Task-driven Webpage Saliency

Anonymous ECCV submission

Paper ID 1672

1 Comparison with Prior Work

A recent work in [1] proposed a method for predicting webpage saliency under free-viewing condition. Unfortunately, we did not get their code for quantitative comparison on our evaluation dataset. Therefore, we apply our model on their tested webpages, and make visual comparison with the prediction results reported in Fig. 5 (*i.e.,* Ours+Position) in their paper. Fig. 1 shows the comparison. We show the predictions of our model under two tasks, "form filling" and "information browsing", according to the nature of their tested webpages. Compared with [1], while both methods are able to predict visually salient regions on the webpages, our model can predict different human attention behaviors under different tasks. For example, at the fourth row of Fig. 1, while both methods identify the images at the center as salient, our method can allocate some saliency around the input box at the top-right corner under "form filling", and attend more to the images under task "information browsing".

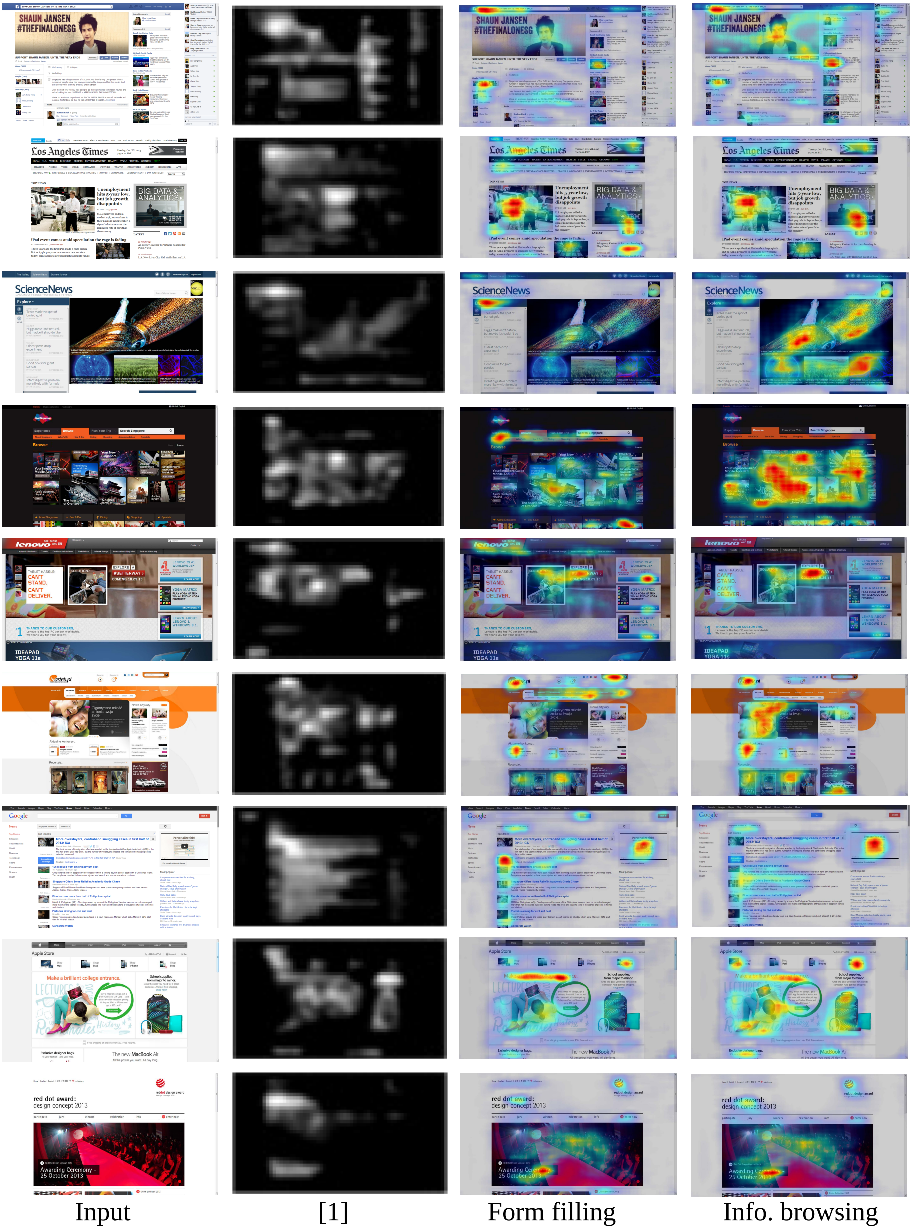


Fig. 1: Qualitative comparison of our results with those of [1] on the tested webpages used in Fig. 5 (*i.e.*, Ours+Position) of [1]. We show the overlay of our predicted saliency maps on the input webpages in the last two columns. Note that [1] predicts the saliency maps under free-viewing condition, while our saliency maps are predicted under task-driven conditions (*i.e.*, "form filling" and "information browsing").

2 Per-task Performance in Ablation Study

We show the per-task performance of the ablated models in various evaluation metrics in Tables 1-3. The best results are highlighted in red, while the second best results are in blue.

