TPP-Gaze: Modelling Gaze Dynamics in Space and Time with Neural Temporal Point Processes

Supplementary Material

[1]Alessandro D'Amelio, [2]Giuseppe Cartella, [2]Vittorio Cuculo,
[1]Manuele Lucchi, [2]Marcella Cornia, [2]Rita Cucchiara, [1]Giuseppe Boccignone
[1]University of Milan, Italy [2]University of Modena and Reggio Emilia, Italy
[1]{name.surname}@unimi.it, [2]{name.surname}@unimore.it

We introduced TPP-Gaze, a scanpath prediction method that models gaze dynamics as a neural temporal point process. In the following sections, we provide additional results showing evidence of the superiority of our proposed approach compared to the state-of-the-art. Additionally, we describe how the proposed approach can be extended for the visual search task.

A. Additional Quantitative Results

Additional Metrics on OSIE, NUSEF, and FiFa. As a complement of Table 3 of the main paper, we report in Table 5 the results on OSIE, NUSEF, and FiFa datasets in terms of SS and SED. Also for these metrics, TPP-Gaze achieves the best results when compared with models trained under the same settings and datasets. It is also worth noting that, especially for the NUSEF and FiFa datasets, our approach can achieve the best overall results in terms of SS with and without duration.

Scanpath Statistics on MIT1003, NUSEF, and FiFa. As discussed in the main paper, TPP-Gaze features scanpath statistics that better align with human behavior when compared with IOR-ROI-LSTM [3], DeepGazeIII [7] and Scanpath-VQA [2]. The same trend is appreciable from Fig. 8. Notably, even when tested on MIT1003, NUSEF, and FiFa, TPP-Gaze effectively models the long-tail distribution of both fixation durations and saccade amplitudes. In contrast, other methods tend to capture only the average human gaze dynamics. An exception is DeepGazeIII on NUSEF, which achieves comparable results for saccade amplitudes but does not model fixation duration.

Return Fixations Analysis on MIT1003, NUSEF, and FiFa. Fig. 9 complements the analysis reported in the main paper by showing the distribution of return fixations (RFs) for the MIT1003, NUSEF, and FiFa datasets. In these settings as well, TPP-Gaze demonstrates its ability to model RF patterns effectively, generally presenting an RF distribution that aligns better with human observers compared to other methods.

	OSIE			NUSEF			FiFa		
	SS (KL-Div) ↓		SED ↓	SS (KL-Div) ↓		SED ↓	SS (KL-Div) ↓		SED ↓
	w/ Dur	w/o Dur	Avg	w/ Dur	w/o Dur	Avg	w/ Dur	w/o Dur	Avg
Itti-Koch [4]	-	3.93	9.07	-	1.89	9.97	-	14.70	8.65
CLE (Itti) [1, 4]	-	3.24	9.29	-	1.40	10.16	-	12.62	8.86
CLE (DG) [1, 8]	-	3.65	9.23	-	1.35	10.07	-	14.38	8.83
G-Eymol [9]	12.28	2.95	8.00	1.99	0.53	8.02	13.17	5.00	6.13
IOR-ROI-LSTM [3]	0.20	2.84	8.82	0.06	1.10	9.69	0.30	12.44	8.27
DeepGazeIII [7]	-	2.51	8.47	-	1.04	9.38	-	12.08	7.97
Scanpath-VQA [2]	0.02	0.09	7.55	0.02	0.10	8.39	0.03	0.44	6.81
DeepGazeIII [7]	-	2.52	**8.57**	-	1.04	9.42	-	12.25	8.00
Scanpath-VQA [2]	0.29	0.31	9.70	0.06	0.18	10.61	0.35	0.90	9.73
TPP-Gaze (GRU)	**0.25**	**0.30**	8.05	**0.02**	**0.03**	8.41	**0.15**	**0.24**	**7.00**
TPP-Gaze (Transf.)	0.29	0.35	8.10	**0.02**	0.04	**8.40**	0.21	0.31	7.05

Table 5. Additional results on OSIE, NUSEF, and FiFa datasets. **Gray color** indicates models trained under the same settings and datasets. Within this group, **bold** values represent the best performance for each metric. Underline values indicate the overall best performance across all models and metrics.

B. Extending the Model to Visual Search

We extend the TPP-Gaze architecture to handle the visual search task by forcing the model to learn a task-specific semantic representation of the input image (see Fig. 10). Recall that the TPP-Gaze's semantic representation module consists of a DenseNet201 CNN backbone and a learnable readout network composed of three 1×1 convolutional layers with 8, 16, and 1 channels, respectively. The obtained spatial priority map is then projected to a fixed-dimensional vector, $\mathbf{z}_j$, to obtain the j-th image semantic representation. Specifically, the last layer performing a 1×1 convolution is responsible for learning a (non-linear) combination of the feature maps from the previous layers.

To guide the model toward a specific search objective, we redefine the architecture to enable TPP-Gaze to learn such a combination conditioned on a given text string. To this end, we first obtain a linguistic embedding of the search target using the RoBERTa language model. Let $\mathbf{F}_{target}$

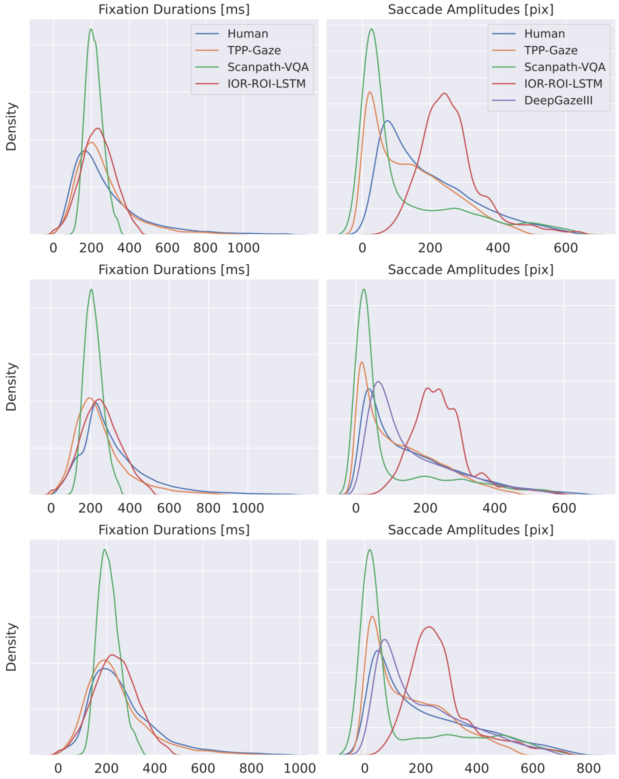


Figure 8. Statistical properties exhibited by TPP-Gaze and other methods relative to those of human observers, in terms of empirical fixation durations and saccade amplitudes on MIT1003 (top row), NUSEF (middle row) and FiFa (bottom row) datasets. For consistency with the main paper, comparison against DeepGazeIII on MIT1003 is omitted.

