Supplementary Material for "Scanpath Prediction on Information Visualisations"

Yao Wang, Mihai Bâce, and Andreas Bulling

This document contains the implementation details of finetuning MD-SEM and training PathGAN (Section 1), distribution of scanpath length and fixation duration from the MASSVIS dataset (Figure 1), The accumulated fixation distribution from the MASSVIS dataset (Figure 2), transition matrices of two viewers in the MASSVIS (Figure 3), example annotations from the MASSVIS (Figure 4), full scanpath metrics of Figure 6 in the main manuscript (Figure 6), example element fixation density (EFD) maps and predictions of MD-EAM in MASSVIS (Figure 5), three example scanpath predictions of our UMSS model (Figure 7 – 9), an example questionnaire interface from our user study (Figure 10), quantitative results on scanpath prediction for the full 10-second ground truth (Table 1), and MASSVIS [1] dataset split (Table 2).

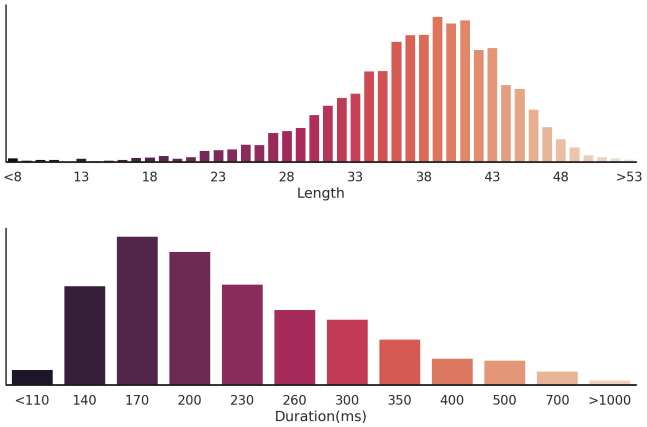
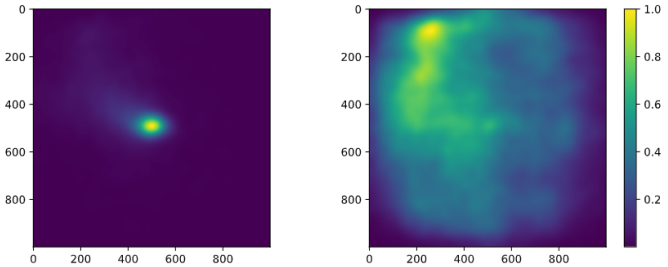


Fig. 1: Distributions of scanpath length (top) and fixation duration (bottom) from the MASSVIS [1, 9] dataset.



(a) (b)

Fig. 2: Accumulated fixation distribution from the MASSVIS dataset. (a) The first fixations of all viewers. (b) The rest fixations except the first fixations of all viewers. There is a strong centre bias within the first fixations across all viewers. This is due to the experiment setting, where a fixation cross shows up right before the image appears on the screen [1].

1 Implementation Details

1.1 Fine-tuning MD-SEM

We followed original setting for fine-tuning MD-SEM [2]. The loss weights combination was CCM = 3, KL = 10, CC = -5 and NSS = -1. Normalized Scanpath Saliency (NSS) [3] calculates the performance of a saliency map model is defined to be the average saliency value of fixated pixels in the normalized saliency maps. CCM is the Pearson's Correlation Coefficient (CC) [4] on pairs of saliency maps at adjacent durations, which is computed as the difference between the ground truth and predicted scores [2]. Kullback-Leibler divergence (KL) computes the Kullback-Leibler divergence between the empirical saliency maps and the model saliency maps after converting both of them into probability distributions[5]. Hyperparameters were batch size = 8, and initial learning rate = $1E-4$. Adam optimiser [6] was used for gradient descent.

1.2 Training PathGAN

The Root Mean Squared Propagation (RMSprop) optimizer and Binary Cross Entropy loss with learning rate = $1E-4$, and rho = 0.9, epsilon = $1E-08$, decay = $1E-07$ are used for both training and fine-tuning. During fine-tuning, we randomly mixed 5 % of training data from SALICON [7] in each epoch to prevent forgetting [8]. We trained PathGAN for 125 epochs on SALICON and 40 epochs on MASSVIS.

TABLE 1: Quantitative evaluation on MASSVIS for the full 10-second ground truth in terms of Dynamic Time Warping (DTW) and scaled Time Dimension Embedding (sTDE) metrics. Best results are shown in **bold**, second best are underlined.

