Predicting Human Scanpaths in Visual Question Answering (Supplementary Materials)

Xianyu Chen Ming Jiang Qi Zhao
University of Minnesota
{chen6582, mjiang}@umn.edu, qzhao@cs.umn.edu

1. Introduction

The main paper has introduced the proposed scanpath prediction model that accurately generates human-like visual scanpath in visual question answering, free-viewing, and visual search tasks. The supplementary materials support our main findings with further evidence and report additional implementation details of the proposed method:

1) We present additional results to investigate the effects of hyperparameters, visual encoder backbones, machine attention mechanisms, and more (Section 2). These results suggest that our method is not only generalizable across multiple tasks, but also flexible to work with different visual encoders and task guidance maps. The results also suggest that our predicted scanpaths can fixate task-relevant objects in both VQA and visual search.

2) We present additional qualitative results in comparison with the state-of-the-art scanpath prediction methods. They demonstrate the superior performance of our method on three datasets: AiR [4] (VQA), OSIE [11] (free-viewing) and COCO-Search18 [12] (visual search) datasets (Section 3).

3) We present the detailed design of our network and how it is adapted to predict scanpaths in different tasks (Section 4).

2. Supplementary Results and Analyses

2.1. Ablation Studies of Hyperparameters

As introduced in the main paper, our learning objective consists of two hyperparameters: λ and γ. In supervised learning, λ balances the cross-entropy loss for the selection of action and the negative logarithmic likelihood estimation of the duration parameters. In reinforcement learning, γ determines the contribution of the Consistency-Divergence loss (CDL). Since λ and γ balance the loss terms in different stages, they can be optimized separately: Tab. 1 and Tab. 2 show the model performances under different settings on the AiR validation set. The best hyperparameters are determined by the harmonic mean of the four ScanMatch [5] scores (*i.e.* with or without duration, correct or incorrect).

We first investigate the effects of λ that balances the cross-entropy loss for the selection of action and the negative logarithmic likelihood estimation of the duration parameters. On the one hand, with a small λ, the objective function puts a lower weight on the prediction of fixation duration. In this case (*i.e.* $\lambda = 0.5$), it leads to sub-optimal results of the duration prediction (see ScanMatch w/ Dur. and MultiMatch-Duration in Tab. 1). On the other hand, with a large λ, the objective function reduces the relative weight of the cross-entropy loss for the selection of action. In this case (*i.e.* $\lambda = 2.0$ or $\lambda = 5.0$), the scores of the other evaluation metrics would drop significantly. Setting $\lambda = 1.0$ results in a reasonable trade-off between the learning of fixation sequence and durations.

Next, we study the effect of γ that determines the contributions of the CDL in the reinforcement learning. As shown in Tab. 2, when $\gamma = 2.0$, our method achieves the best results and its performance on the prediction of incorrect scanpaths is maximized. In contrast, when γ is too small or too large, our method either cannot differentiate the correct scanpaths from the incorrect ones, or does not gain sufficient performance improvement based on the self-critical sequence training (SCST).

Based on these ablation studies, we use $\lambda = 1.0$ and $\gamma = 2.0$ for all the experiments in the main paper and the

λ	ScanMatch ↑	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	Harmonic Mean	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
0.5	0.304	0.294	0.326	0.926	0.714	0.915	0.844	0.527	8.262	6.472	0.841	0.868
		0.283	0.317	0.919	0.718	0.908	0.837	0.556	8.762	7.573	0.828	0.849
1.0	**0.306**	0.290	0.324	0.925	0.713	0.914	0.842	0.532	8.283	6.457	0.839	0.866
		0.290	0.321	0.920	0.717	0.910	0.840	0.557	8.737	7.533	0.830	0.851
2.0	0.292	0.276	0.308	0.924	0.707	0.912	0.834	0.531	8.533	6.769	0.832	0.859
		0.277	0.309	0.919	0.716	0.909	0.828	0.561	8.944	7.741	0.822	0.846
5.0	0.291	0.276	0.309	0.922	0.710	0.910	0.832	0.533	8.574	6.833	0.829	0.857
		0.277	0.305	0.919	0.711	0.909	0.825	0.563	9.088	7.901	0.819	0.843

Table 1. Ablation study of different values of hyperparameter λ on the AiR dataset. We select the best hyperparameter based on the harmonic mean of the four ScanMatch scores. Best results are highlighted in bold.

γ	ScanMatch ↑	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	Harmonic Mean	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
0.5	0.365	0.377	0.373	0.948	0.709	0.931	0.868	0.618	7.637	5.790	0.861	0.893
		0.352	0.358	0.942	0.709	0.925	0.861	0.642	7.884	6.437	0.851	0.880
1.0	0.367	0.372	0.370	0.947	0.705	0.931	0.866	0.619	7.693	5.853	0.857	0.891
		0.360	0.366	0.943	0.716	0.926	0.865	0.643	7.925	6.600	0.849	0.875
2.0	**0.370**	0.375	0.373	0.947	0.708	0.930	0.865	0.615	7.687	5.832	0.857	0.890
		0.363	0.369	0.943	0.713	0.925	0.867	0.640	7.881	6.589	0.853	0.876
5.0	0.367	0.373	0.370	0.948	0.704	0.930	0.867	0.611	7.615	5.771	0.859	0.891
		0.359	0.364	0.943	0.709	0.926	0.863	0.613	7.912	6.547	0.846	0.874
10.0	0.367	0.377	0.372	0.947	0.710	0.930	0.867	0.613	7.657	5.790	0.858	0.890
		0.356	0.363	0.942	0.714	0.925	0.862	0.638	7.863	6.463	0.848	0.877

Table 2. Ablation study of different values of hyperparameter γ on the AiR dataset. We select the best hyperparameter based on the harmonic mean of the four ScanMatch scores. Best results are highlighted in bold.

following analyses.

2.2. Ablation Studies of Visual Encoder Backbones

To demonstrate the generalizability of our network architecture, we compare the performances of our method based on different visual encoder backbones (*i.e.* VGG-16 [9] and ResNet-50 [6]). We conduct comparative experiments on the three datasets: AiR [4], OSIE [11] and COCO-Search [12] datasets. The experimental results are reported in Tab. 3, Tab. 4 and Tab. 5, respectively.

Overall, ResNet-50 results in a better performance than the VGG-16 backbone. Specifically, on the AiR dataset [4], the difference between ResNet-50 and VGG-16 is larger when applying SCST and CDL compared to that in the supervised learning. This difference may arise from the better representation of the feature extracted from ResNet-50 which can be further improved by the SCST and CDL. Similarly, on the OSIE dataset [11] and COCO-Search18 dataset [12], SCST also shows significant performance improvements with both ResNet-50 and VGG-16 backbones, which suggests the generalizability of our method under different backbones and vision tasks. In sum, our approach can generalize to different visual encoder backbones on three different datasets (AiR [4], OSIE [11] and COCO-Search18 [12]) and achieve consistent performances in all of the experiments.

2.3. Ablation Studies of Machine Attention Mechanisms

To verify the effectiveness of different machine attention mechanisms for providing task guidance, we conduct an ablation study using four different machine attention mechanisms [1, 4, 7, 8] for scanpath prediction. The specific implementations of these attention mechanisms follow Chen *et al*. [4]. As shown in Tab. 6, in general, a better machine attention mechanism is more helpful in the guidance of the scanpath prediction. The accuracy of the AiR [4] attention achieves the top performance on most of the evaluation metrics due to its specific attention supervision based on the ground-truth object annotations and the reasoning process. UpDown [1] achieves an acceptable performance because of its use of implicitly supervised object-based attention.