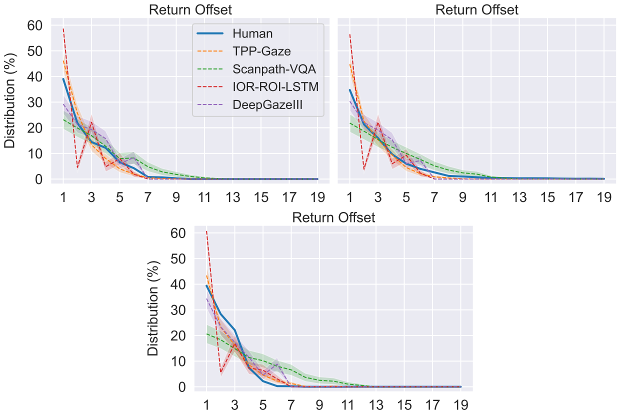


Figure 9. Return fixations analysis comparing TPP-Gaze with other methods and human observers. Results are shown on MIT1003 (top-left plot), NUSEF (top-right plot), and FiFa (bottom plot) datasets.

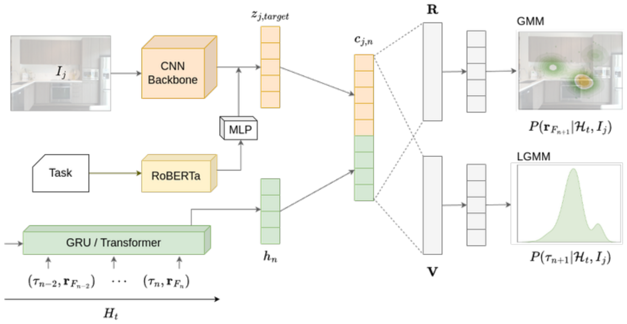

I_j
CNN Backbone
$z_{j,target}$
MLP
Task
RoBERTa
GRU / Transformer
$(\tau_{n-2}, \mathbf{r}_{F_{n-2}})$
$\cdots$
$(\tau_n, \mathbf{r}_{F_n})$
H_t
h_n
$c_{j,n}$
R
V
GMM
LGMM
$P(\mathbf{r}_{F_{n+1}}|\mathcal{H}_t, I_j)$
$P(\tau_{n+1}|\mathcal{H}_t, I_j)$

Figure 10. Overview of TPP-Gaze model architecture extended to handle the visual search task. A linguistic embedding (RoBERTa) of the search target is employed to learn a task-drive semantic representation (z_j). The latter, together with the history of past events (h_n), is used to simulated the next fixation position and duration.

be the embedding vector representing the search objective. The readout network for the visual search model consists of three 1×1 convolutional layers with 16, 64, and 256 channels, respectively. Thus, it is modified to output $M = 256$ feature maps. Let $\mathbf{X} = [\mathbf{x}_0; \cdots; \mathbf{x}_M] \in \mathbb{R}^{M \times d}$ represent the matrix of flattened image features. The task-specific semantic representation for the j-th image, $\mathbf{z}_{j,target}$, is then obtained as follows:

$$w = \text{softplus}(\text{MLP}(\mathbf{F}_{target}))$$
$$\mathbf{z}_{j,target} = \sum_{i=1}^{M} w_i \mathbf{x}_i. \qquad (1)$$

C. Additional Qualitative Results

Additional qualitative results are depicted from Fig. 11 to Fig. 15 on COCO-FreeView, MIT1003, OSIE, NUSEF, and FiFa datasets, respectively. Each fixation is represented by a circle, with its diameter proportional to the fixation duration. For methods that do not model fixation duration, circles are shown with a uniform size. The first fixation of each scanpath is omitted. The qualitative results support the findings of the main paper, highlighting the accuracy of TPP-Gaze in predicting human-like scanpaths. Other methods, instead, either overfit on a few salient objects, especially people and faces in the case of Scanpath-VQA, or predict scanpath trajectories containing fixations on unlikely locations (see the bottom sample in Fig. 13 or the top sample in Fig. 14).

In the main paper, we also quantitatively assess the performance of the scanpath models on the saliency prediction task. In particular, given a sample image, we construct the aggregated saliency map by convolving a Gaussian kernel over all the locations of predicted fixations [6]. To support our quantitative analysis, we present the saliency prediction of our model against the competitors from Fig. 16 to Fig. 20 on COCO-FreeView, MIT1003, OSIE, NUSEF, and FiFa datasets, respectively. Note that we include DeepGazeIII in the comparison for reference even though its results are not directly comparable. Indeed, DeepGazeIII was specifically trained on a large scale dataset [5] to predict saliency maps along with scanpaths. Nevertheless, TPP-Gaze outperforms DeepGazeIII and the other models in many cases, demonstrating better alignment with humans.

Finally, in Fig. 21 we show additional qualitative results on sample images from COCO-Search18 for the visual search task. As can be observed, TPP-Gaze can effectively simulate human-like goal-directed visual attention patterns for various target objects. The model demonstrates its ability to adapt, with a simple architectural variation, from a free-viewing setting to a task-specific visual search scenario. The results illustrate how TPP-Gaze generates plausible attention trajectories that focus on regions likely to contain the target object, mimicking the efficient search strategies employed by humans when looking for specific items in complex scenes.

References

[1] Giuseppe Boccignone and Mario Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, 331(1-2):207–218, 2004. 1

[2] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting Human Scanpaths in Visual Question Answering. In *CVPR*, 2021. 1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13

[3] Zhenzhong Chen and Wanjie Sun. Scanpath Prediction for Visual Attention using IOR-ROI LSTM. In *IJCAI*, 2018. 1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13

[4] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. PAMI*, 20:1254–1259, 1998. 1

[5] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, 2015. 2

[6] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, 2009. 2

[7] Matthias Kümmerer, Matthias Bethge, and Thomas SA Wallis. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5):7–7, 2022. 1, 4, 6, 7, 8, 9, 11, 12, 13

[8] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet. *arXiv preprint arXiv:1411.1045*, 2014. 1

[9] Dario Zanca, Stefano Melacci, and Marco Gori. Gravitational laws of focus of attention. *IEEE Trans. PAMI*, 42(12):2983–2995, 2020. 1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13

