, pp. 1–6.

[192] Rafael Monroy et al., “Salnet360: Saliency maps for omni-directional images with cnn”, *in*: *Signal Processing: Image Communication* 69 (2018), pp. 26–34.

[193] Pierre Lebreton and Alexander Raake, “GBVS360, BMS360, ProSal: Extending existing saliency prediction models from 2D to omnidirectional images”, *in*: *Signal Processing: Image Communication* 69 (2018), pp. 69–78.

[194] Mikhail Startsev and Michael Dorr, “360-aware saliency estimation with conventional image saliency predictors”, *in*: *Signal Processing: Image Communication* 69 (2018), pp. 43–52.

[195] Jing Ling et al., “A saliency prediction model on 360 degree images using color dictionary based sparse representation”, *in*: *Signal Processing: Image Communication* 69 (2018), pp. 60–68.

[196] Yucheng Zhu, Guangtao Zhai, and Xiongkuo Min, “The prediction of head and eye movement for 360 degree images”, *in*: *Signal Processing: Image Communication* 69 (2018), pp. 15–25.

[197] Ziheng Zhang et al., “Saliency detection in 360 videos”, *in*: *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 488–503.

[198] Roopak R Tamboli et al., “Super-multiview content with high angular resolution: 3D quality assessment on horizontal-parallax lightfield display”, *in*: *Signal Processing: Image Communication* 47 (2016), pp. 42–55.

[199] Roopak R Tamboli et al., “Objective Quality Assessment of 2D Synthesized Views for Light-Field Visualization”, *in*: *2018 International Conference on 3D Immersion (IC3D)*, IEEE, 2018, pp. 1–7.

[200] Vamsi Kiran Adhikarla et al., “Towards a quality metric for dense light fields”, *in*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 58–67.

[201] Yuming Fang et al., “Light Filed Image Quality Assessment by Local and Global Features of Epipolar Plane Image”, *in*: *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, IEEE, 2018, pp. 1–6.

[202] Irene Viola and Touradj Ebrahimi, “Valid: Visual quality assessment for light field images dataset”, *in*: *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2018, pp. 1–3.

[203] Pradip Paudyal, Federica Battisti, and Marco Carli, “Reduced Reference Quality Assessment of Light Field Images”, *in*: *IEEE Transactions on Broadcasting* 65.*1* (2019), pp. 152–165.

[204] Chen Li et al., “Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model”, *in*: *arXiv preprint arXiv:1807.10990* (2018).

[205] Mai Xu et al., “Assessing visual quality of omnidirectional videos”, *in*: *IEEE Transactions on Circuits and Systems for Video Technology* (2018).

[206] Yashas Rai, Patrick Le Callet, and Philippe Guillotel, “Which saliency weighting for omni directional image quality assessment?”, *in*: *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2017, pp. 1–6.

[207] Wenyan Yang et al., “Object detection in equirectangular panorama”, *in*: *2018 24th International Conference on Pattern Recognition (ICPR)*, IEEE, 2018, pp. 2190–2195.

[208] Deng-Ping Fan et al., “Shifting more attention to video salient object detection”, *in*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8554–8564.

[209] Jan Sokalski, Toby P Breckon, and Ian Cowling, “Automatic salient object detection in UAV imagery”, *in*: *Proc. 25th International Unmanned Air Vehicle Systems* (2010), pp. 11–1.

[210] Mahdi S Hosseini et al., “Atlas of Digital Pathology: A Generalized Hierarchical Histological Tissue Type-Annotated Database for Deep Learning”, *in*: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11747–11756.

[211] David Minnen, Johannes Ballé, and George D Toderici, “Joint autoregressive and hierarchical priors for learned image compression”, *in*: *Advances in Neural Information Processing Systems*, 2018, pp. 10771–10780.