Method	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
VGG-16 [9]	0.289	0.322	0.924	0.716	0.910	0.844	0.539	8.608	6.937	0.838	0.857
	0.275	0.307	0.918	**0.717**	0.905	0.828	0.550	9.015	7.933	0.824	0.843
ResNet-50 [6]	0.296	0.329	0.927	**0.719**	0.914	0.849	0.533	8.438	6.733	0.841	0.862
	0.288	0.317	0.922	**0.717**	0.910	0.837	0.546	8.749	7.682	0.831	0.850
VGG-16* [9]	0.360	0.365	0.948	0.709	**0.934**	0.865	0.592	7.809	5.937	0.860	0.886
	0.347	0.353	0.942	0.707	0.928	0.854	0.615	8.151	7.060	0.848	0.868
ResNet-50* [6]	**0.394**	**0.391**	**0.950**	0.717	0.933	**0.879**	**0.615**	**7.523**	**5.701**	**0.869**	**0.893**
	0.365	**0.368**	**0.946**	0.705	**0.930**	**0.864**	**0.632**	**7.955**	**6.772**	**0.856**	**0.877**

Table 3. Ablation study of different visual encoder backbones on the AiR dataset. Asterisks indicate the application of SCST and CDL. In each sub-panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths. Best results are highlighted in bold.

Method	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
VGG-16 [9]	0.340	0.320	0.910	0.668	0.888	0.821	0.667	7.799	5.626	0.825	0.865
ResNet-50 [6]	0.349	0.329	0.913	**0.669**	0.892	0.830	0.652	7.890	5.709	0.830	0.870
VGG-16* [9]	0.377	0.370	0.937	0.657	0.918	0.838	0.669	7.326	4.989	0.847	0.898
ResNet-50* [6]	**0.383**	**0.377**	**0.943**	0.651	**0.924**	**0.847**	**0.684**	**7.155**	**4.579**	**0.852**	**0.905**

Table 4. Ablation study of different visual encoder backbones on the OSIE datasaet. Asterisks indicate the application of SCST. Best results are highlighted in bold.

Method	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
VGG-16 [9]	0.421	0.361	0.924	0.651	0.912	0.887	0.690	2.538	1.127	0.894	0.942
ResNet-50 [6]	0.434	0.372	0.926	0.656	0.912	0.894	0.698	2.433	1.005	0.901	0.947
VGG-16* [9]	0.511	0.469	0.938	**0.709**	0.924	0.905	0.712	2.036	0.669	0.914	0.957
ResNet-50* [6]	**0.554**	**0.510**	**0.941**	0.706	**0.927**	**0.914**	**0.721**	**1.852**	**0.484**	**0.923**	**0.965**

Table 5. Ablation study of different visual encoder backbones on the COCO-Search18 dataset. Asterisks indicate the application of SCST. Best results are highlighted in bold.

The low performances achieved by HAN [8] and MLB [7] may arise from the misalignment of the attention ground-truth from a specific group of questions and spatial attention instead of the object-wise attention. Fig. 1 shows a qualitative comparison of using different machine attention mechanisms for the task guidance [1, 4, 7, 8]. It can be seen that only the AiR [4] attention map can precisely highlight the relevant object (*i.e.* phone). Therefore, with AiR [4], the predicted scanpaths corresponding to the correct answer can successfully fixate the phone, while the others [1, 7, 8] fail.

2.4. Supplementary Results on Fixated Regions

To demonstrate the effectiveness of each proposed technique in localizing task-relevant objects, we extend the Tab 3 in our main paper with ablation studies of the proposed techniques *i.e.* task guidance (TG), SCST, CDL. As shown in Tab. 7, this analysis computes the percentage of fixations in each type of regions (*i.e.* region of interest (ROI), non-ROI, and background). We can observe that our baseline (*i.e.* a task-ignorant supervised-learning variant of our method) obtains a significantly better performance than other state-of-the-art approaches [2, 3, 10] by placing more fixations inside the task-relevant ROIs, which demonstrates the effectiveness of our network design and objective function. Further, as the proposed techniques add up, the per-

VQA Model	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
MLB [7]	0.369	0.367	0.949	0.714	0.932	0.866	0.602	7.822	6.025	0.858	0.882
	0.350	0.349	0.942	0.716	0.925	0.853	0.631	8.120	7.030	0.845	0.866
HAN [8]	0.366	0.367	0.949	0.707	0.932	0.866	**0.618**	7.718	5.970	0.858	0.881
	0.353	0.352	0.943	0.711	0.926	0.853	0.618	7.885	**6.632**	0.848	0.870
UpDown [1]	0.382	0.376	0.949	**0.717**	**0.933**	0.872	0.593	7.560	**5.638**	0.866	0.887
	0.353	0.360	0.942	**0.717**	0.928	0.852	0.543	**7.879**	6.838	0.843	0.863
AiR [4]	**0.394**	**0.391**	**0.950**	**0.717**	**0.933**	**0.879**	0.615	**7.523**	5.701	**0.869**	**0.893**
	0.365	**0.368**	**0.946**	0.705	**0.930**	**0.864**	**0.632**	7.955	6.772	**0.856**	**0.877**

Table 6. Ablation study of different VQA models (*i.e.* MLB [7], HAN [8], UpDown [1] and AiR [4]) on the AiR dataset. In each panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths. Best results are highlighted in bold.