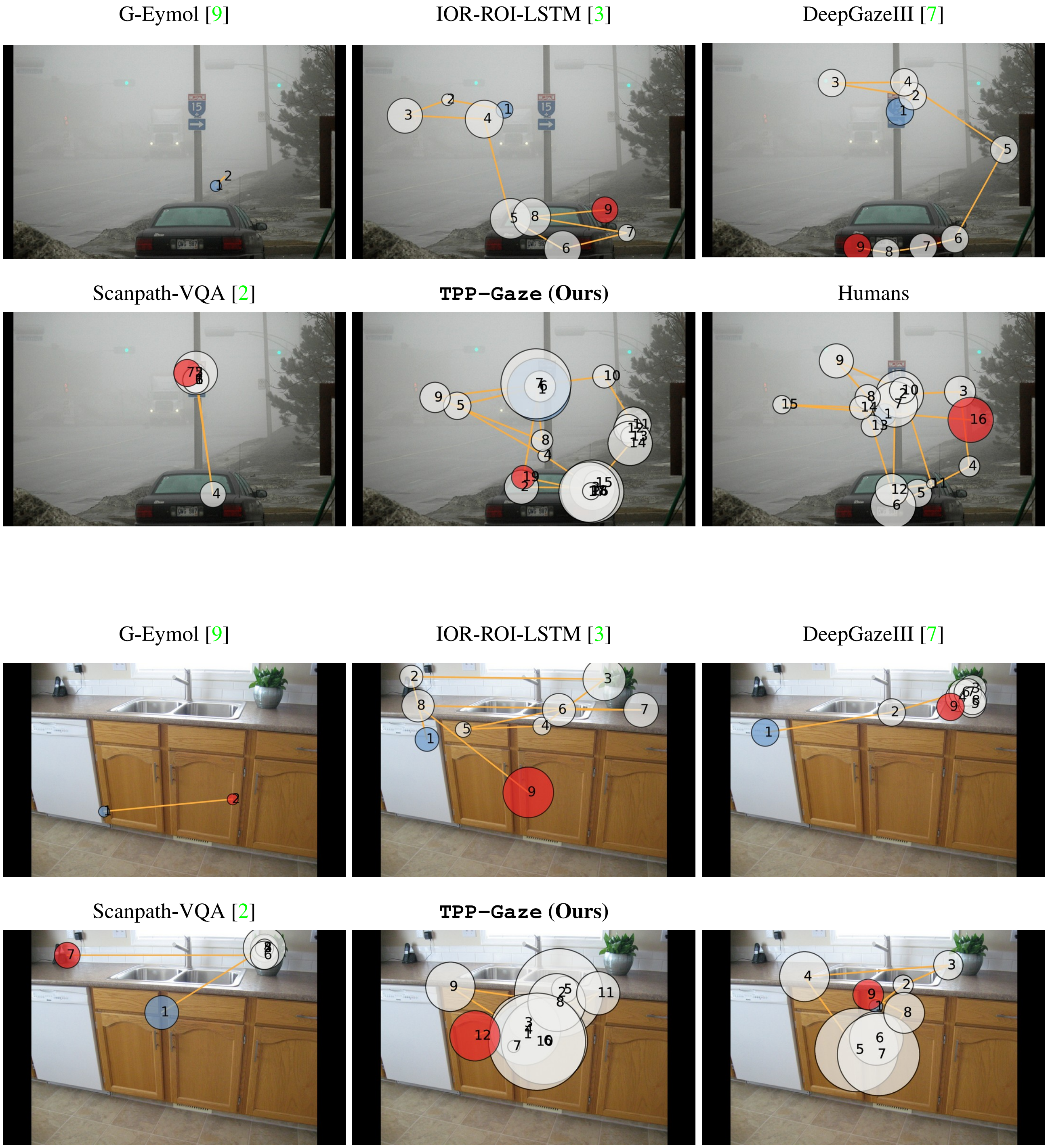
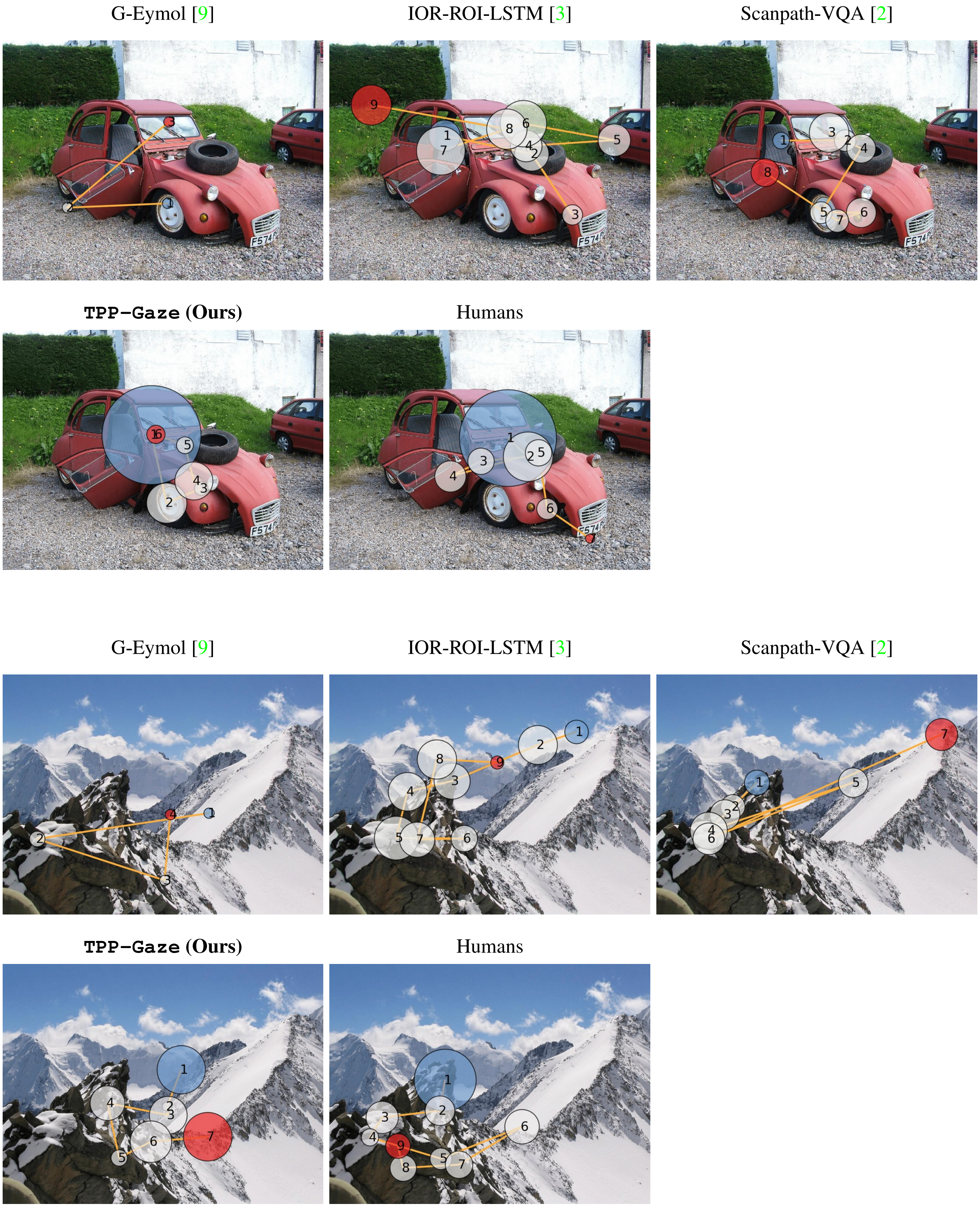


Figure 11. Qualitative comparison of simulated and human scanpaths on the COCO-FreeView dataset.



