Panoramic convolutions for 360º single-image saliency prediction

Daniel Martin Ana Serrano Belen Masia
Universidad de Zaragoza, I3A
{danims, anase, bmasia}@unizar.es

Figure 1. We propose a CNN-based saliency predictor built over panoramic convolutions, which are a type of dilated convolutions that works on 360º equirectangular panoramas, a distorted projection of a 3D scene. Our model learns the intrinsic relations between pixels in a panorama given its gnomonic projection (i.e., points on the surface of the sphere are projected from the sphere's center to a tangent plane) to then extract a saliency map. Our system outperforms current state-of-the-art saliency predictors for 360º equirectangular panoramas.

Abstract

We present a convolutional neural network based on panoramic convolutions for saliency prediction in 360º equirectangular panoramas. Our network architecture is designed leveraging recently presented 360º-aware convolutions that represent kernels as patches tangent to the sphere where the panorama is projected, and a spherical loss function that penalizes prediction errors for each pixel depending on its coordinates in a gnomonic projection. Our model is able to successfully predict saliency in 360º scenes from a single image, outperforming other state-of-the-art approaches for panoramic content, and yielding more precise results that may help in the understanding of users' behavior when viewing 360º VR content. Our project is publicly available in `https://github.com/DaniMS-ZGZ/Panoramic-CNN-360-Saliency`

1. Introduction

Over the last years, we are witnessing an unprecedented growth in the field of virtual reality (VR), which is increasing its presence in consumers' homes. However, much remains unclear about what works and what does not work in terms of content for VR experiences. This is intrinsically tied to how users behave in virtual environments (VE). Some previous works have analyzed users' behavior in VR [16], providing some useful insights about how do people explore virtual environments, and in particular about visual saliency, gaze, and fixation behaviors. With the recent development of deep learning techniques, many data-driven approaches for saliency estimation have been developed, that yield reasonable good results [13, 5]. Nevertheless, most of those techniques have been trained with conventional images. Thus, when employed with equirectangular panoramas, such as those employed in cinematic VR, their performance drastically drops.

To tackle this, some recent works have proposed different approaches to allow traditional saliency systems to work with equirectangular panoramas. Within them, two main types can be identified: heuristic and data-driven approaches [20]. Heuristic approaches usually share a common framework consisting of a pre-processing stage (e.g., projection changes), a feature extraction stage (e.g., 2D saliency estimation), and a post-processing stage (e.g., equator - center bias) that ends up with a 360º saliency map [9, 17, 8, 11, 1, 23]. Data-driven approaches, on the other hand, have adapted deep learning techniques to face some VR open problems such as classification [4] or saliency detection in 360º video [22, 10, 21, 18, 14]. In contrast to the latter, predicting saliency in static scenes is harder, since no motion nor optical flow cues are available. However, many VR applications as virtual tourism, industrial or urban design, or some exploring experiences only provide static sce-

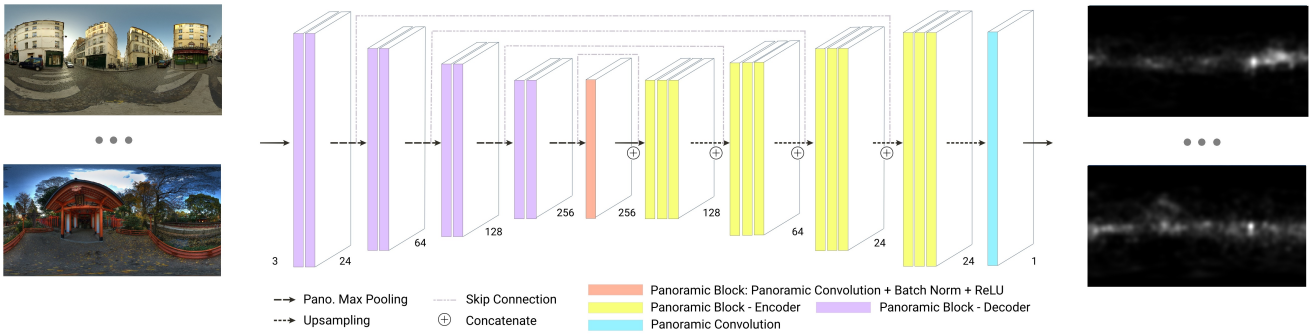

Figure 2. Proposed model. Our approach is based on an encoder-decoder architecture, where the first half extracts features from our input (RGB panoramas), and the second half learns how to reconstruct the output (a probabilistic saliency map) from that set of features. For each block, 360º-aware kernels are used in convolutions and pooling operations, so that relations between pixels are calculated in spherical space instead of in image space, thus taking into account panorama distortions.