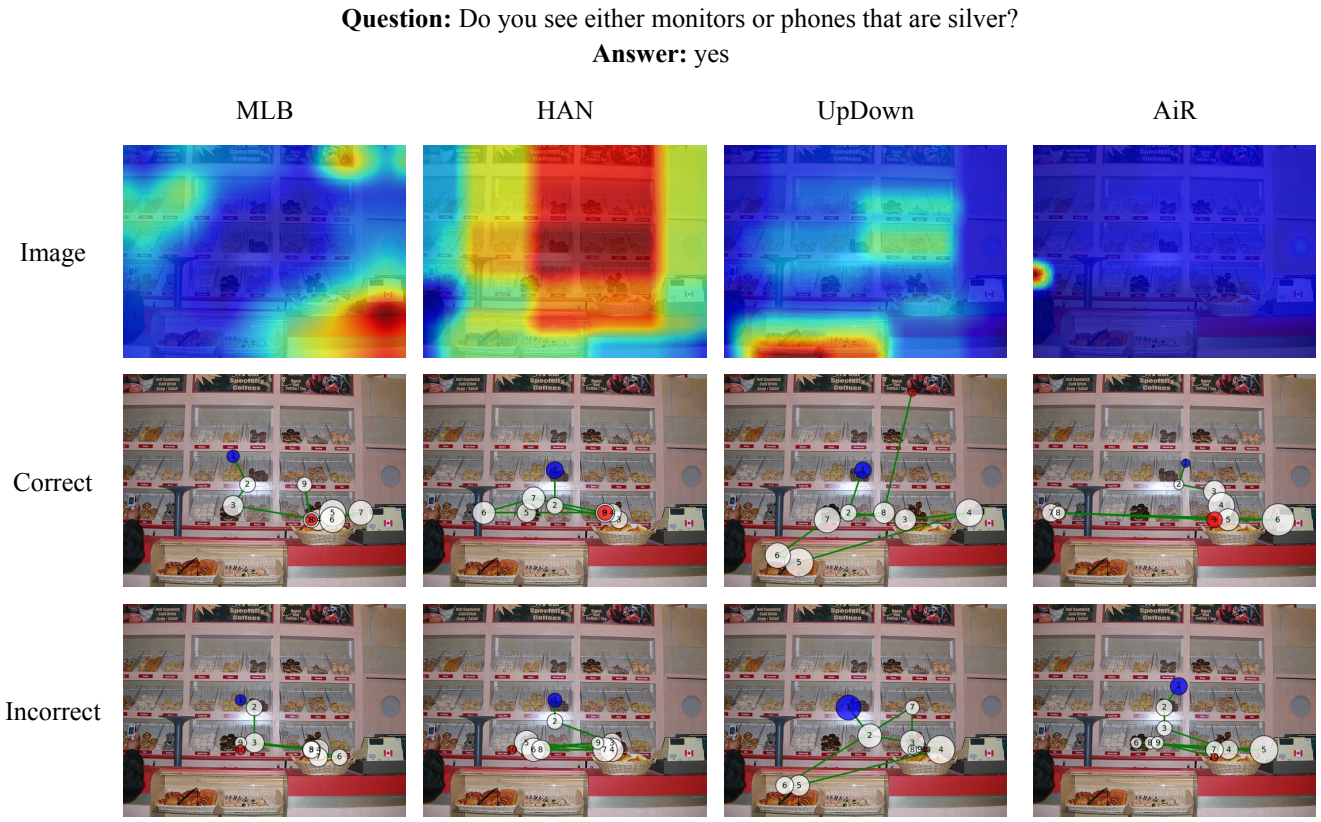

Figure 1. Qualitative comparison of different VQA models [1, 4, 7, 8] (*i.e.* MLB [7], HAN [8], UpDown [1] and AiR [4]). The first row presents the attention maps overlaid on the image. The second and third rows demonstrate the predicted scanpaths corresponding to correct and incorrect answers, respectively.

centage of fixations in ROIs gradually increases. This increasing trend agrees with the scanpath evaluation metrics, suggesting that better scanpath prediction models not only align with human eye movements better, they also fixate more important objects for answering the questions.

In terms of comparing across different attention mechanisms, we also observe a similar trend that higher-performance models generate more fixations in the ROIs. As shown in Tab. 8, the most accurate machine attention maps obtained from AiR [4] allow our model to achieve the performance closest to that of humans. The ranks of these VQA models are consistent with those of the scanpath pre-

Method			Fixations Position %		
TG	SCST	CDL	ROI ↑	Non-ROI ↓	Background ↓
Human			26.43 21.60	67.48 71.92	6.09 6.48
SaltiNet [3]			4.17 3.96	77.88 78.49	17.95 17.55
PathGAN [2]			7.82 7.17	84.34 86.10	7.83 6.73
IOR-ROI [10]			9.14 9.79	82.99 82.53	7.87 7.67
			16.15 15.08	77.04 78.89	6.81 6.04
✓			17.75 17.12	75.61 76.43	6.65 6.45
	✓		21.68 19.79	73.45 74.97	4.87 5.24
	✓	✓	19.77 19.41	74.27 75.32	5.96 5.28
✓	✓		25.63 21.99	69.96 73.34	4.41 4.67
✓	✓	✓	25.04 22.33	69.70 72.27	5.26 5.40

Table 7. Percentage of fixations in ROI, non-ROI, and background. In each panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths.

VQA Models	Fixations %		
	ROI ↑	Non-ROI ↓	Background ↓
Human	26.43 21.60	67.48 71.92	6.09 6.48
MLB [7]	20.57 21.32	73.75 73.18	5.67 5.50
HAN [8]	21.87 20.54	73.99 75.72	4.14 3.74
UpDown [1]	25.01 22.36	70.61 72.20	4.39 5.44
AiR [4]	25.04 22.33	69.70 72.27	5.26 5.40

Table 8. Percentage of fixations in ROI, non-ROI, and background for different VQA models. In each panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths.

diction performance shown in Tab. 6.

2.5. Ablation Studies on OSIE and COCO-Search18 Datasets

We further present additional ablation studies of our proposed method on the OSIE [11] and COCO-Search18 [12] datasets. For the free-viewing task, since TG and CDL are not applicable, this experiment compares our proposed method with the supervised baseline without SCST. As shown in Tab. 9, SCST can significantly improve the performance on the OSIE dataset [11]. Similarly, for the visual search task, where TG and SCST are applicable but not the CDL, we can also observe the significant impact of SCST on the model's performance (see Tab. 10). These observations confirm the finding on the AiR dataset that SCST can help the model to generate scanpaths that are more consistent with that from human. Furthermore, Tab. 10 also suggests that TG plays an important role in guiding the fixation to the final targets, thus further increasing the performance.

2.6. Evaluating Visual Search Performances on COCO-Search18 with Additional Metrics

To demonstrate the performance of our predicted scanpaths in visual search, we evaluate the scanpaths with the search targets based on the three evaluation metrics (*i.e.* Target Fixation Probability AUC (TFP-AUC), Probability Mismatch and Scanpath Ratio) provided by the COCO-Search18 dataset [12]. Specifically, the **TFP-AUC** describes the effectiveness of the target searching process. It measures the area under the *target fixation curve* that shows the cumulative probability corresponding to the number of fixations made to target. The **Probability Mismatch** metric refers to the sum of the absolute differences of cumulative probabilities of target fixation. It is used to describe the discrepancy of search patterns between the human scanpaths and the predicted scanpaths. The **Scanpath Ratio** is obtained by the ratio of Euclidean distance between the initial fixation location and the center of the target to the length of the scanpath, which is used to measure the search efficiency. In Tab. 11, we compare our method with the state-of-the-art methods based on these metrics. As can be seen, consistent with human scanpaths, our method outperforms the state-of-the-art approaches [2, 3, 10, 12] by a large margin in all the three metrics, which demonstrates the high efficiency of our method in localizing the search targets.

Method	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
Human	0.390	0.386	0.941	0.695	0.931	0.851	0.621	7.486	5.001	0.844	0.906
Ours w/o SCST	0.349	0.329	0.913	**0.669**	0.892	0.830	0.652	7.890	5.709	0.830	0.870
Ours	**0.383**	**0.377**	**0.943**	0.651	**0.924**	**0.847**	**0.684**	**7.155**	**4.579**	**0.852**	**0.905**

Table 9. Ablation study of SCST on the OSIE dataset. The best results are highlighted in bold. Underlines indicate scores above the human performance.

Method		ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
TG	SCST	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
Human		0.526	0.490	0.944	0.755	0.934	0.913	0.685	2.181	0.359	0.920	0.974
		0.430	0.368	0.924	0.658	0.910	0.893	0.699	2.441	1.023	0.899	0.946
✓		0.434	0.372	0.926	0.656	0.912	0.894	0.698	2.433	1.005	0.901	0.947
	✓	0.552	0.509	0.939	0.699	0.926	**0.914**	0.710	1.861	**0.484**	0.922	0.964
✓	✓	**0.554**	**0.510**	**0.941**	**0.706**	**0.927**	**0.914**	**0.721**	**1.852**	**0.484**	**0.923**	**0.965**

Table 10. Ablation study of TG and SCST on the COCO-Search18 dataset. The best results are highlighted in bold. Underlines indicate scores above the human performance.

Method	TFP-AUC ↑	Probability Mismatch ↓	Scanpath Ratio ↑
Human	5.161	-	0.852
SaltiNet [3]	0.527	4.634	0.665
PathGAN [2]	0.366	4.794	0.926
IOR-ROI [10]	1.690	3.471	0.502
IRL [12]	4.558	0.964	0.853
Ours	**4.785**	**0.809**	**0.948**

Table 11. Additional quantitative results on the COCO-Search18 dataset. The best results are highlighted in bold. Underlines indicate scores above human performance. This table is complementary to the Tab 5 in the main paper.

3. Additional Qualitative Results

This section presents additional qualitative results of our method in comparison with the state-of-the-art scanpath prediction models and humans. These qualitative results consist of three different tasks and experimental settings: VQA, free-viewing, and visual search.

Fig. 2–Fig. 4 present qualitative examples of the predicted correct and incorrect scanpaths on the AiR dataset [4] under the VQA task. Note that subtle differences of scanpaths can determine the correctness of answers: the incorrect scanpaths consistently miss important objects (*i.e.* phone, dog and knives). While the state-of-the-art scanpath prediction models look at salient objects in general, our predicted scanpaths align better with task-related objects and the human eye-movement behavior regarding fixation positions, durations, and orders.