Table 1: Per-task performance of the ablated models in KL↓.

method	Signing	Form	Info.	Shopping	Comm.	Average
	up	filling	browsing		joining	
No task-specific subnet	1.3780	1.4338	1.2336	1.2670	1.3407	1.3302
No task-free subnet	0.7052	1.0775	0.8042	0.7983	0.6653	0.8103
Separate encoders	0.9706	1.2995	0.8645	1.0115	0.9243	1.0130
Individual CNNs	1.2551	1.3472	1.1117	1.1962	1.2676	1.2352
Train only on synthetic data	2.3348	4.7550	2.1966	1.9611	2.3486	2.7229
No pre-train on synthetic data	9.9380	16.6584	4.7476	9.7775	11.1243	10.4285
Ours	0.8672	1.1515	0.7314	0.8614	0.8106	0.8834

Table 2: Per-task performance of the ablated models in sAUC↑.

method	Signing	Form	Info.	Shopping	Comm.	Average
	up	filling	browsing		joining	
No task-specific subnet	0.5807	0.5602	0.5809	0.5858	0.5799	0.5779
No task-free subnet	0.6438	0.6185	0.6212	0.6275	0.6341	0.6289
Separate encoders	0.6513	0.6026	0.6281	0.6326	0.6317	0.6291
Individual CNNs	0.6262	0.5944	0.5915	0.6176	0.6009	0.6058
Train only on synthetic data	0.6282	0.5975	0.6098	0.6179	0.6187	0.6143
No pre-train on synthetic data	0.5582	0.5211	0.5890	0.5472	0.5475	0.5528
Ours	0.6538	0.6331	0.6435	0.6414	0.6519	0.6448

Table 3: Per-task performance of the ablated models in NSS↑.

method	Signing	Form	Info.	Shopping	Comm.	Average
	up	filling	browsing		joining	
No task-specific subnet	0.4065	0.4001	0.4165	0.4300	0.4056	0.4116
No task-free subnet	0.6123	0.5363	0.5168	0.5524	0.5804	0.5592
Separate encoders	0.6355	0.4838	0.5614	0.5602	0.5922	0.5664
Individual CNNs	0.5707	0.4481	0.4449	0.5567	0.4741	0.4974
Train only on synthetic data	0.6050	0.4884	0.5359	0.5609	0.5724	0.5521
No pre-train on synthetic data	0.3702	0.1983	0.4749	0.2962	0.3423	0.3374
Ours	0.6458	0.5936	0.6238	0.6066	0.6381	0.6217

