ScanpathNet: A Recurrent Mixture Density Network for Scanpath Prediction (Supplementary Material)

Ryan Anthony Jalova de Belen, Tomasz Bednarz, and Arcot Sowmya
The University of New South Wales, Sydney, Australia
r.debelen@unsw.edu.au, t.bednarz@unsw.edu.au, a.sowmya@unsw.edu.au

1. ScanpathNet

ScanpathNet is a deep learning model inspired by Guided Search 6 (GS6) [12], the latest theoretical framework of visual search. Since this work focusses on free-viewing tasks, we present a modified GS6 without compromising its theoretical underpinnings (refer to [12] for a complete discussion of the theory). Nevertheless, ScanpathNet provides explicit extensibility for search-based tasks.

2. Dataset Used

Experiments were done on the OSIE [13], MIT1003 [7] and CAT2000 [3] from the MIT/Tubingen Saliency Benchmark[1]. For CAT2000, only the train set was used since the eye-tracking data for the test set were held out.

3. Comparison against the state-of-the-art

The state-of-the-art models for comparison were chosen due to the public availability of their implementations. They also represent the diversity of traditional and deep learning approaches to scanpath prediction.

- ScanpathNet: the code will be available on GitHub.[2]
- VQA [4]: we used the implementation on GitHub.[3]
- IOR-ROI [8]: we obtained the implementation on GitHub.[4] We used the Mask^X-RCNN [5][5] with a threshold of 0.5 as mentioned in the original paper to extract the semantic segmentation masks.
- PathGAN [1]: we used the implementation on Github.[6]
- SaltiNet [2]: we used the implementation on GitHub.[7]
- Star-FC [11]: we used the Python implementation on GitHub.[8]
- SGC [9, 10]: we used the Matlab implementation.[9]
- Itti [6]: we used the Matlab implementation.[10]

4. More experimental information

As mentioned in Section 3.2, we performed empirical tests to investigate the effect of different spatial masks on the IOR mechanism. More specifically, we compared the performance of using (1) a Gaussian spatial mask and (a) a spatial mask with 0 values around the fixation locations. After training on all datasets, we found that the model did not converge (i.e., the loss function did not decrease) when setting 1 was used. Setting 2 was used for further experiments because the model converged. It is important to note that the inhibition of return mechanism implemented here is different from the commonly used method of direct spatial inhibition of the saliency map. Here, we apply the inhibition in the feature space through an element-wise product. This design choice is inspired by the GS6 paper. Figure 3 in [12] shows that information from the world is represented in the visual system and suggests that succeeding operations happen in the feature space.

5. More qualitative results

We compare ScanpathNet against a wide selection of traditional and deep learning scanpath models. Extensive comparisons are provided in Figures 1, 2, 3 and 4. ScanpathNet predictions resemble human scanpaths and are qualitatively better than the predictions made by the other models. It is important to note that there are times when ScanpathNet prediction locations are not perfectly aligned with the ground truth fixation locations (Image 1 in Figure 2, Image 1 & 3 in Figure 4).

[1] https://saliency.tuebingen.ai/datasets.html
[2] https://github.com/ryanxdebelen/ScanpathNet
[3] https://github.com/chenxy99/Scanpaths
[4] https://github.com/sunwj/scanpath
[5] https://github.com/ronghanghu/seg_every_thing
[6] https://github.com/imatge-upc/pathgan
[7] https://github.com/massens/saliency-360salient-2017
[8] https://github.com/ykotseruba/pySTAR-FC
[9] https://github.com/XiaoshuaiSun/SGP
[10] http://www.saliencytoolbox.net