Fig. 5–Fig. 8 present qualitative examples of the predicted scanpaths on the OSIE dataset [11] under the free-viewing task. While the state-of-the-art scanpath prediction models look at some of the salient objects, our predicted scanpaths are almost indistinguishable from the human eye-movement behavior.

Fig. 9–Fig. 12 present qualitative examples of the predicted scanpaths on the COCO-Search18 dataset [12] under the visual search task. As can be seen, most of the state-of-the-art scanpath prediction models fail to look at the search targets. Differently, our predicted scanpaths always successfully and efficiently find the targets (*i.e.* stop sign, cup, oven, and fork) with only 3 fixations.

Question: What is the device on top of the nightstand made of wood?
Answer: phone

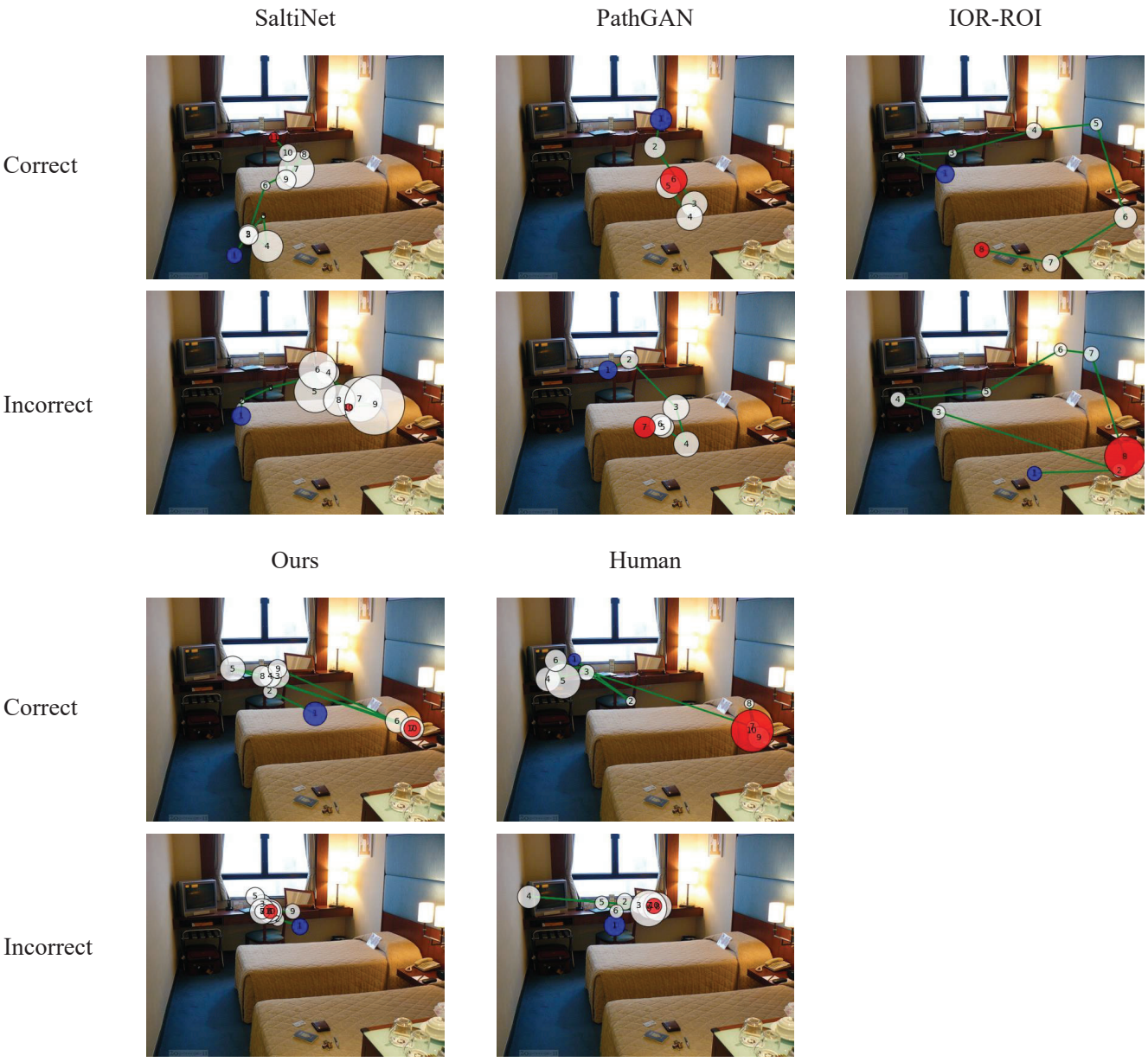

Figure 2. Qualitative example on the AiR dataset.

Question: Is the blue bike to the right or to the left of the dog on the left?
Answer: left

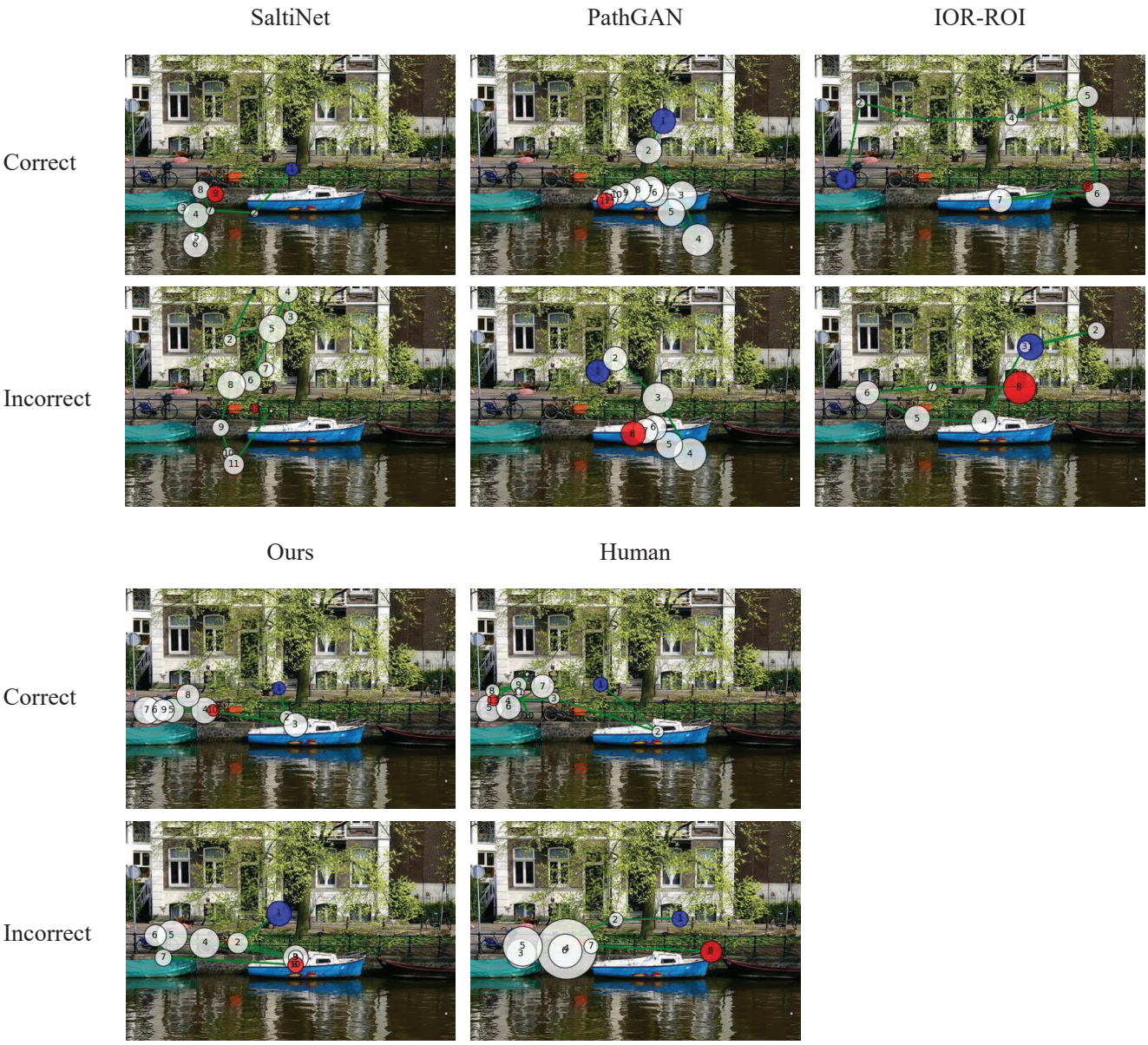

Figure 3. Qualitative example on the AiR dataset.