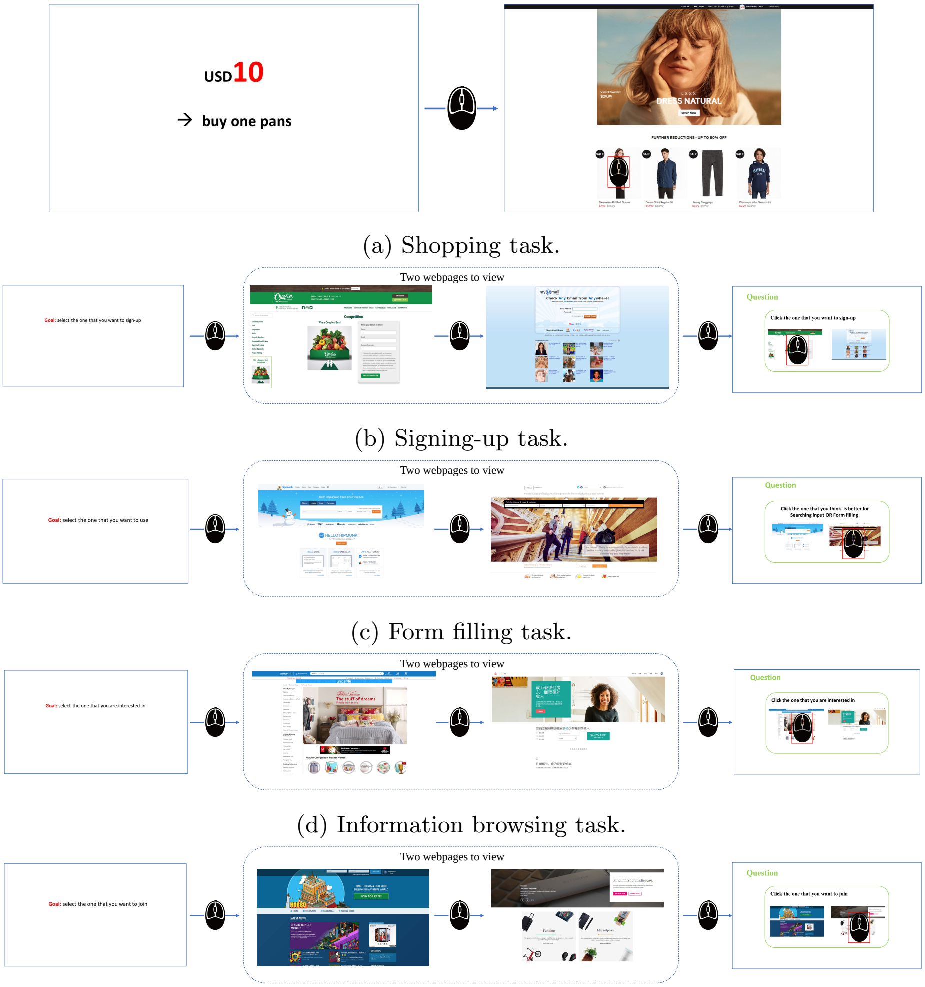
3 Our Evaluation Dataset

Our evaluation dataset consists of 200 web pages scraped from the web. We show the distributions of proportions of semantic components on the webpages in Fig. 3, which demonstrates that our web pages contain contents of varying densities, from little text to heavy text and from few images to many images. In Fig. 4, we plot the spatial distributions of semantic components and white space on a normalized webpage space. The semantic components tend to be uniformly distributed over the webpage space, which indicates the diversity of the webpages in layout. Some sample webpages from our dataset are shown in Fig. 5.

To collect task-driven saliency data on our dataset, we performed an eye-tracking experiment to collect the gaze data of different viewers on our webpages under different task conditions, using a Tobii T60 eye tracker. We adopted the same methodology as [2]. In each viewing session, given a task-specific goal, participants were asked to view one or two webpage images and then complete the corresponding task. Specifically, for the "shopping" task (Fig. 2a), participants were first shown a goal, *i.e.,* , buying a favorite product with a given budget. They were then shown a webpage image to view and asked to select a product item to buy from the webpage. For the other tasks (Fig. 2b-2e), participants were first shown a task-related goal, followed by two webpages to view in sequence. They were then asked to choose one of the two webpages that they preferred to use in order to accomplish the given goal. Note that the task order and webpage presentation order were randomized for each participant, and each participant viewed a given webpage at most once. For each webpage, we collected eye gaze data from 10 participants, and then aggregated all the eye gaze data to form its saliency map.