G-Eymol [9]

IOR-ROI-LSTM [3]

Scanpath-VQA [2]

`TPP-Gaze` (Ours)

Humans

G-Eymol [9]

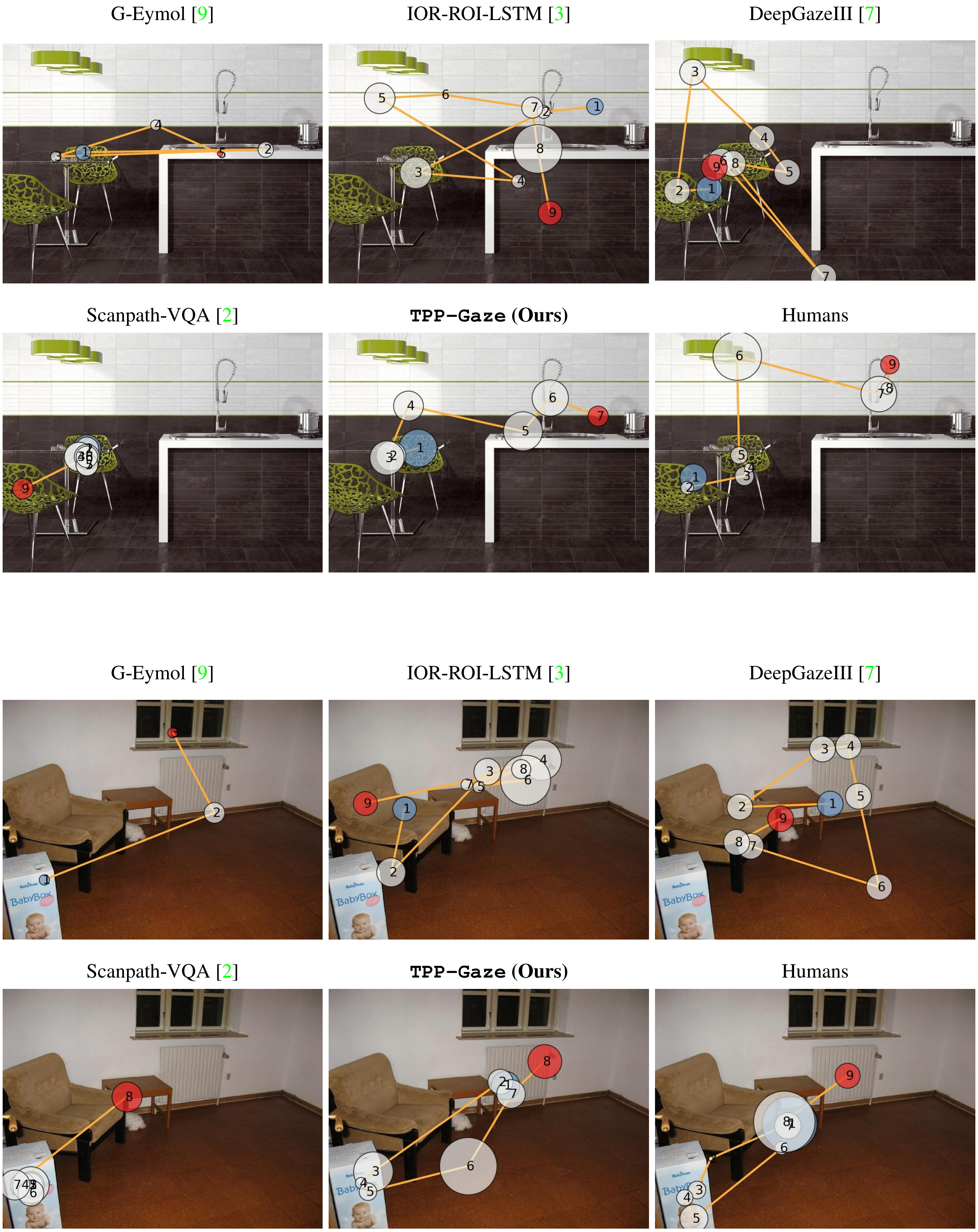
IOR-ROI-LSTM [3]

Scanpath-VQA [2]

`TPP-Gaze` (Ours)

Humans

Figure 12. Qualitative comparison of simulated and human scanpaths on the MIT1003 dataset. We omit DeepGazeIII for consistency with the experimental settings described in the main paper.



G-Eymol [9]

IOR-ROI-LSTM [3]

DeepGazeIII [7]

Scanpath-VQA [2]

`TPP-Gaze` (Ours)

Humans

G-Eymol [9]

IOR-ROI-LSTM [3]

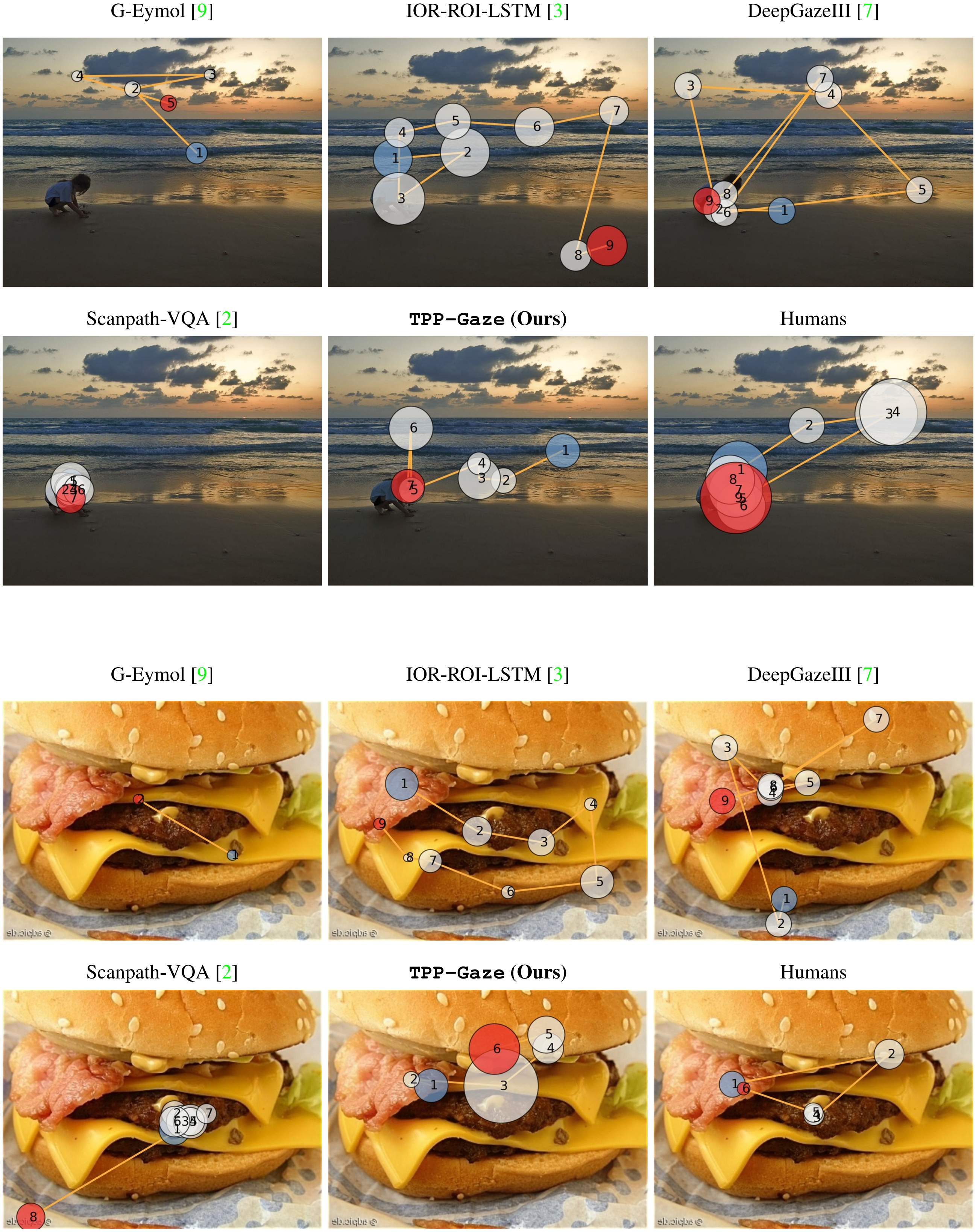
DeepGazeIII [7]

Scanpath-VQA [2]

`TPP-Gaze` (Ours)

Humans

Figure 13. Qualitative comparison of simulated and human scanpaths on the OSIE dataset.