Question: Are there both knives and spoons in the picture?
Answer: yes

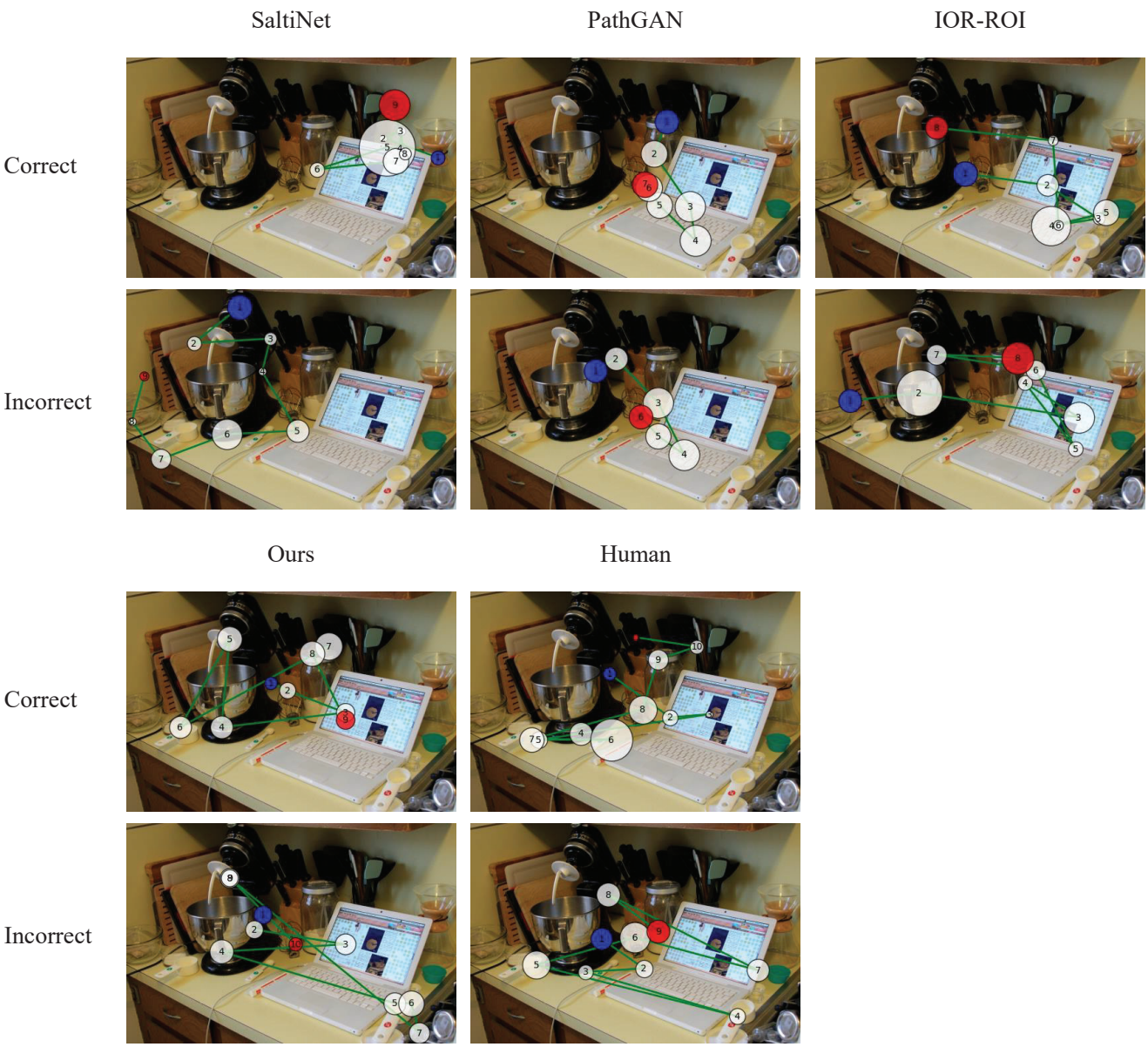

Figure 4. Qualitative example on the AiR dataset.