Methods	DTW (2D) ↓	sTDE ↑
Human	8978.57	0.932
PathGAN [10]	10394.58	0.866
PathGAN-official [10]	18396.09	0.764
DCSM [11]	**9822.26**	0.876
Saltinet [12]	13916.36	0.878
DVS+Saltinet [12, 13]	13556.94	0.884
MDSEM+Saltinet [2, 12]	13763.52	0.889
UMSS (Ours)	10040.11	**0.903**

TABLE 2: MASSVIS [1] Dataset split by visualisation source and type.

		Train	Evaluation
Source	Government	83 (25.4%)	17 (25.8%)
	Infographics	77 (23.5%)	15 (22.7%)
	News	101 (30.9%)	21 (31.8%)
	Scientific	66 (20.2%)	13 (19.7%)
Type	Bar	67 (20.5%)	17 (25.8%)
	Pie	11 (3.4%)	5 (7.6%)
	Line	57 (17.4%)	6 (9.1%)
	Scatter	13 (4.0%)	2 (3.0%)
	Table	28 (8.6%)	4 (6.1%)
	Combination	18 (5.5%)	4 (6.1%)
	Other	133 (40.7%)	28 (42.4%)
	Sum	327 (100%)	66 (100%)

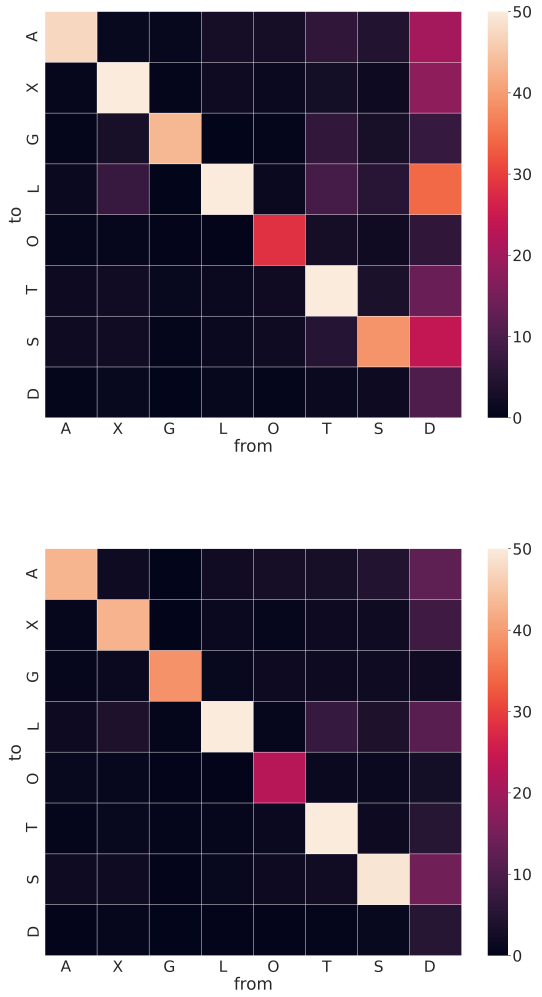
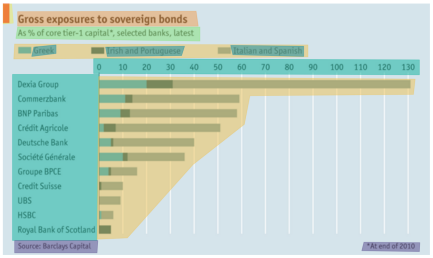
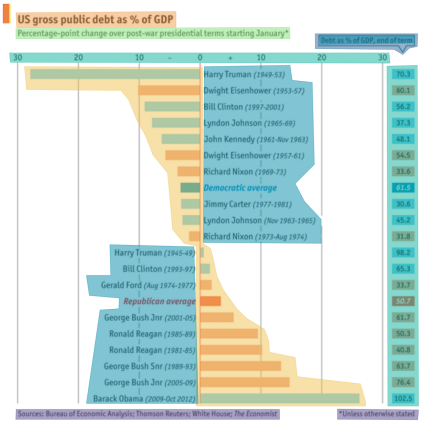
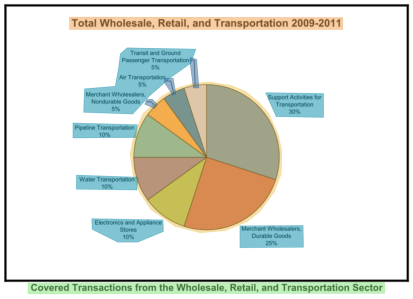
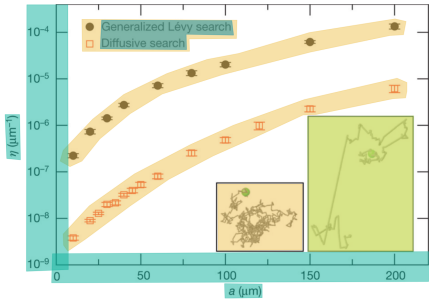