G-Eymol [9]

IOR-ROI-LSTM [3]

DeepGazeIII [7]

Scanpath-VQA [2]

`TPP-Gaze` (Ours)

Humans

G-Eymol [9]

IOR-ROI-LSTM [3]

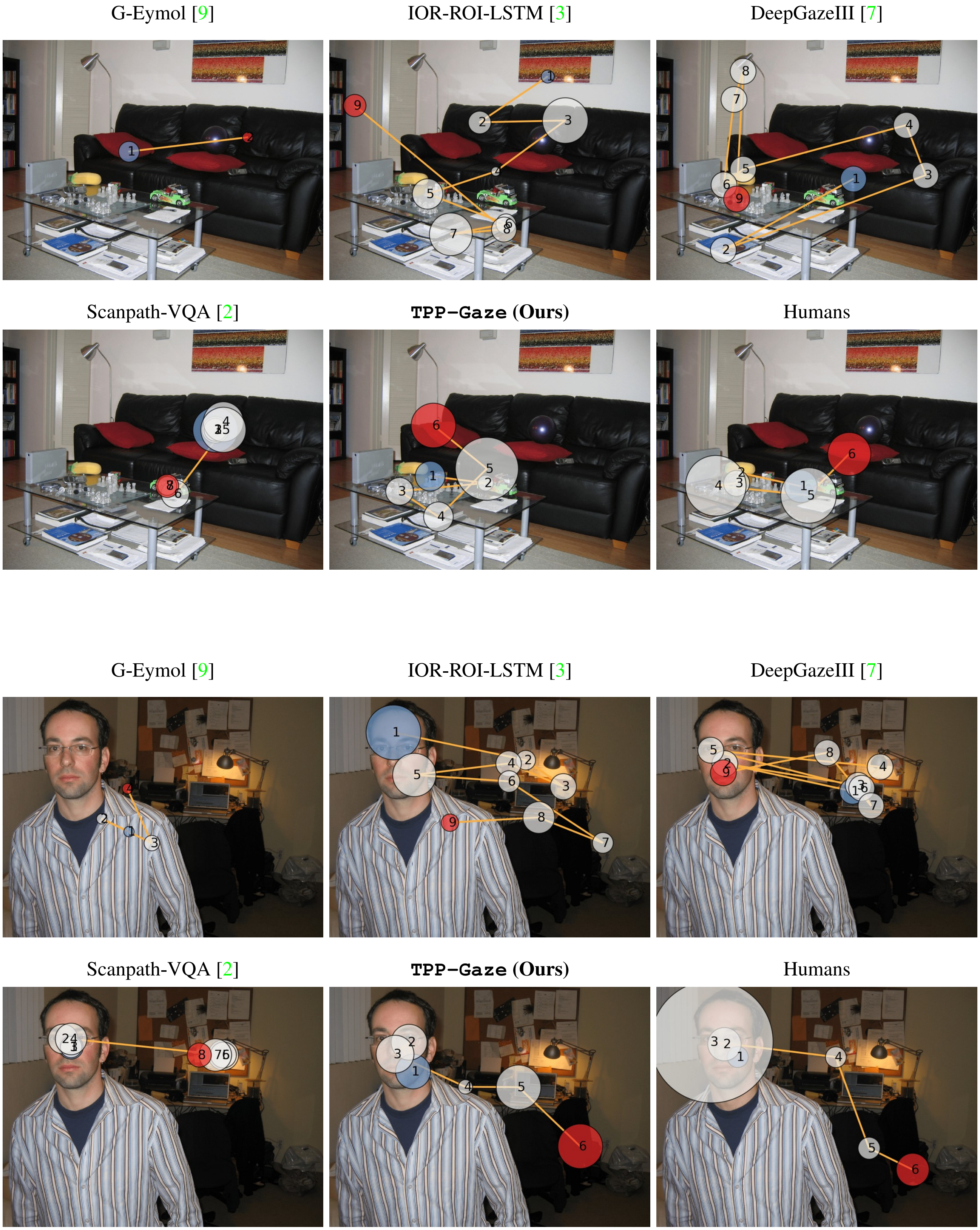
DeepGazeIII [7]

Scanpath-VQA [2]

`TPP-Gaze` (Ours)

Humans

Figure 14. Qualitative comparison of simulated and human scanpaths on the NUSEF dataset.


G-Eymol [9]

IOR-ROI-LSTM [3]

DeepGazeIII [7]

Scanpath-VQA [2]

`TPP-Gaze` (Ours)

Humans

G-Eymol [9]

IOR-ROI-LSTM [3]

DeepGazeIII [7]

Scanpath-VQA [2]

`TPP-Gaze` (Ours)

Humans

Figure 15. Qualitative comparison of simulated and human scanpaths on the FiFa dataset.