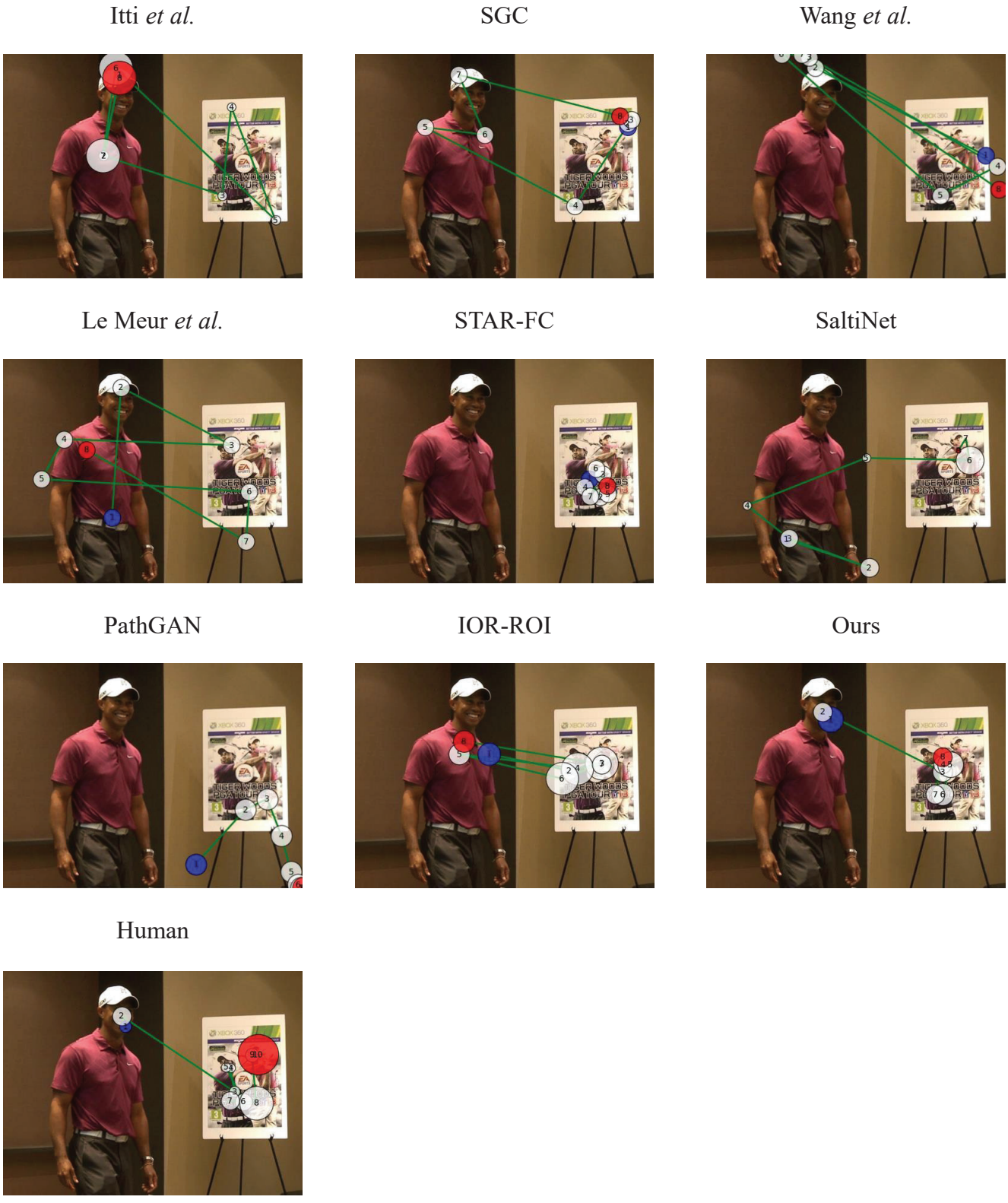

Figure 5. Qualitative example on the OSIE dataset.

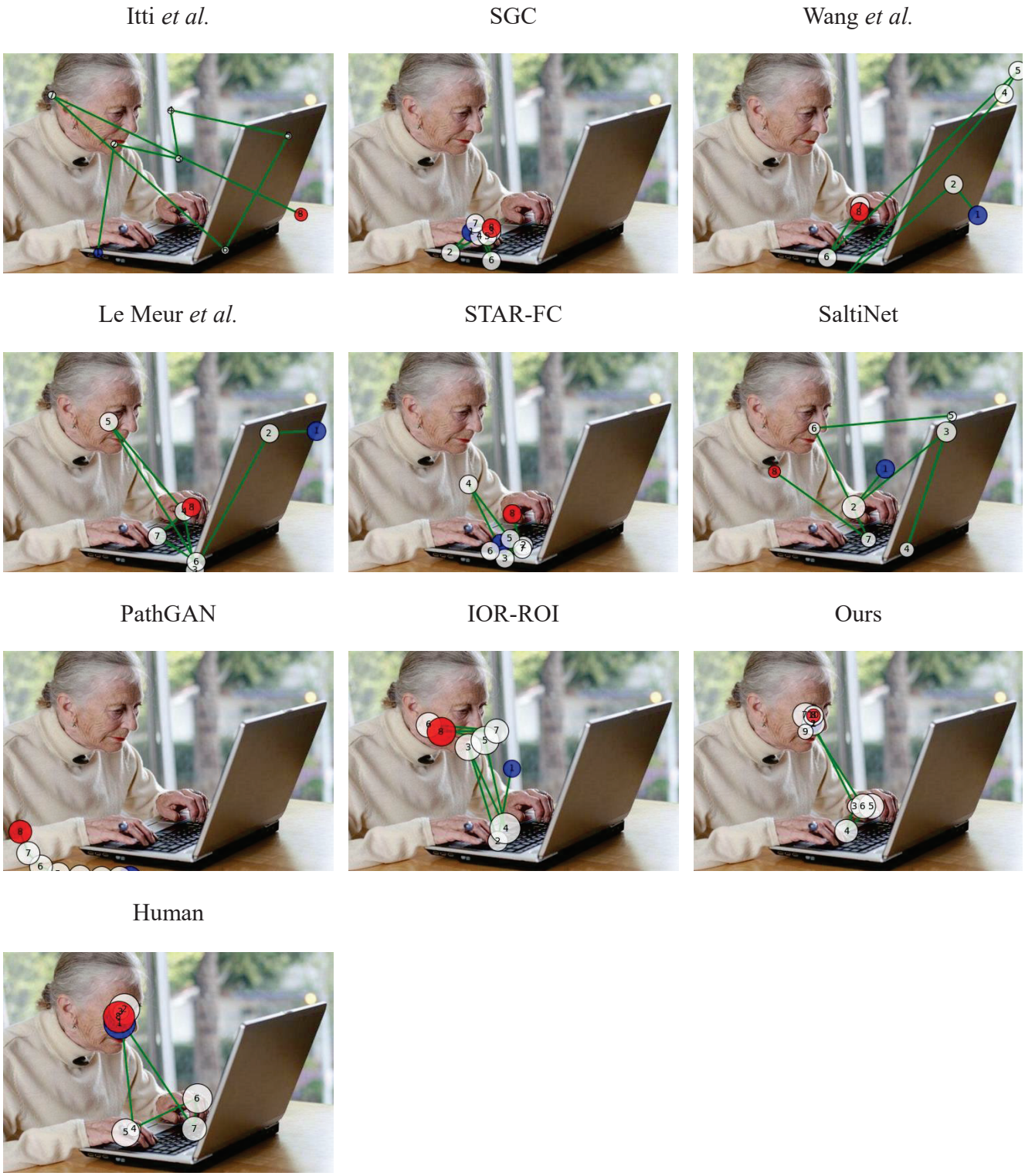

Figure 6. Qualitative example on the OSIE dataset.

Figure 7. Qualitative example on the OSIE dataset.

Figure 8. Qualitative example on the OSIE dataset.

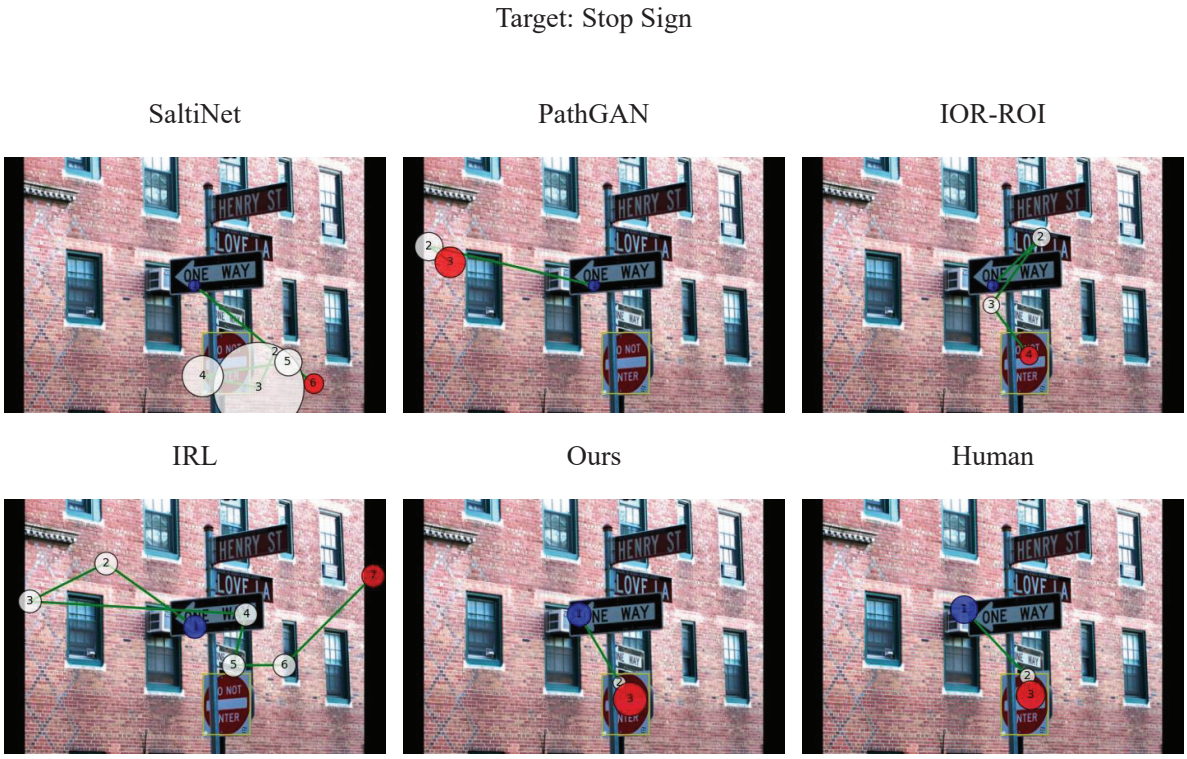

Figure 9. Qualitative example on the COCO-Search18 dataset.