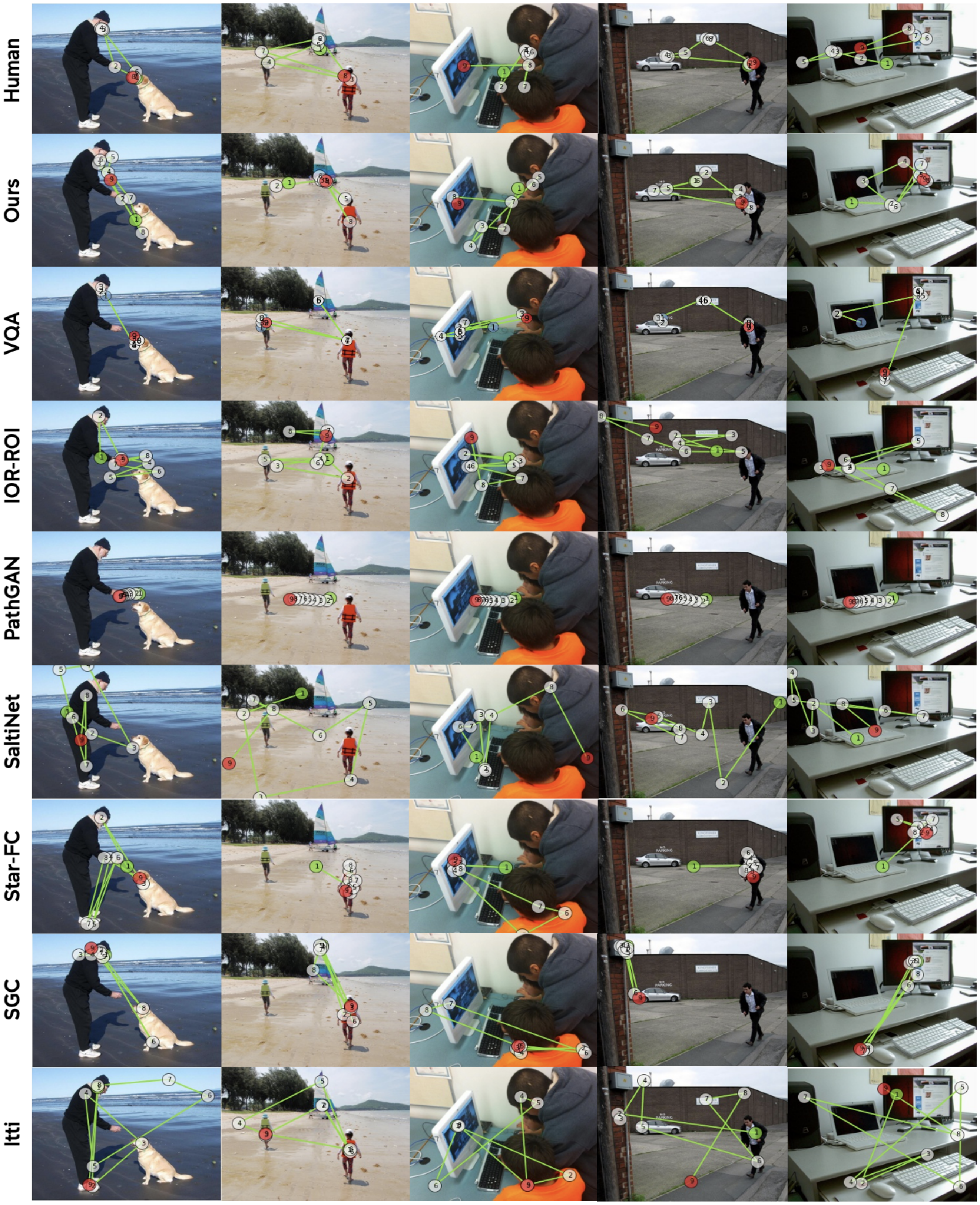

Figure 1. Visualisation of the generated scanpaths from each scanpath model on OSIE images with increasing complexity.