4 More Qualitative Results

More qualitative results of our method and the prior methods are shown in Fig. 6 and 7.


USD10
→ buy one pans
Goal: select the one that you want to sign-up
Two webpages to view
Question
Click the one that you want to sign-up
Goal: select the one that you want to use
Two webpages to view
Question
Click the one that you think is better for Searching input OR Form filling
Goal: select the one that you are interested in
Two webpages to view
Question
Click the one that you are interested in
Goal: select the one that you want to join
Two webpages to view
Question
Click the one that you want to join

(a) Shopping task.

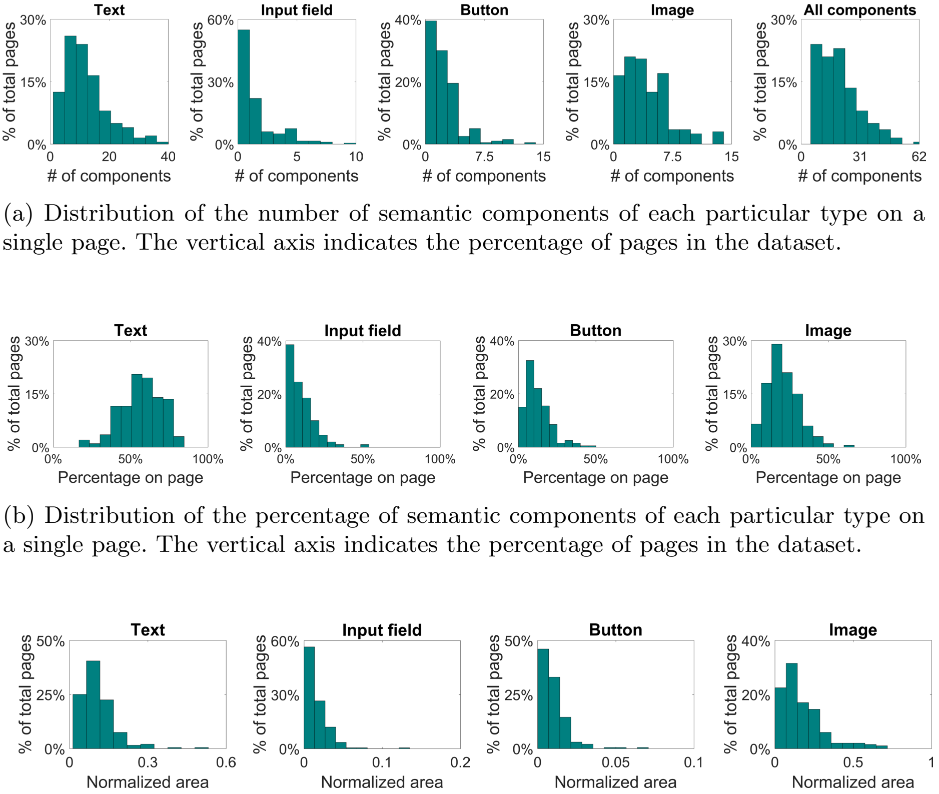
(b) Signing-up task.

(c) Form filling task.

(d) Information browsing task.

(e) Community joining task.

Fig. 2: The workflow used to collect eye gaze data under different tasks in our eye-tracking experiment.

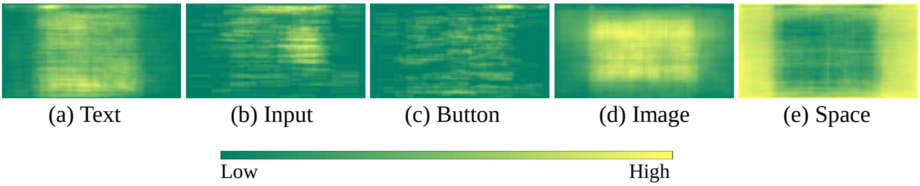


(a) Distribution of the number of semantic components of each particular type on a single page. The vertical axis indicates the percentage of pages in the dataset.

(b) Distribution of the percentage of semantic components of each particular type on a single page. The vertical axis indicates the percentage of pages in the dataset.

