SALGAN360: VISUAL SALIENCY PREDICTION ON 360 DEGREE IMAGES WITH GENERATIVE ADVERSARIAL NETWORKS

Fang-Yi Chao, Lu Zhang, Wassim Hamidouche, Olivier Deforges

Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, F-35000 Rennes, France
{fang-yi.chao, lu.ge, wassim.hamidouche, olivier.deforges}@insa-rennes.fr

ABSTRACT

Understanding visual attention of observers on 360° images gains interest along with the booming trend of Virtual Reality applications. Extending existing saliency prediction methods from traditional 2D images to 360° images is not a direct approach due to the lack of a sufficient large 360° image saliency database. In this paper, we propose to extend the SalGAN, a 2D saliency model based on the generative adversarial network, to SalGAN360 by fine tuning the SalGAN with our new loss function to predict both global and local saliency maps. Our experiments show that the SalGAN360 outperforms the tested state-of-the-art methods.

Index Terms— 360° image, omnidirectional image, saliency prediction, deep convolutional neuron network, generative adversarial network (GAN)

1. INTRODUCTION

The 360° images, or omnidirectional images, which capture the scene in all directions around a given center play an important role on the development of Virtual Reality (VR). By wearing Head-Mounted Displays (HMD), it enables to create an immersive experience to observers by allowing them to rotate their heads in the three axis (roll, pitch, yaw) to explore the perspective content according to the viewpoints. Considering that observers may be only interested in one part of 360° images and neglect other parts [1], visual saliency prediction becomes essential to understand user behavior. Visual saliency prediction outputs a saliency map estimating the probability distribution of human visual attention over an image. The generated information can be used in a wide range of computer vision applications [2] including compression, segmentation and retargeting.

Saliency models are based on datasets collected by tracking human fixation locations on images when human observers explore the images with no specific intention. Saliency prediction techniques on traditional 2D images have been recently widely investigated. With the advance of Deep Convolutional Neuron Network (DCNN) and the availability of large scale saliency datasets, saliency models based on DCNN achieved top performance on MIT Saliency Benchmark [3]. However, these models are not directly applicable on 360° images due to the serious distortion on its top and bottom regions caused by equirectangular projection. Moreover, the lack of a sufficient large 360° image saliency dataset results in obstacle to train a new saliency model. In order to take advantages from traditional 2D models, we propose the SalGAN360, which is based on the SalGAN [4] and adapted for 360° images via transfer learning. The main contributions of this paper are: 1) both local and global saliency distributions of the 360° image are estimated and fused ; 2) the 360° image is projected from equirectangular format into multiple cubic format to simulate undistorted contents presented to observers on HMD; 3) a new loss function taking into account more evaluation metrics is introduced to fine tune the layers in SalGAN to optimize its performance on 360° images.

The rest of paper is structured as follows. Section 2 gives a review on the state-of-the-art saliency prediction methods for 360° images. Section 3 presents the whole architecture of our SalGAN360 from preprocessing to transfer learning on the SalGAN. Section 4 describes the performance evaluation experiment settings and compares the SalGAN360 with the state-of-the-art models. Section 5 concludes the paper and provides ideas for future work.

2. RELATED WORK

The most existing saliency prediction methods for 360° images were extended from those designed for traditional 2D images. Their procedures can be separated into two parts: projection and saliency detection. Maugey *et al.* [5] projected a 360° image into double cubes, then estimated their saliency via an aggregation of feature extraction models consisting of Graph-Based Visual Saliency (GBVS) [6], Image Signature (ImgSig) [7], Adaptive Whitening Saliency (AWS) model [8], multi-scale rarity-based saliency detection (RARE2012) model [9], Boolean Map Approach (BMS) [10] and a face detector [11]. Considering that observers tend to look at more the equator area, Lebreton *et al.* proposed the GBVS360 [12], a model combining the adaptive equatorial prior with the saliency map predicted from the GBVS on rectilinear images projected from a 360° images. Monroy

978-1-5386-1737-3/18/$31.00

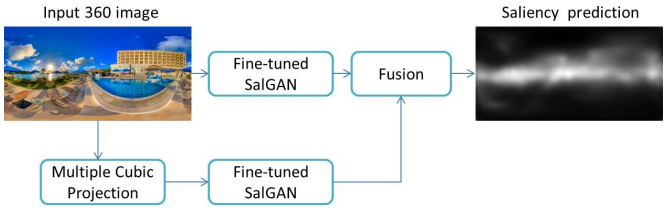


Fig. 1. Diagram of SalGAN360.

et al. presented the SalNet360 [13] including two parts: the first part predicts a saliency map for a 2D image (inspired by the SalNet [14]) and the second part refines the saliency by merging the output of the first part with the spherical coordinates. De Abreu *et al.* introduced the Fused Saliency Maps (FSM) [15] postprocessing method which deals with the center prior limitation of current saliency models. It can be used after any saliency models to average saliency maps predicted from four horizontal translated 360° images.