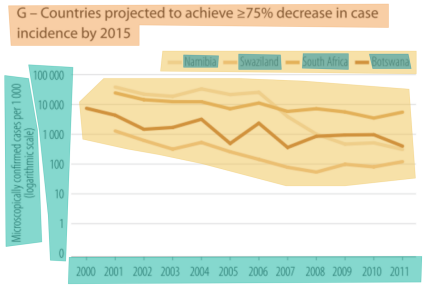
Fig. 3: Transition matrices of two viewers in MASSVIS. Viewers tend to look at Title and Legend continuously before jumping to other regions, while they tend to read Data in cooperation with Annotation, Axis, and Legend. A: Annotation, X: Axis, G: Graphics, L: Legend, O: Object, T: Title, S: Source etc., D: Data.











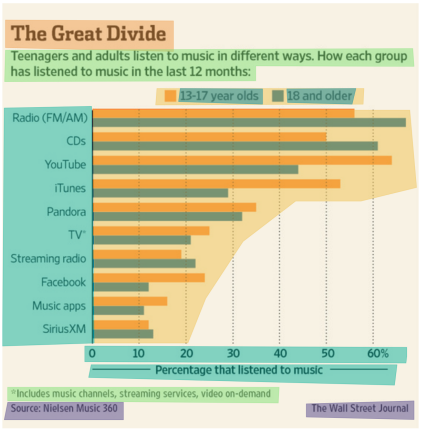
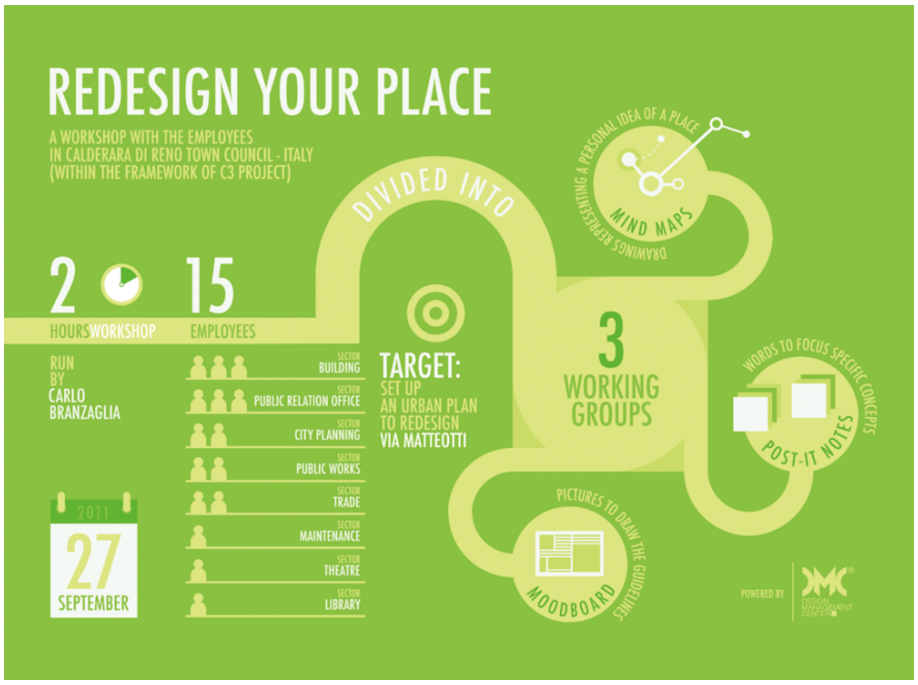
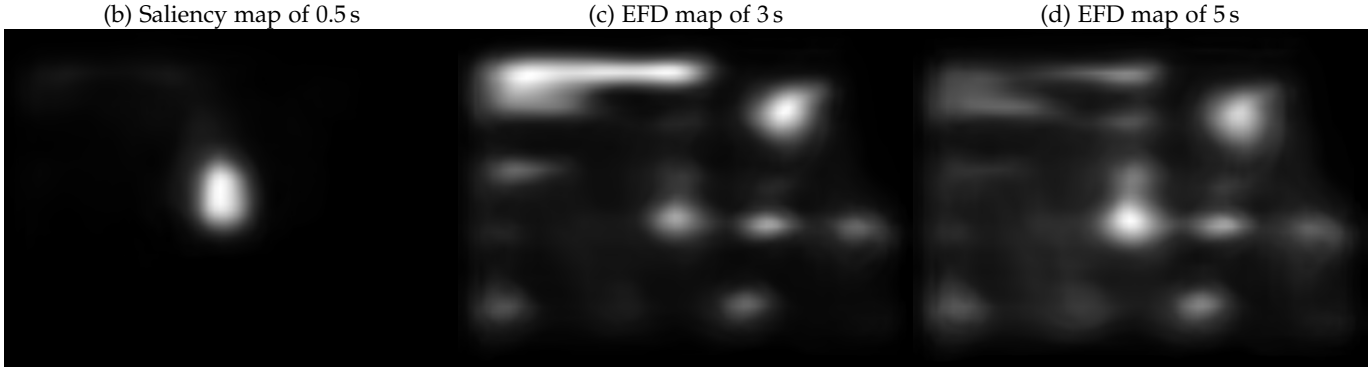