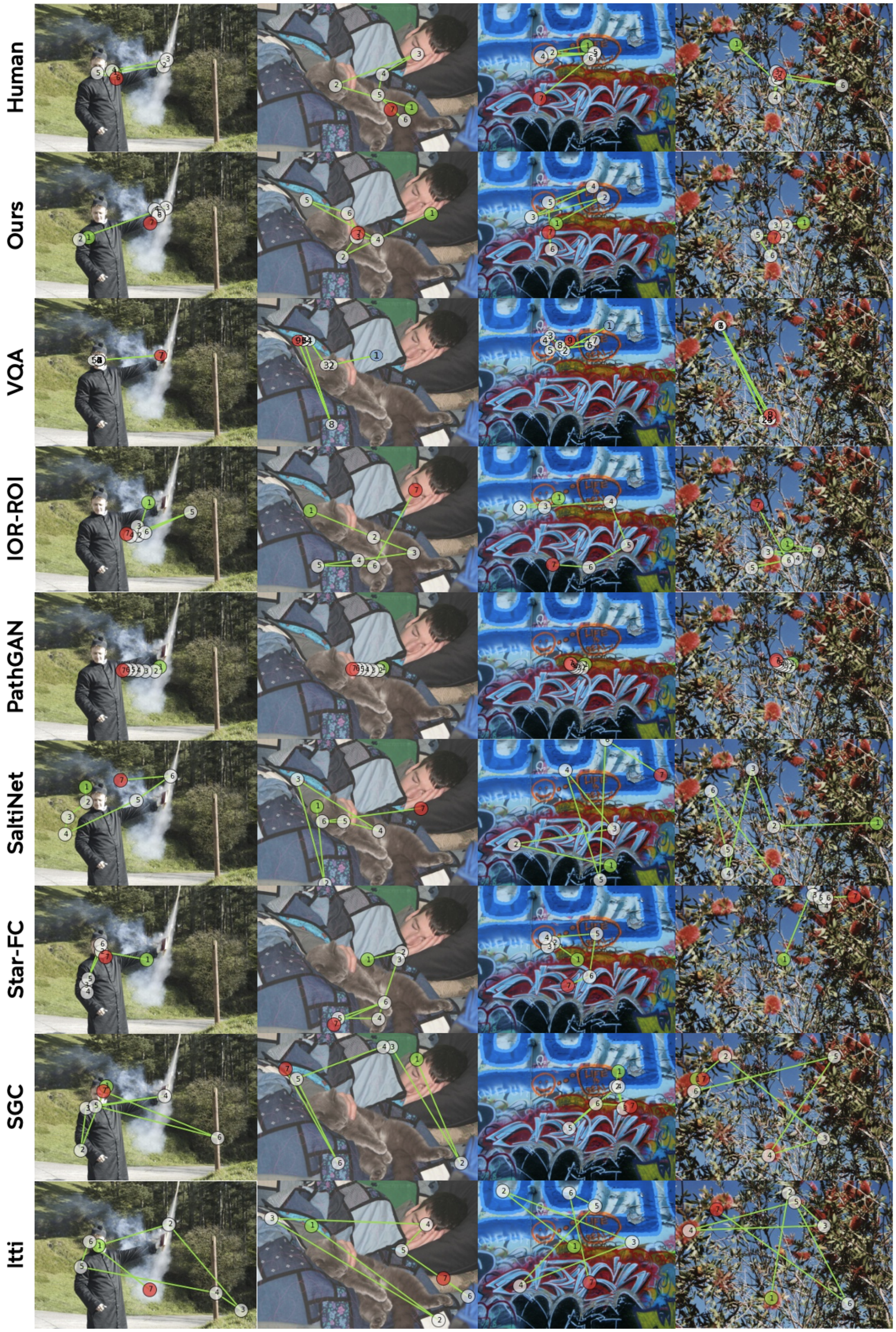

Figure 2. Visualisation of the generated scanpaths from each scanpath model on MIT1003 images with increasing complexity.