narios. Although other works exist for this kind of content [3, 12, 19], we propose a panoramic CNN that learns feature relations from a gnomonic, non-distorted space. The gnomonic projection is a nonconformal map projection obtained by projecting some points on the surface of sphere from the sphere's center to their corresponding points in a plane that is tangent to that sphere (see Figure 1). With this technique, neither reprojections (e.g., cubemaps) nor patch decomposition is needed, hence global and local structure and spatial dimensions are preserved.

2. Model architecture

To address saliency prediction in panoramas, we propose an U-Net-like encoder-decoder architecture [15], (extensively used in many scenarios, such as segmentation or depth estimation problems), that takes a single equirectangular panorama as input and outputs a probabilistic saliency map, where each pixel p is assigned a value $s_p \in [0, 1]$, with a higher value yielding a higher probability of the user directing their attention to it. This kind of architecture has been widely exploited due to its ability to extract inherent features in the data and extrapolate information from them. Our model architecture can be seen in Figure 2, and has two clearly differentiated parts, encoder and decoder, separated by a bottleneck layer. Encoder layers are composed by two convolutional blocks, whereas decoder layers are composed by three convolutional blocks to facilitate the inference steps. Each convolutional block consists of a panoramic convolution, a batch normalization layer, and a ReLU. Each encoder-decoder analogous layers include skip connections to facilitate the saliency prediction process.

2.1. Panoramic convolutions

A key aspect of our architecture is that instead of common convolutions, we leverage recent advances in panoramic convolutions [4] to build our model. Those represent kernels as patches tangent to the sphere (see Figure 1) where the panorama would be projected. Each point s in a unit sphere surface is defined as a function of its latitude $\phi \in [\frac{-\pi}{2}, \frac{\pi}{2}]$ and longitude $\theta \in [\pi, \pi]$. For each of the points, there exists a tangent plane located at s, whose coordinates $x, y \in \mathbb{R}^2$ are related to a point on the sphere by its gnomonic projection. The relation between pixels in a panorama is thus established depending on the related pixels in the sphere, given the tangent plane for each of the pixels. We build both convolutions and pooling layers in our model following this strategy.

2.2. Loss function

Since equirectangular panoramas are strongly distorted, errors in each pixel should not penalize equally for all pixels: errors in the top and bottom regions of the panoramas are less critical than errors on the equator. This is even more relevant given the equator bias in VR [16]. Therefore, we utilize a 360º-aware loss function $\mathcal{L}$ [22] which is an adaptation of a traditional L2 loss function, such that:

$$\mathcal{L} = \sum_{\phi=0,\theta=0}^{\Phi,\Theta} \omega_{\phi,\theta}(S_{\phi,\theta} - P_{\phi,\theta})^2 \qquad (1)$$

where S and P are the ground-truth and the predicted saliency maps, respectively; ϕ and θ represent each pixel's latitude and longitude in the spherical geometry, and $\omega_{\phi,\theta}$ is the weight of each pixel's error proportional to its solid angle in the sphere (i.e., how large each part of the image appears to an observer looking from that pixel). Given the nature of this problem, other loss functions (e.g, Riemannian distance loss [7]) could be further investigated.