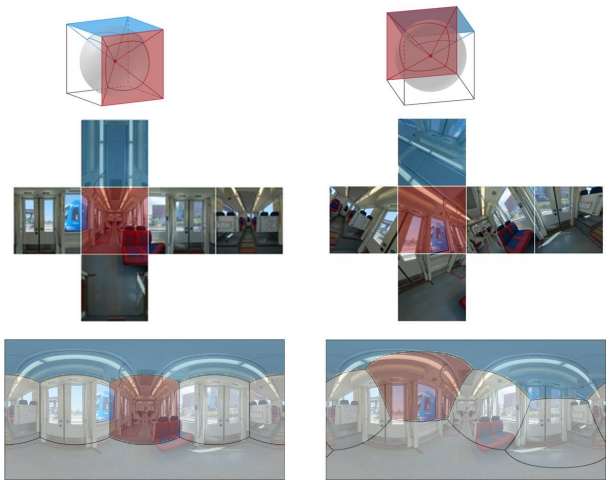
In this paper, we present a new method that fuses global and local saliency maps predicted from a fine-tuned SalGAN. Unlike the SalGAN trained on the Binary Cross Entropy (BCE), we fine-tuned our model via a new loss function taking into account various evaluation metrics at the same time.

3. PROPOSED SALGAN360 MODEL

This section describes the architecture of the SalGAN360 which predicts the saliency map for 360° images. The overall diagram is shown in **Fig. 1**. In its 1st part illustrated on the top of **Fig. 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.

3.1. 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° Field of View (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 HMD, we transfer an equirectangular image into multiple cubic maps by rotating the center of cube to multiple horizontal and vertical angles. **Fig. 2** shows



(a) (b)

Fig. 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.

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 **Fig. 2**(b)). Each rectilinear image is provided as an input to the saliency prediction model independently to estimate local saliency maps.

3.2. Fine tuning of SalGAN

The central element of SalGAN360 is extended from the SalGAN, a Generative Adversarial Networks (GAN) composed of two 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 dataset, we fine tune the network by retraining SalGAN initialized with pretrained weights. **Table 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:

$$L = \mu_{BCE} + \sigma_{BCE}(L') \quad (1)$$

$$L' = L^{KLdiv}_{normal}(\hat{S}, S^{den}) - L^{CC}_{normal}(\hat{S}, S^{den}) - L^{NSS}_{normal}(\hat{S}, S^{fix}) \quad (2)$$

Table 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	

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 - Kullback-Leibler divergence (KL) , Pearson's Correlation Coefficient (CC) and Normalized Scanpath Saliency (NSS) - normalized as follows:

$$L_{normal}(\hat{S}, S^{den}) = \frac{L(\hat{S}, S^{den}) - \mu}{\sigma} \tag{3}$$

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 (BCE), which is defined as:

$$L_{BCE} = -\frac{1}{N}\sum_{j=1}^{N} S_j^{den} \log \hat{S}_j + (1 - S_j^{den}) \log (1 - \hat{S}_j) \tag{4}$$

As in [4], the final loss function for the generator during adversarial training can be expressed as:

$$L_{GAN} = \alpha L - \log D(I, \hat{S}) \tag{5}$$

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

3.3. 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:

$$\hat{S}_{360} = 0.5\hat{S}_{Global} + 0.5\hat{S}_{Local} \tag{6}$$

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.

4. RESULTS

4.1. Experimental Setup

The dataset used to fine tune the network contains 40 images of head and eye movements provided by the University of Nantes [1]. We use 30 images to train and 10 images to validate our model. In the MCP step, for each training and validation image, we rotate cube in every 30°, that is 0°, 30°, 60° both horizontally and vertically (note that rotating 90° equals 0°). There are thus $3 \times 3 = 9$ rotations to obtain $9 \times 6 = 54$ patches for each image. Totally, we produced $30 \times 54 = 1620$ training patches and $10 \times 54 = 540$ validation patches. Note that although a smaller rotation angle produces more patches in different viewpoint, it also causes more overlaps between patches which may give rise to overfitting in our model. We set an initial learning rate to one tenth of previous.

For the performance comparison, all the methods are tested on the 25 images provided by Salient360! Grand Challenge ICME2017 [16], which are different from the images in [1]. Note that for the test images, cube is rotated in every 10° in the SalGAN360 to get a smoother map, thus a test image is divided into $9 \times 9 \times 6 = 486$ patches. The performance is evaluated by KL, CC, NSS and the Area Under the receiver operating characteristic Curve (AUC) as in [17]. The details of the four metrics are provided in [18].

4.2. Ablation Analysis and comparison to the state-of-the-art

The contribution of the fine-tuning step and the fusion step are shown in **Table 2**, where most of the four tested metrics results have been improved after fine tuning.