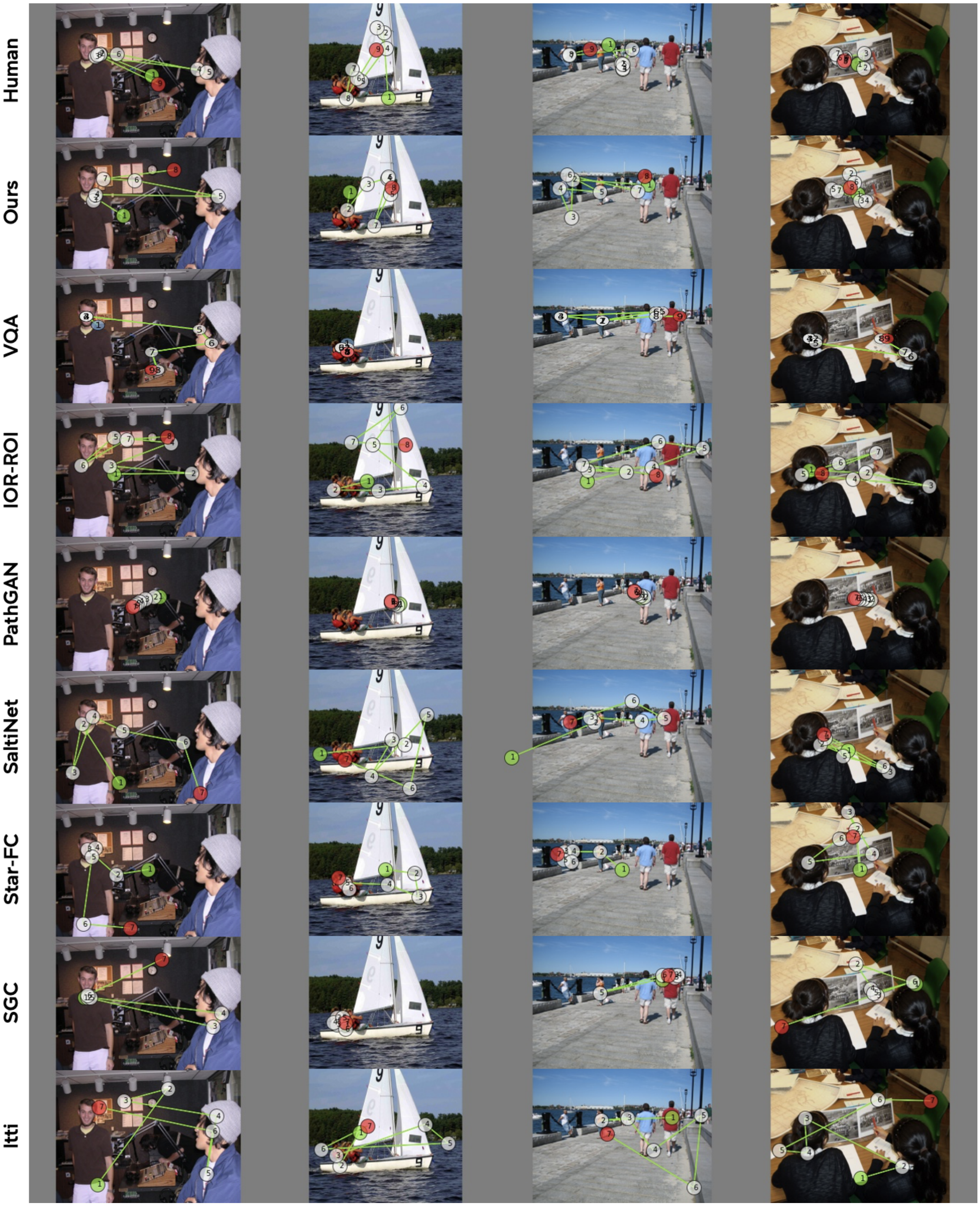

Figure 3. Visualisation of the generated scanpaths from each scanpath model on CAT2000 images with increasing complexity.