2.3. Dataset

One of the main limitations of most 360º learning systems is the limited amount of data available. Panoramas are complex to generate and capturing hardware is not as

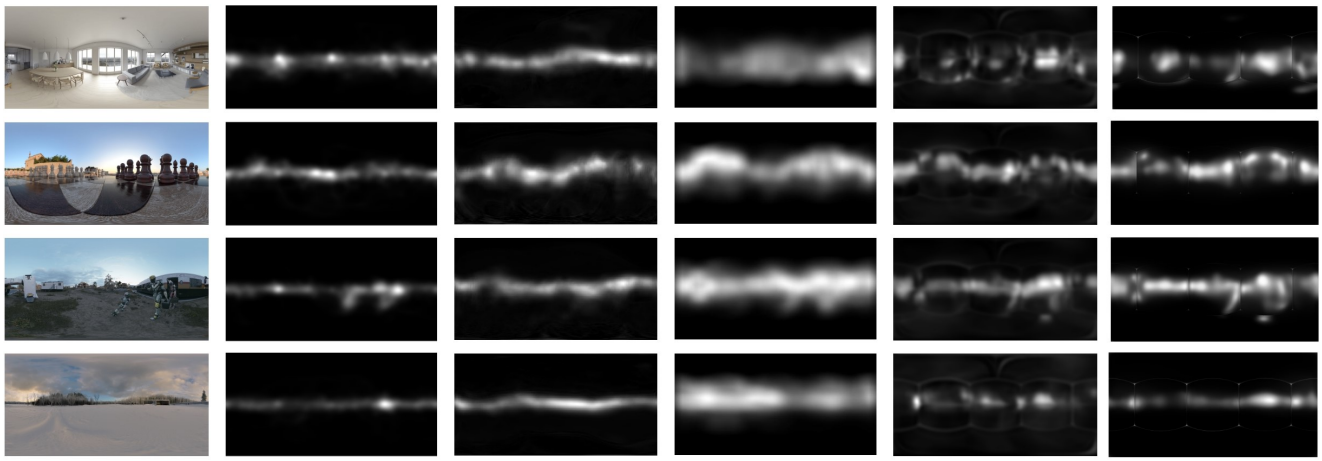

Figure 3. Results comparison between our model results and current state-of-the-art 360º saliency predictors. From left to right: RGB original panorama, ground truth saliency map, our method, Xia et al. (whose method is an adaptation of their own previous work [19] to work with 360º panoramas), Monroy et al. (SalNet360 [12]) and Chao et al. (SalGan360 [3]). Note that our system (third column, in bold) yields the most precise results, avoiding the overestimation that other solutions suffer of.

	Similarity ↑	CC ↑	AUC_Borji ↑	AUC_Judd ↑	KLDiv ↓	NSS ↑	IG ↑
Xia et al. 2016 [19]	0.5935	0.6269	0.8793	0.8816	0.7657	1.6008	2.0721
Monroy et al. 2017 [12]	0.5147	0.5596	0.8401	0.8636	0.7613	2.4898	2.0721
Chao et al. 2018 [3]	0.6385	0.7095	0.9006	0.9304	1.2000	**3.3000**	2.9064
Ours	**0.6469**	**0.7212**	**0.9624**	**0.9650**	**0.4411**	3.0309	**2.9727**

Table 1. Comparison of different metric scores proposed by Bylinskii et al. [2] (we refer the reader to this reference for an in-depth explanation of these metrics). For each score, we indicate with an arrow which is better (a higher or lower value). The best result for each of the metrics is indicated in bold. Note that our model outperforms all other current state-of-the-art methods in most of the cases.

widespread as traditional cameras. To feed our model, we use two public, 360º saliency datasets that include 85 [6] and 22 [16] panoramas with ground-truth saliency map computed from multiple users' gaze data. To alleviate the small amount of ground-truth data, we applied data augmentation to our collection. Rotating, cropping or shearing the images is not feasible since it would break the structure of the panorama. Thus, the only operation this type of content allows is flipping, either vertically, horizontally, or both ways. In addition, adding noise to the panoramas also benefits the network training process. For each original panorama we obtain 7 additional RGB + saliency map pairs (3 flips, and Gaussian, Poisson, salt-pepper and speckle noise), yielding a final dataset of $(85+22)*8 = 856$ images. We kept 7 images for testing and used the rest (with augmentation, a total of 800 samples) for training.