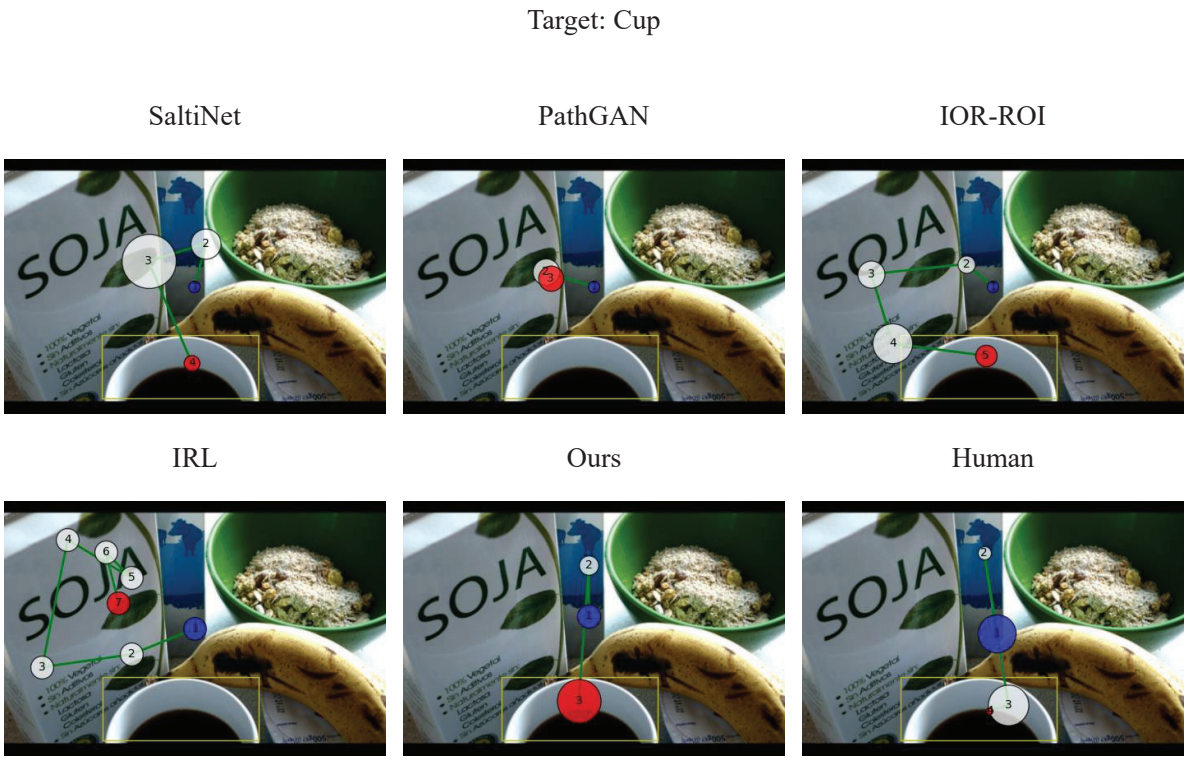

Figure 10. Qualitative example on the COCO-Search18 dataset.

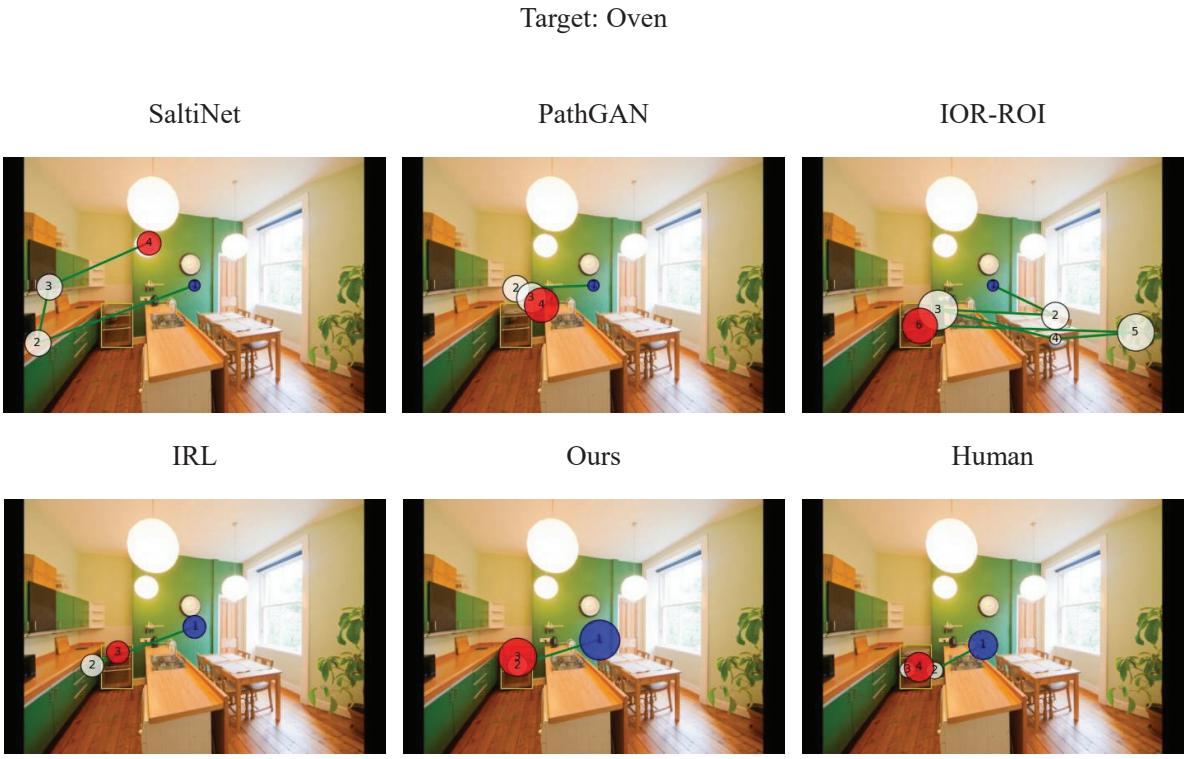

Figure 11. Qualitative example on the COCO-Search18 dataset.

Figure 12. Qualitative example on the COCO-Search18 dataset.

4. Supplementary Methods

In this section, we elaborate the detailed design of our network, as well as the specific implementation details to adapt this network for free-viewing and visual search tasks.

4.1. Network Architecture

Attention Mechanism. First, we elaborate the detailed design of the selective attention mechanism f_{att} shown in Fig. 2 of the main paper. Given the memorized features $X_t = M_t \circ X$ and the most recent features in memory $x_{t-1} = m_{t-1} \circ X$, where $\circ$ indicates the Hadamard product, the attention mechanism is designed to recall the most relevant spatial information $u_t^s \in \mathbb{R}^{h \times w}$ and channel information $u_t^c \in \mathbb{R}^d$ from the memory of M_t.