Fig. 4: Example visualisations from the MASSVIS dataset as well as visualisation element annotations highlighted in colour. Each visualisation element (e.g. Title or Label) have a unique colour and the colouring policy is consistent with Figure 2 from the main manuscript.



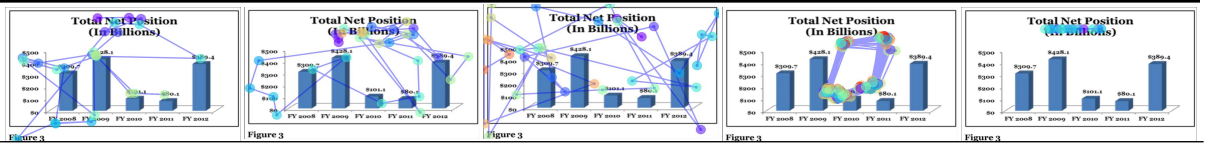
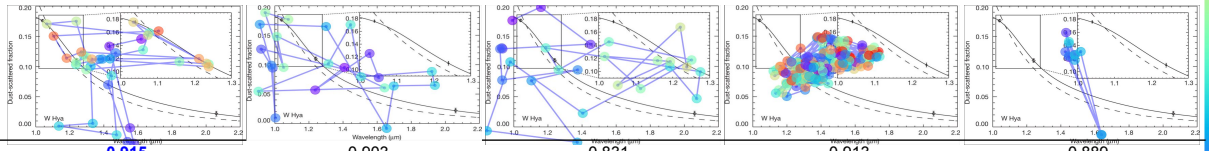
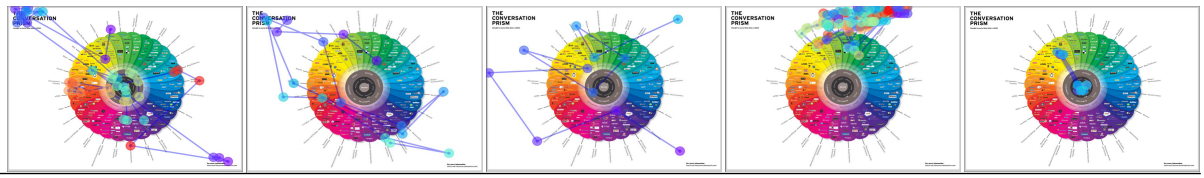
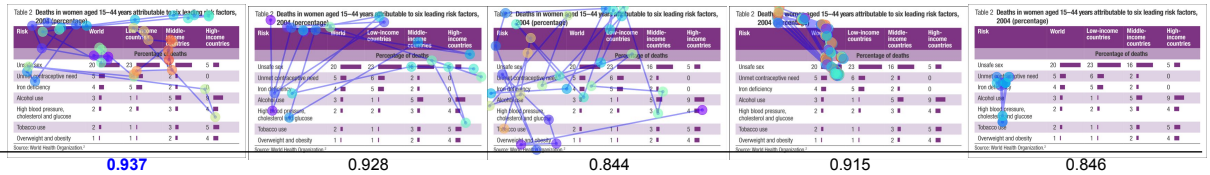
(a) Example stimuli