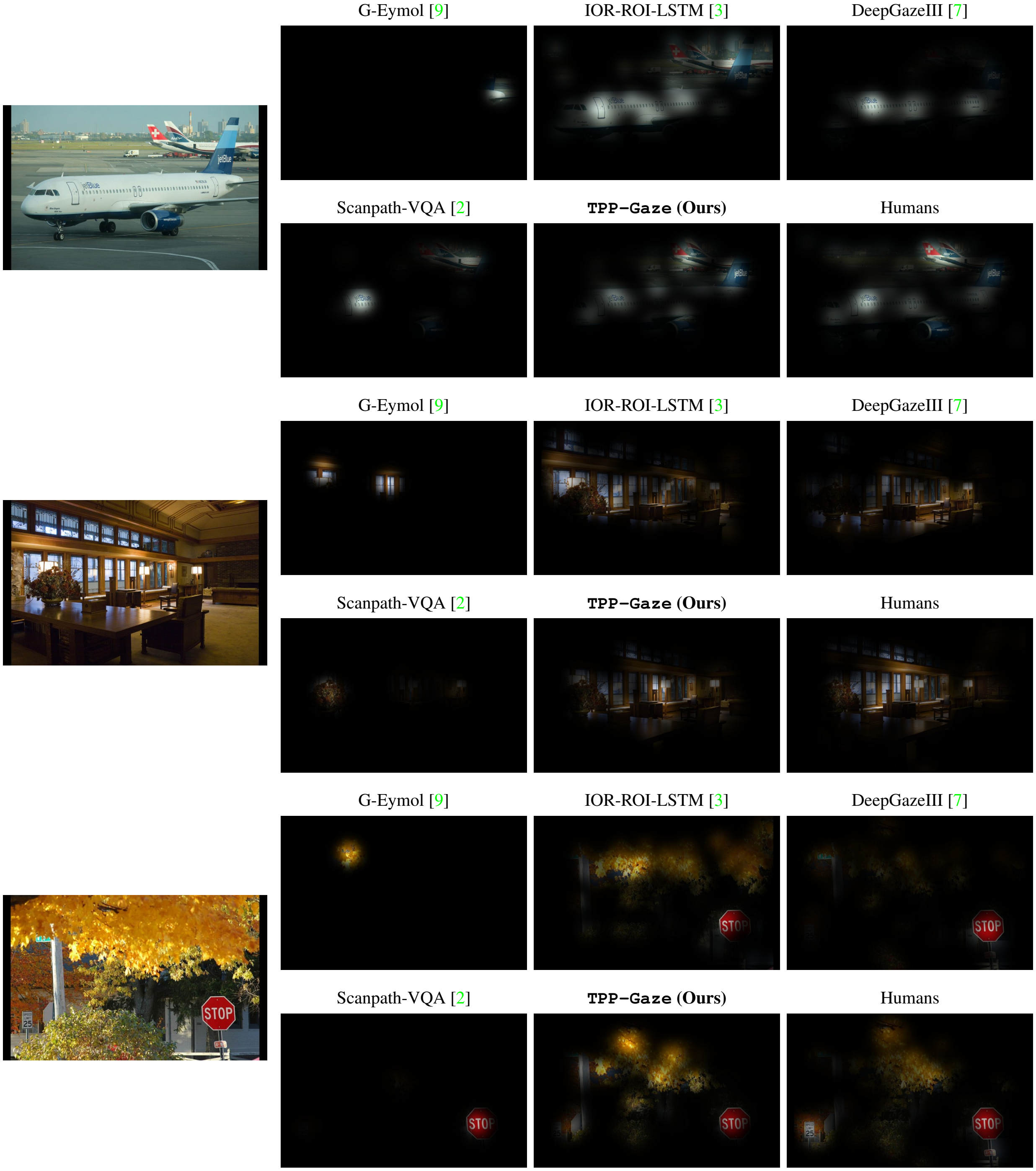


Figure 16. Saliency maps of sample images from COCO-FreeView dataset computed from the fixations generated by the considered scanpath models. For completeness, we include DeepGazeIII, but note that its training procedure also involves saliency prediction.

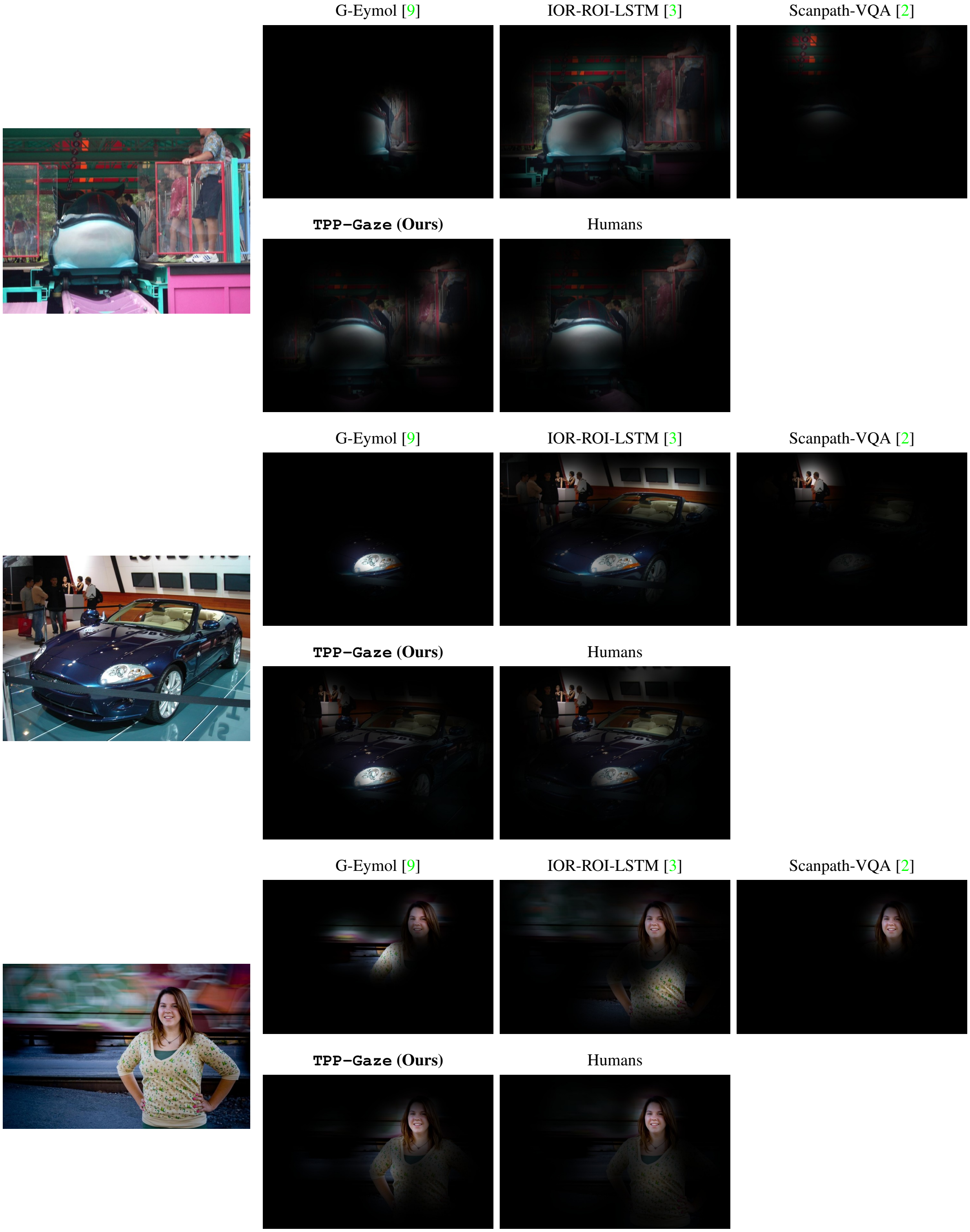


Figure 17. Saliency maps of sample images from MIT1003 dataset computed from the fixations generated by the considered scanpath models. We omit DeepGazeIII for consistency with the experimental settings described in the main paper.