(c) Distribution of the normalized area of semantic components of each particular type on a single page. The area is normalized with respect to the entire page area. The vertical axis indicates the percentage of pages in the dataset.

Fig. 3: Statistics of evaluation dataset



(a) Text (b) Input (c) Button (d) Image (e) Space

Fig. 4: Spatial distributions of semantic components (a-d) and white space (e) in our dataset over a normalized webpage space. The color of a pixel indicates the probability of a semantic component at the pixel.

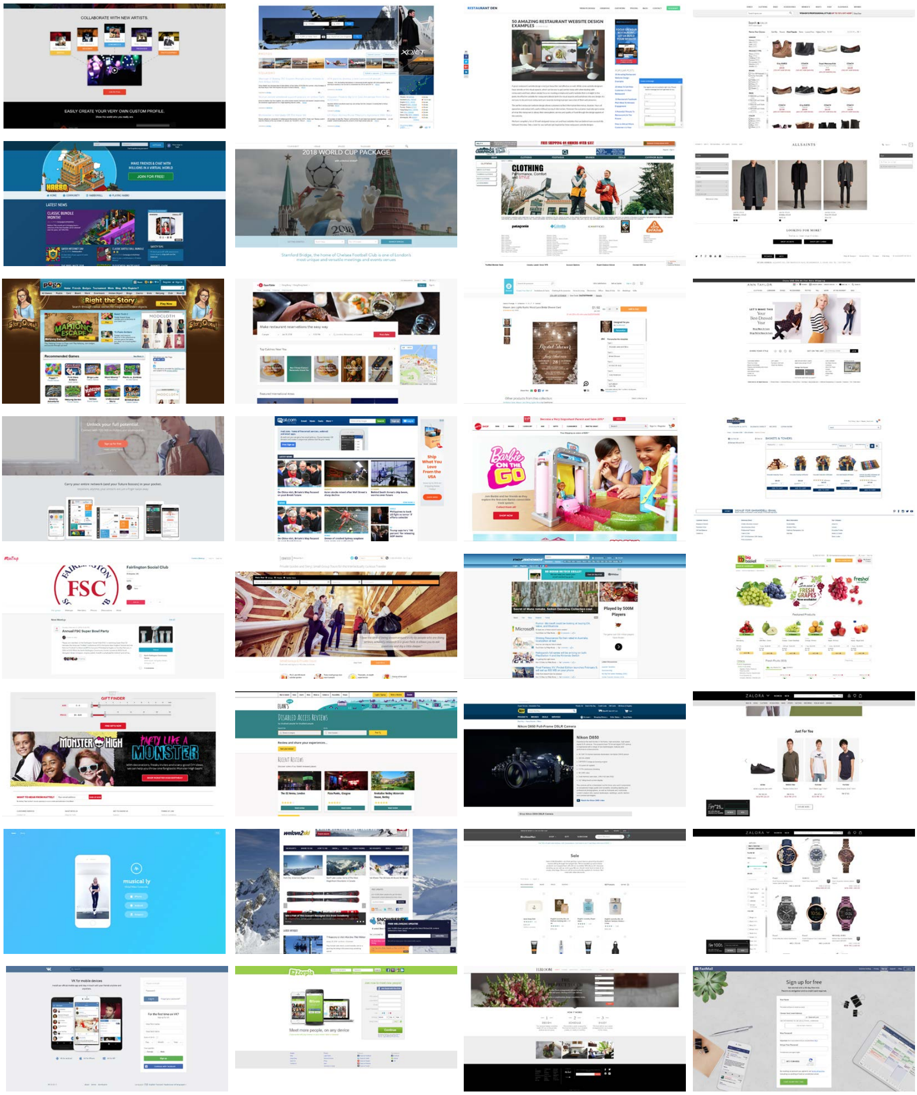


Fig. 5: Samples from our evaluation dataset.