The SalGAN360 is compared in **Table 3** with the SalGAN and four state-of-the-art models for 360° images. It shows that the SalGAN360 outperforms all the tested methods for all the evaluation metrics. The method of Maugay *et al.*, the SalNet360 and the GBVS360 were the participants of the Salient360! Grand Challenge ICME2017, their performances are provided by the organizers of the challenge. The method of Maugay *et al.* submitted to the challenge is different from [5]. It only aggregates RARE2012, BMS and the face detector. **Table 4** shows the performance validated by the organizers of the Grand Challenge Salient360! ICME2018. The model here is trained on the dataset [17] which contains 85 images. We take 60 images to train and 25 images to test. To prevent the model from overfitting, we set the rotation angle in the MCP step to 45°, and change the parameters in the

fusion step to $\hat{S}_{360} = 0.25\hat{S}_{Global} + 0.75\hat{S}_{Local}$ from experiment.

Table 2. Results of Local SalGAN Saliency Map before and after fine-tuning & Results of global, local and fused saliency Map (here "ft" indicates "fine-tuned". In bold - the best performance in each sub-table)

Method	KL↓	CC↑	NSS↑	AUC↑
Local SalGAN	0.477	0.648	0.611	0.665
Local ft SalGAN	**0.426**	**0.690**	**0.880**	**0.726**

Method	KL↓	CC↑	NSS↑	AUC↑
Global ft SalGAN	0.642	0.516	0.922	0.741
Local ft SalGAN	**0.426**	**0.690**	0.880	0.726
Fused ft SalGAN	0.431	0.659	**0.971**	**0.746**

Table 3. Comparison results on dataset [16]

Method	KL↓	CC↑	NSS↑	AUC↑
SalGAN	1.236	0.452	0.810	0.708
SalGAN&FSM [15]	0.896	0.512	0.910	0.723
Maugey *et al.*	0.585	0.448	0.506	0.644
SalNet360[13]	0.458	0.548	0.755	0.701
GBVS360 [12]	0.698	0.527	0.851	0.714
SalGAN360	**0.431**	**0.659**	**0.971**	**0.746**

Table 4. Results of SalGAN360 on [17]

KL↓	CC↑	NSS↑	AUC↑
0.739	0.642	1.585	0.820

5. CONCLUSIONS

In this work we have presented the SalGAN360, a new model predicting the saliency map for 360° images. We show that the SalGAN360 has better performance than the tested state-of-the-art models, by applying the multiple cubic projection, fusing the global and local saliency maps and fine tuning with our new loss function.

6. REFERENCES

[1] 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 (MMSys'17)*, Taipei, Taiwan, Jun. 2017, pp. 205–210.

[2] J. Li and W. Gao, "Saliency-based applications," in *Visual Saliency Computation. Lecture Notes in Computer Science*. 2014, vol. 8404, Springer, Cham.

[3] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba, "Mit saliency benchmark," http://saliency.mit.edu/.

[4] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i-Nieto, "SalGAN: Visual saliency prediction with generative adversarial networks," in *arXiv*, January 2017.

[5] Thomas Maugey, Olivier Le Meur, and Zhi Liu, "Saliency-based navigation in omnidirectional image," in *IEEE 19th International Workshop on Multimedia Signal Processing*. IEEE, 2017.

[6] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proceedings of Neural Information Processing Systems (NIPS)*. MIT Press, 2006.

[7] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," in *IEEE transactions on pattern analysis and machine intelligence*, 2012, vol. 34, pp. 194–201.

[8] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," in *Image and Vision Computing*, 2012, vol. 30, pp. 51–64.

[9] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit, "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," in *Signal Processing: Image Communication*, 2013, vol. 28, pp. 642–58.

[10] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 153–160.

[11] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 2879–2886.

[12] Pierre Lebreton and Alexander Raake, "GBVS360, BMS360, ProSal: Extending existing saliency prediction models from 2D to omnidirectional images," *Signal Processing: Image Communication*, 2018.

[13] Rafael Monroy, Sebastian Lutz, Tejo Chalasani, and Aljosa Smolic, "SalNet360: Saliency maps for omni-directional images with cnn," in *Signal Processing: Image Communication*. Elsevier, September 2017.

[14] Junting Pan, Elisa Sayrol, Xavier Giro-i-Nieto, Kevin McGuinness, and Noel E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[15] Ana De Abreu, Cagri Ozcinar, and Aljosa Smolic, "Look around you: Saliency maps for omnidirectional images in VR applications," *Proceedings of the IEEE Ninth International Conference on Quality of Multimedia Experience (QoMEX'17)*, pp. 1–6, Jun. 2017.

[16] University of Nantes, "Salient360!: Visual attention modeling for 360 images grand challenge," in *Multimedia and Expo (ICME), 2017 IEEE International Conference on*, 2017.

[17] J. Gutiérrez, E. David, Y. Rai, and P. Le Callet, "Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360 still images," in *Signal Processing: Image Communication*, 2018.

[18] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand, "What do different evaluation metrics tell us about saliency models?," *arXiv preprint arXiv:1604.03605*, 2016.