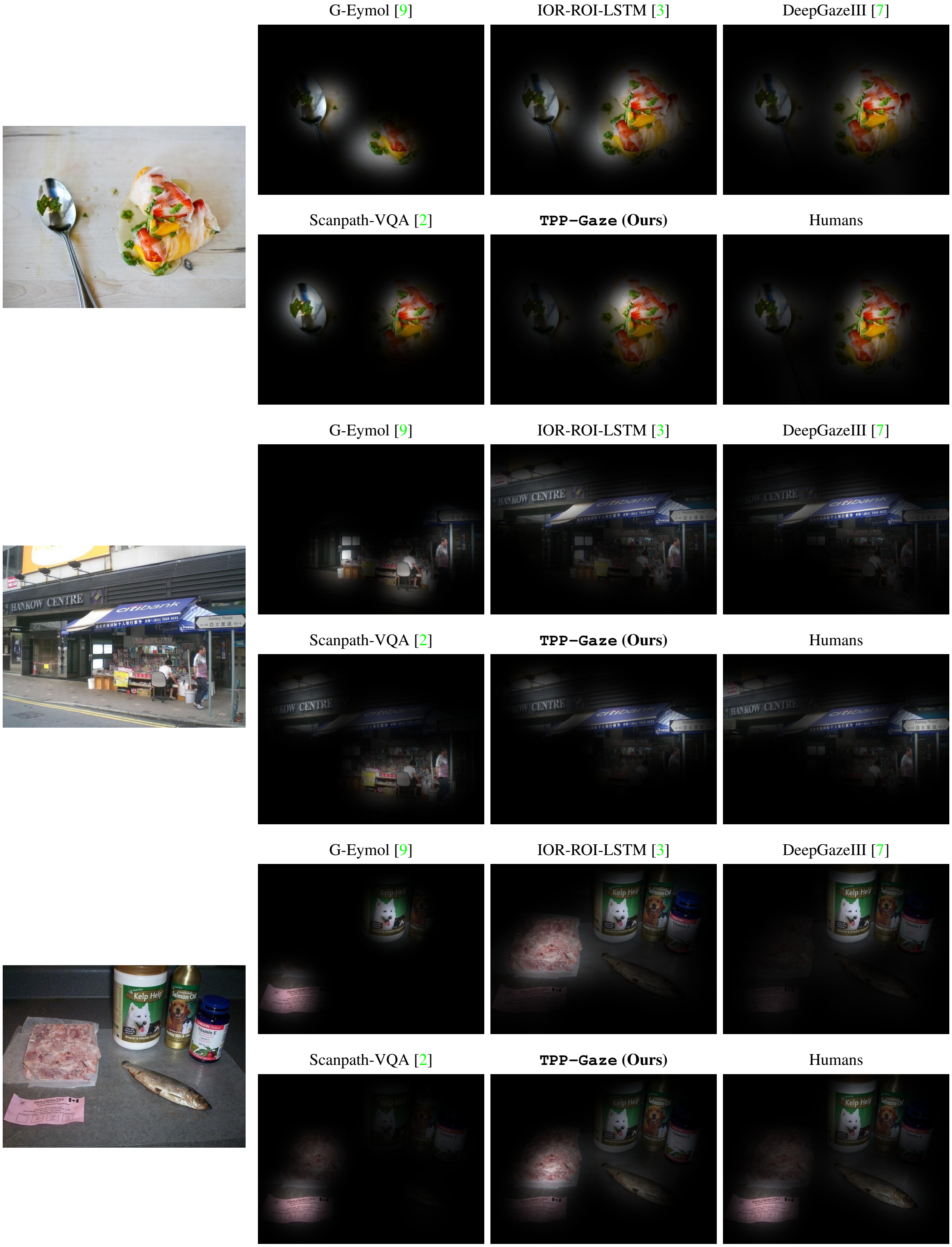


Figure 18. Saliency maps of sample images from OSIE dataset computed from the fixations generated by the considered scanpath models. For completeness, we include DeepGazeIII, but note that its training procedure also involves saliency prediction.

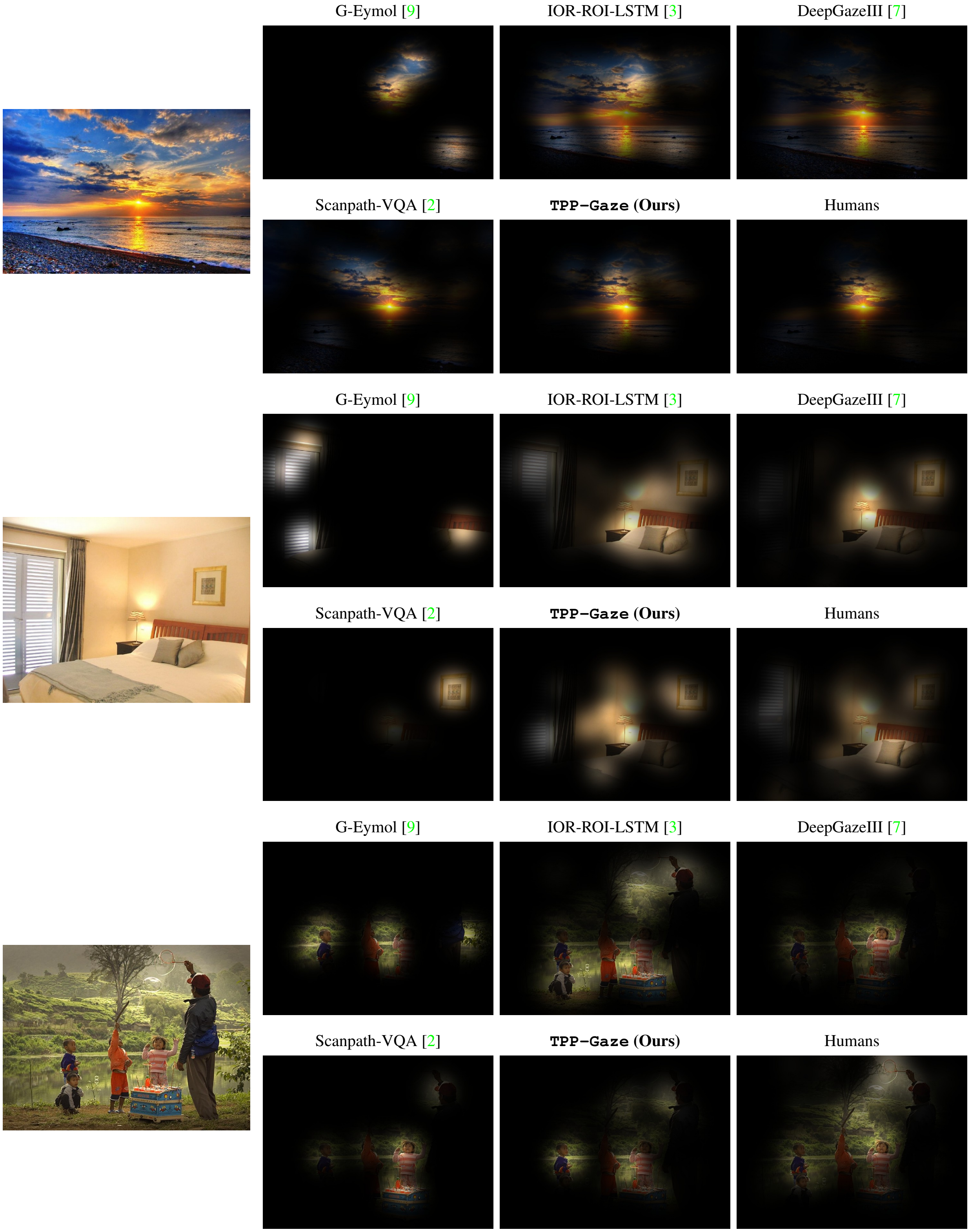


Figure 19. Saliency maps of sample images from NUSEF dataset computed from the fixations generated by the considered scanpath models. For completeness, we include DeepGazeIII, but note that its training procedure also involves saliency prediction.