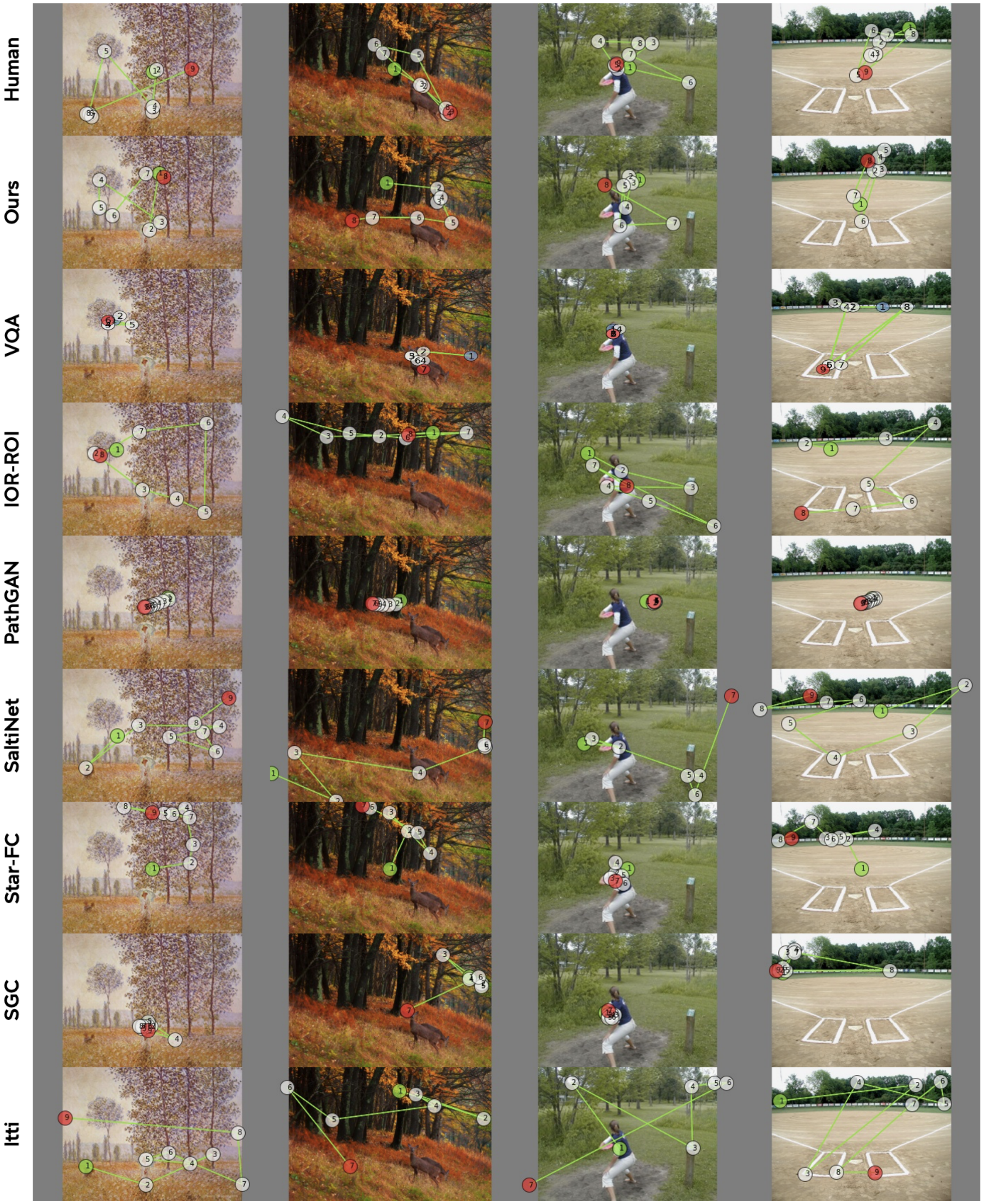

Figure 4. Visualisation of the generated scanpaths from each scanpath model on CAT2000 images with increasing complexity.

References

[1] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Pathgan: Visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[2] Marc Assens Reina, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2331–2338, 2017. 1

[3] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015. 1

[4] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10876–10885, 2021. 1

[5] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4233–4241, 2018. 1

[6] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998. 1

[7] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009. 1

[8] Wanjie Sun, Zhenzhong Chen, and Feng Wu. Visual scanpath prediction using ior-roi recurrent mixture density network. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1

[9] Xiaoshuai Sun, Hongxun Yao, and Rongrong Ji. What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1552–1559. IEEE, 2012. 1

[10] Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, and Xian-Ming Liu. Toward statistical modeling of saccadic eye-movement and visual saliency. *IEEE Transactions on Image Processing*, 23(11):4649–4662, 2014. 1

[11] Calden Wloka, Iuliia Kotseruba, and John K Tsotsos. Active fixation control to predict saccade sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3184–3193, 2018. 1

[12] Jeremy M Wolfe. Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, pages 1–33, 2021. 1

[13] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014. 1