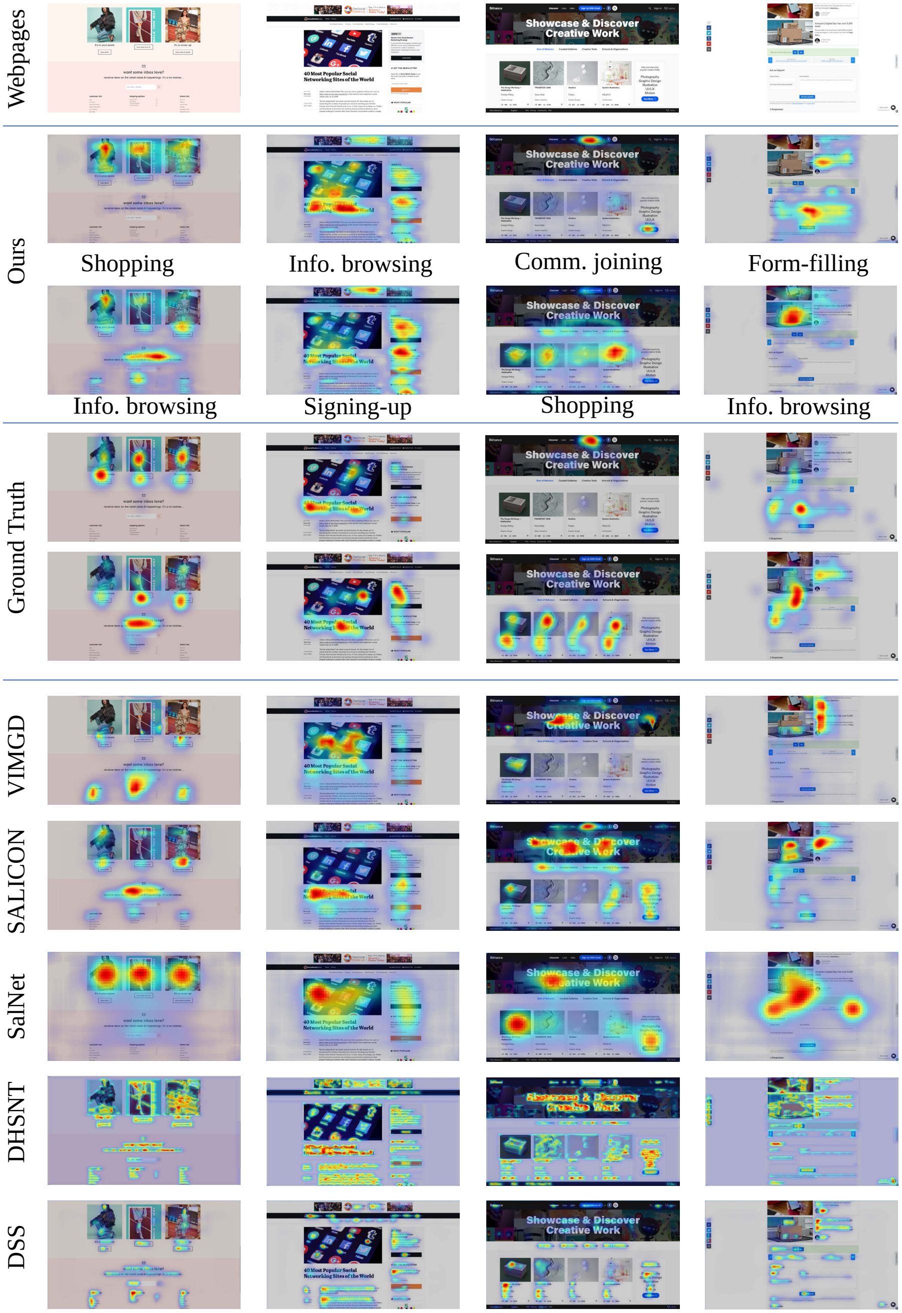


Fig. 6: Saliency prediction results of our method and prior methods under different task conditions.