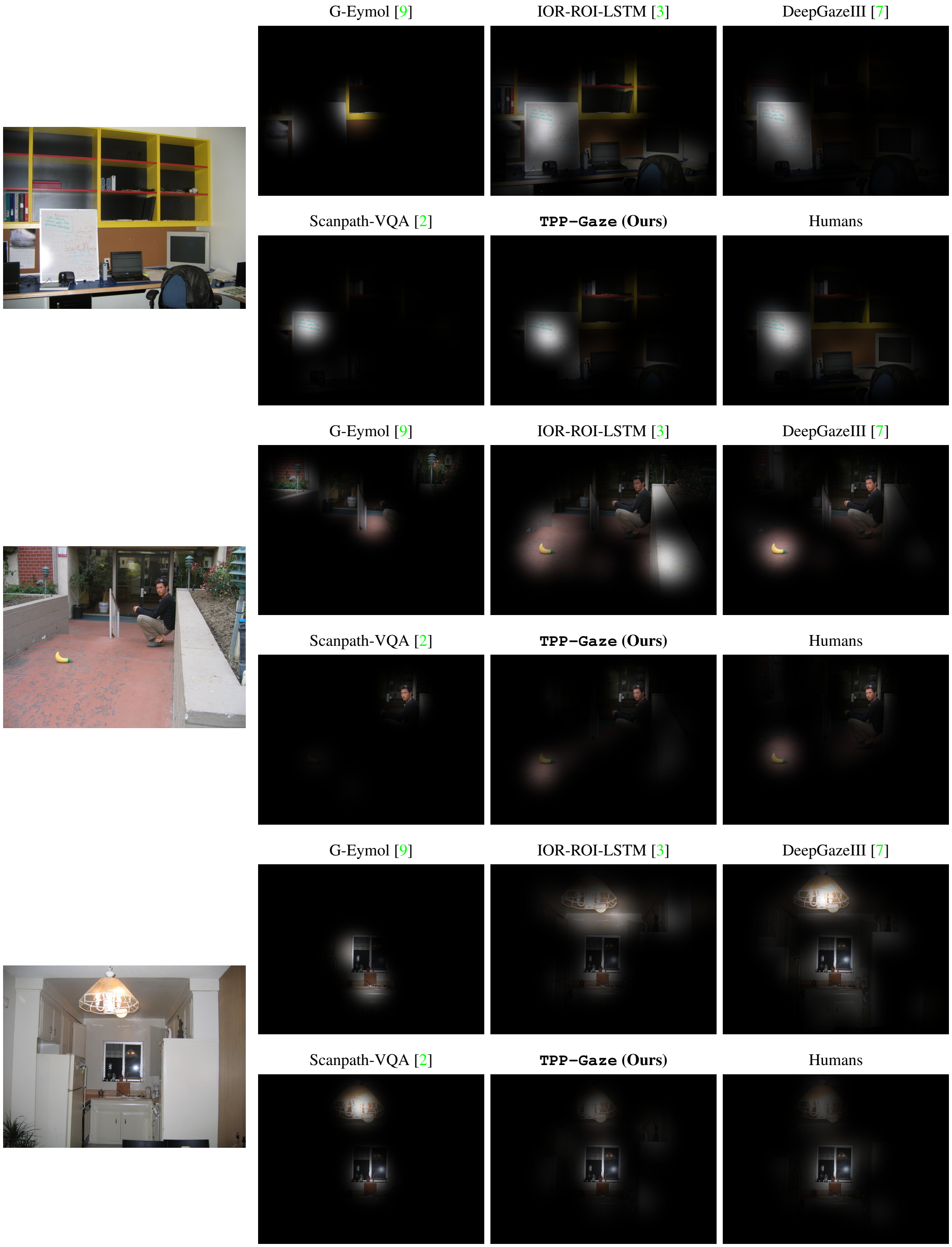


Figure 20. Saliency maps of sample images from FiFa dataset computed from the fixations generated by the considered scanpath models. For completeness, we include DeepGazeIII, but note that its training procedure also involves saliency prediction.

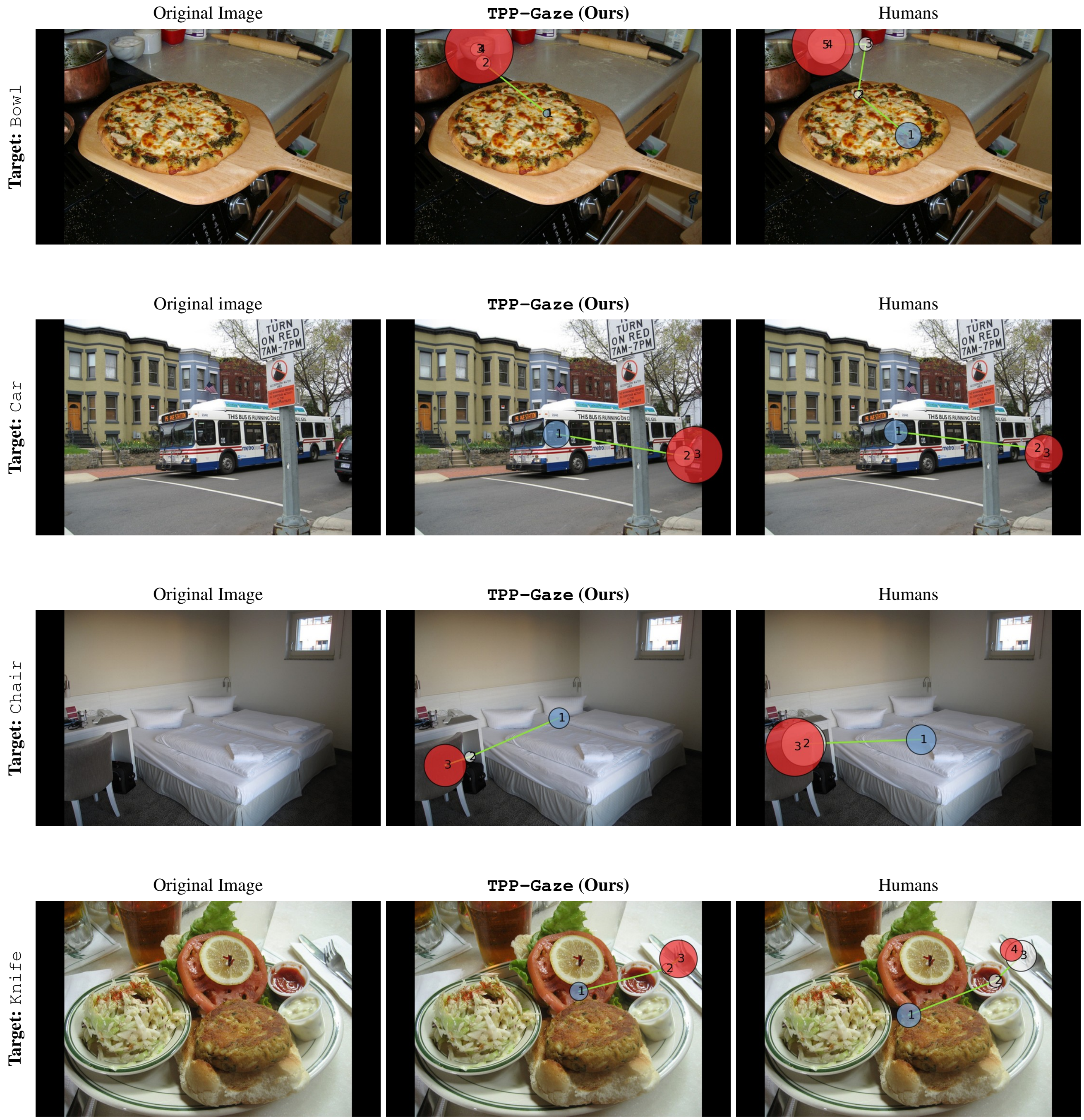


Figure 21. Qualitative comparison of simulated and human scanpaths on the COCO-Search18 dataset for the visual search task.