2.4. Training

We trained our model for 700 epochs, batching it in sets of 6 images. We applied Xavier's uniform initialization in our model, and our training parameters are: learning ratio = 10^{-4}, momentum = 0.9, and weight decay = 10^{-5}. The whole process was supported by a checkpointing system that saved all training parameters each iteration, and the model whenever it improved from previous iterations, following a cross-validation strategy.

3. Results

Figure 3 shows our model's predictions for four panoramas from our test set, compared to some state-of-the-art static saliency predictors, namely Xia et al.'s 360º saliency predictor based on their previous work [19], Monroy et al.'s SalNet360 [12] and Chao et al.'s SalGan360 [3]. For all four methods, we quantitatively evaluate some well-established metrics for saliency comparison [2]. Even though each metric is designed to measure different aspects of the prediction (i.e., location-based vs. distribution-based metrics), our system outperforms current state-of-the-art saliency predictors, since it is able to extract real, intrinsic feature relations in equirectangular panoramas more precisely, and predict visual attention based on them.

4. Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (CHAMELEON project, grant agreement No 682080).

References

[1] Federica Battisti, Sara Baldoni, Michele Brizzi, and Marco Carli. A feature-based approach for saliency estimation of omni-directional images. *Signal Processing: Image Communication*, 69:53–59, 2018.

[2] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2018.

[3] Fang-Yi Chao, Lu Zhang, Wassim Hamidouche, and Olivier Deforges. SAlGAN360: visual saliency prediction on 360 degree images with generative adversarial networks. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 01–04. IEEE, 2018.

[4] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 518–533, 2018.

[5] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3488–3493. IEEE, 2016.

[6] Jesús Gutiérrez, Erwan J David, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. 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)*, pages 1–3. IEEE, 2018.

[7] Benjamin Hou, Nina Miolane, Bishesh Khanal, Matthew CH Lee, Amir Alansary, Steven McDonagh, Jo V Hajnal, Daniel Rueckert, Ben Glocker, and Bernhard Kainz. Computing cnn loss and gradients for pose estimation with riemannian geometry. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 756–764. Springer, 2018.

[8] Pierre Lebreton and Alexander Raake. GBVS360, BMS360, ProSal: Extending existing saliency prediction models from 2D to omnidirectional images. *Signal Processing: Image Communication*, 69:69–78, 2018.

[9] Chen Li, Mai Xu, Xinzhe Du, and Zulin Wang. Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model. In *Proceedings of the 26th ACM International Conference on Multimedia*, pages 932–940, 2018.

[10] Wen-Chih Lo, Ching-Ling Fan, Jean Lee, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 360 video viewing dataset in head-mounted virtual reality. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 211–216, 2017.

[11] Guilherme Luz, João Ascenso, Catarina Brites, and Fernando Pereira. Saliency-driven omnidirectional imaging adaptive coding: Modeling and assessment. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2017.

[12] Rafael Monroy, Sebastian Lutz, Tejo Chalasani, and Aljosa Smolic. Salnet360: Saliency maps for omni-directional images with cnn. *Signal Processing: Image Communication*, 69:26–34, 2018.

[13] 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.

[14] 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)*, pages 1–6. IEEE, 2017.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015.

[16] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein. How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 2017.

[17] Mikhail Startsev and Michael Dorr. 360-aware saliency estimation with conventional image saliency predictors. *Signal Processing: Image Communication*, 69:43–52, 2018.

[18] Tatsuya Suzuki and Takao Yamanaka. Saliency map estimation for omni-directional image considering prior distributions. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2079–2084. IEEE, 2018.

[19] Chen Xia, Fei Qi, and Guangming Shi. Bottom–up visual saliency estimation with deep autoencoder-based sparse reconstruction. *IEEE Transactions on Neural Networks and Learning Systems*, 27(6):1227–1240, 2016.

[20] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. State-of-the-art in 360 video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):5–26, 2020.

[21] Mai Xu, Yuhang Song, Jianyi Wang, MingLang Qiao, Liangyu Huo, and Zulin Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2693–2708, 2018.

[22] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao. Saliency detection in 360 videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 488–503, 2018.

[23] Yucheng Zhu, Guangtao Zhai, and Xiongkuo Min. The prediction of head and eye movement for 360 degree images. *Signal Processing: Image Communication*, 69:15–25, 2018.