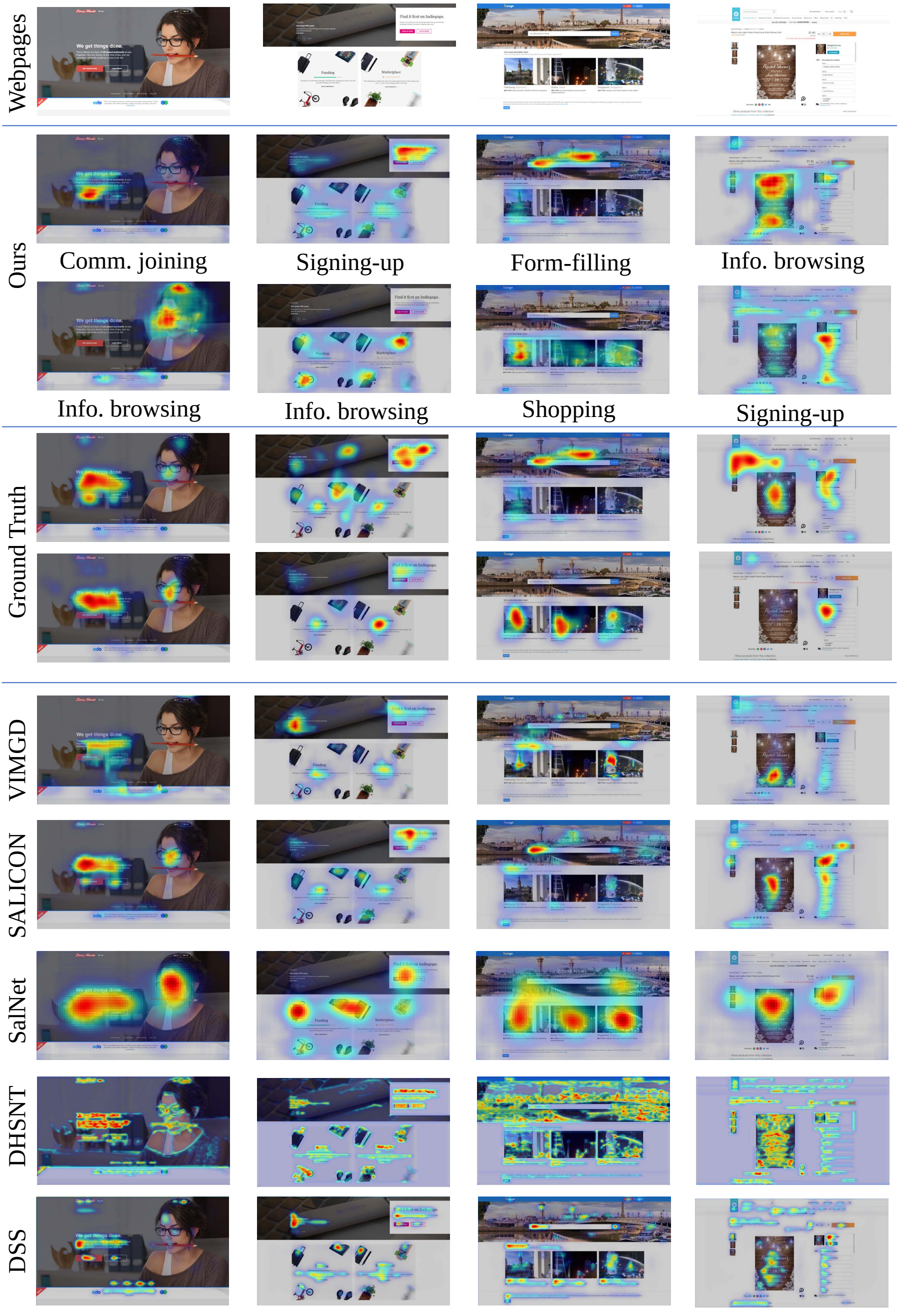


Fig. 7: Saliency prediction results of our method and prior methods under different task conditions.

References

1. Shen, C., Zhao, Q.: Webpage saliency. In: ECCV. (2014)
2. Pang, X., Cao, Y., Lau, R., Chan, A.: Directing user attention via visual flow on web designs. In: SIGGRAPH Asia. (2016)