To recall the most relevant spatial information u_t^s at time t, we compute $h_s(X_t)$ and $h_s(x_{t-1})$ with aggregating the feature map in channel dimension to get the spatial semantic followed by a linear layer and a ReLU activation function to further encode the memorized features. We formulate function $h_s(X_t)$ as

$$h_s(X_t) = \text{ReLU}(W_{hs}\text{AvgPool}(X_t) + b_{hs}), \tag{1}$$

where W_{hs} and b_{hs} are learnable parameters. Hence the selection of the spatial feature is obtained as

$$\alpha_t = W_\alpha\big(W_S h_s(X_t) + W_s h_s(x_{t-1}) + b_s\big), \tag{2}$$
$$u_t^s = \alpha_t h_s(X_t). \tag{3}$$

The attention α_t helps to determine the weights of the previous action maps in the prediction of the next fixation. Here, W_S, W_s, b_s and W_α are learnable parameters to optimize the attention α_t.

Similarly, to recall the most relevant channel information u_t^c at time t, we compute $h_c(X_t)$ and $h_c(x_{t-1})$ with aggregating the feature map in spatial dimension to get the channel semantic followed by a linear layer and a ReLU activation function to further encode the memorized features. We formulate function $h_c(X_t)$ as

$$h_c(X_t) = \text{ReLU}(W_{hc}\text{AvgPool}(X_t) + b_{hc}), \tag{4}$$

where W_{hc} and b_{hc} are learnable parameters. Hence the selection of the spatial feature is obtained as

$$\beta_t = W_\beta\big(W_c h_c(X_t) + W_c h_c(x_{t-1}) + b_c\big), \tag{5}$$
$$u_t^c = \beta_t h_c(X_t). \tag{6}$$

The attention β_t helps to determine the weights the previous action maps in the prediction of the next fixation. The parameters W_C, W_c, b_c and W_β are also trainable parameters to optimize the attention β_t.

With the computed u_t^s and u_t^c, we can get the recalled features containing both the spatial semantics and channel semantics:

$$R_t = u_t^s \otimes u_t^c, \tag{7}$$

where $\otimes$ represents the outer product. The result $R_t \in \mathbb{R}^{d \times h \times w}$ is sent to the ConvLSTM module to predict the next fixation.

ConvLSTM. The recalled feature R_t and the visual feature $X \in \mathbb{R}^{d \times h \times w}$ are processed with a ConvLSTM to encode the spatio-temporal patterns in the scanpaths. Specifically, they are used to adaptively control the gate functions of the ConvLSTM:

$$\begin{aligned} i_t = {} & W_{x_i}X + W_{h_i}h_{t-1} + W_{c_i}c_{t-1} \\ & + W_{r_i}R_t + b_i, \end{aligned} \tag{8}$$

$$\begin{aligned} f_t = {} & W_{x_f}X + W_{h_f}h_{t-1} + W_{c_f}c_{t-1} \\ & + W_{r_f}R_t + b_f, \end{aligned} \tag{9}$$

$$\begin{aligned} o_t = {} & W_{x_o}X + W_{h_o}h_{t-1} + W_{c_o}c_{t-1} \\ & + W_{r_o}R_t + b_o, \end{aligned} \tag{10}$$

where i_t, f_t, o_t denote the input gate, the forget gate and the output gate, respectively. We use h_{t-1} and c_{t-1} to represent the hidden state and the cell state. The learnable weights of the corresponding gate functions are W_{x_i}, W_{h_i}, W_{c_i}, b_i, W_{x_f}, W_{h_f}, W_{c_f}, b_f, W_{x_o}, W_{h_o}, W_{c_o} and b_o, while the learnable weights of the recalled features are W_{r_i}, W_{r_f} and W_{r_o}.

Output Layers. The output of the ConvLSTM is the hidden state h_t that encodes the spatio-temporal information of the scanpaths. We further encode this hidden state h_t with a convolutional layer to obtain the features

$$F_t = \text{ReLU}(W_g h_t + b_g), \tag{11}$$

where W_g and b_g are both the learnable parameters.

Finally, we define two output functions: $f_a(\cdot;\theta_a)$ generates the logit scores of the action maps and end-of-scanpath indicator (m_t, e_t), and $f_\tau(\cdot;\theta_\tau)$ generates the parameters of the fixation duration $[\mu_t, \sigma_t^2]$.

Specifically, for $f_a(\cdot;\theta_a)$, the action maps m_t are generated by a convolutional layer followed a ReLU activation function and the end-of-scanpath indicator e_t is obtained by a convolutional layer followed by a global average pooling:

$$m_t = \text{ReLU}(W_{f_a^1}F_t + b_{f_a^1}), \tag{12}$$
$$e_t = \text{AvgPool}(W_{f_a^2}F_t + b_{f_a^2}), \tag{13}$$

where $W_{f_a^1}$, $W_{f_a^2}$, $b_{f_a^1}$ and $b_{f_a^2}$ are learnable parameters. We define the overall output as the softmax-normalized concatenation of m_t and e_t:

$$p_t^a(a_t|a_{1:t-1}) = \text{softmax}\Big([m_t, e_t]\Big). \tag{14}$$

ξ	ScanMatch ↑	ScanMatch ↑		MultiMatch ↑					SED ↓		STDE ↑	
	Harmonic Mean	w/o Dur.	w/ Dur.	Vector	Direction	Length	Position	Duration	Mean	Best	Mean	Best
0.7	0.526	0.549	0.505	0.938	0.705	0.923	0.913	0.715	1.885	0.541	0.920	0.963
0.8	**0.531**	0.554	0.510	0.941	0.706	0.927	0.914	0.721	1.852	0.484	0.923	0.965
0.9	0.523	0.546	0.501	0.941	0.704	0.927	0.914	0.722	1.882	0.546	0.922	0.963

Table 12. Ablation study of different values of hyperparameter ξ on the COCO-Search18 dataset. We select the best hyperparameter based on the harmonic mean of the two ScanMatch scores. Best results are highlighted in bold.

Moreover, $f_\tau(\cdot;\theta_\tau)$ consists of a convolutional layer, a ReLU activation function followed another convolutional layer. It can be formulated as

$$[\mu_t, \sigma_t^2] = W_{f_\tau^2}\text{ReLU}(W_{f_\tau^1}h_t + b_{f_\tau^1}) + b_{f_\tau^2}, \qquad (15)$$

where $W_{f_\tau^1}$, $W_{f_\tau^2}$, $b_{f_\tau^1}$ and $b_{f_\tau^2}$ are learnable parameters.

With these prediction outputs, we can randomly sample the fixation positions and its duration based on the probability distributions.

4.2. Adapting Task Guidance for Free-Viewing and Visual Search

Task guidance for the OSIE free-viewing dataset. The proposed method can be directly adapted to predict scanpaths in the free-viewing task, by setting the task guidance map Z as an all-zero matrix. This allows the scanpath prediction to be completely driven by the visual information, which is the same condition as the humans experience in the free-viewing experiment of the OSIE dataset.

Task guidance for the COCO-Search18 visual search dataset. To predict scanpaths in the visual searach task, we use a CenterNet [13] object detector to detect the search targets. The CenterNet detects the bounding boxes of the 18 object classes of the COCO-Search18 datasets, and predicts a classification score for each object. We select the objects with classification scores higher than a threshold ξ, and generate the task guidance map by setting the values inside their bounding boxes as 1 and the rest as 0. The optimal threshold $\xi = 0.8$ is obtained from an ablation study on the COCO-Search18 validation set (see Tab. 12). We train a specific model for each of the 18 object classes. While most of their parameters are shared, their output layers are optimized independently. This design is similar to the prediction of correct and incorrect scanpaths in the VQA task.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] 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 Workshop (ECCVW)*, 2018.

[3] Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto, 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 Workshop (ICCVW)*, 2017.

[4] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. AiR: Attention with reasoning capability. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[5] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods (BRM)*, 2010.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[8] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring human-like attention supervision in visual question answering. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[10] 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 (IEEE TPAMI)*, 2019.

[11] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision (JoV)*, 2014.

[12] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[13] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.