(b) Saliency map of 0.5 s (c) EFD map of 3 s (d) EFD map of 5 s



(e) Prediction of 0.5 s (f) Prediction of 3 s (g) Prediction of 5 s

Fig. 5: One example stimulus in MASSVIS (a), and the corresponding saliency map of 0.5 s (b), element fixation density (EFD) maps of 3 s (c), and 5 s (d) time duration, and predictions of MD-EAM of 0.5 s (e), 3 s (f), and 5 s (g) time duration. MD-EAM is able to preserve element-level information. The attention shift from Title to Data is clearly shown between (f) and (g).

	Human	Ours	Saltinet	PathGAN	DCSM
S1					
sTDE	0.918	0.909	0.831	**0.923**	0.883
DTW (2D)	**5227.88**	7088.17	13412.89	5261.85	6550.46
MM-Shape	0.951	0.933	0.888	**0.976**	0.949
MM-Direction	**0.814**	0.738	0.647	0.719	0.753
MM-Length	0.941	0.919	0.849	**0.965**	0.940
MM-Position	**0.805**	0.705	0.653	0.800	0.750
Expert Rating	**1.3 (1st)**	2.0 (2nd)	3.3 (3rd)	4.0 (4th)	4.4 (5th)
S2					
sTDE	**0.922**	0.910	0.825	0.906	0.896
DTW (2D)	4662.87	7287.95	9152.58	7010.13	**6014.97**
MM-Shape	0.937	0.942	0.890	**0.962**	0.947
MM-Direction	0.804	0.750	0.768	0.656	**0.814**
MM-Length	0.933	**0.944**	0.872	0.941	0.936
MM-Position	0.794	**0.810**	0.720	0.749	0.799
Expert Rating	2.0 (2nd)	**1.1 (1st)**	3.2 (3rd)	4.4 (5th)	4.3 (4th)
S3					
sTDE	**0.915**	0.903	0.831	0.913	0.889
DTW (2D)	5766.45	7512.26	10938.40	7386.64	**5554.80**
MM-Shape	0.939	0.928	0.889	**0.964**	0.916
MM-Direction	**0.779**	0.737	0.712	0.701	0.687
MM-Length	0.932	0.924	0.857	**0.952**	0.922
MM-Position	**0.780**	0.748	0.758	0.748	0.774
Expert Rating	**1.2 (1st)**	2.3 (2nd)	3.5 (4th)	4.8 (5th)	3.4 (3rd)
S4					
sTDE	**0.943**	0.908	0.817	0.879	0.897
DTW (2D)	**3166.89**	4581.12	4216.57	8184.79	4483.85
MM-Shape	0.972	0.948	0.879	**0.984**	0.941
MM-Direction	**0.742**	0.720	0.684	0.693	0.691
MM-Length	0.973	0.931	0.795	**0.979**	0.933
MM-Position	**0.896**	0.790	0.640	0.660	0.817
Expert Rating	**1.7 (1st)**	**1.7 (1st)**	3.4 (3rd)	4.0 (4th)	4.2 (5th)
S5					
sTDE	**0.937**	0.928	0.844	0.915	0.846
DTW (2D)	**4859.21**	5595.11	8945.90	5828.74	9074.21
MM-Shape	0.960	0.940	0.895	**0.973**	0.947
MM-Direction	**0.837**	0.732	0.746	0.687	0.747
MM-Length	**0.957**	0.926	0.865	0.955	0.936
MM-Position	**0.856**	0.802	0.687	0.765	0.700
Expert Rating	1.8 (2nd)	**1.7 (1st)**	4.4 (5th)	4.0 (3rd)	4.1 (4th)

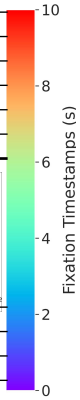


Fig. 6: Full table of examples of mismatches between scanpath prediction performance as seen through the evaluation metrics and visualisation expert ratings. Each row (one visualisation from MASSVIS) shows several metrics that are contradictory to expert rating (orange), or consistent with expert rating (blue).

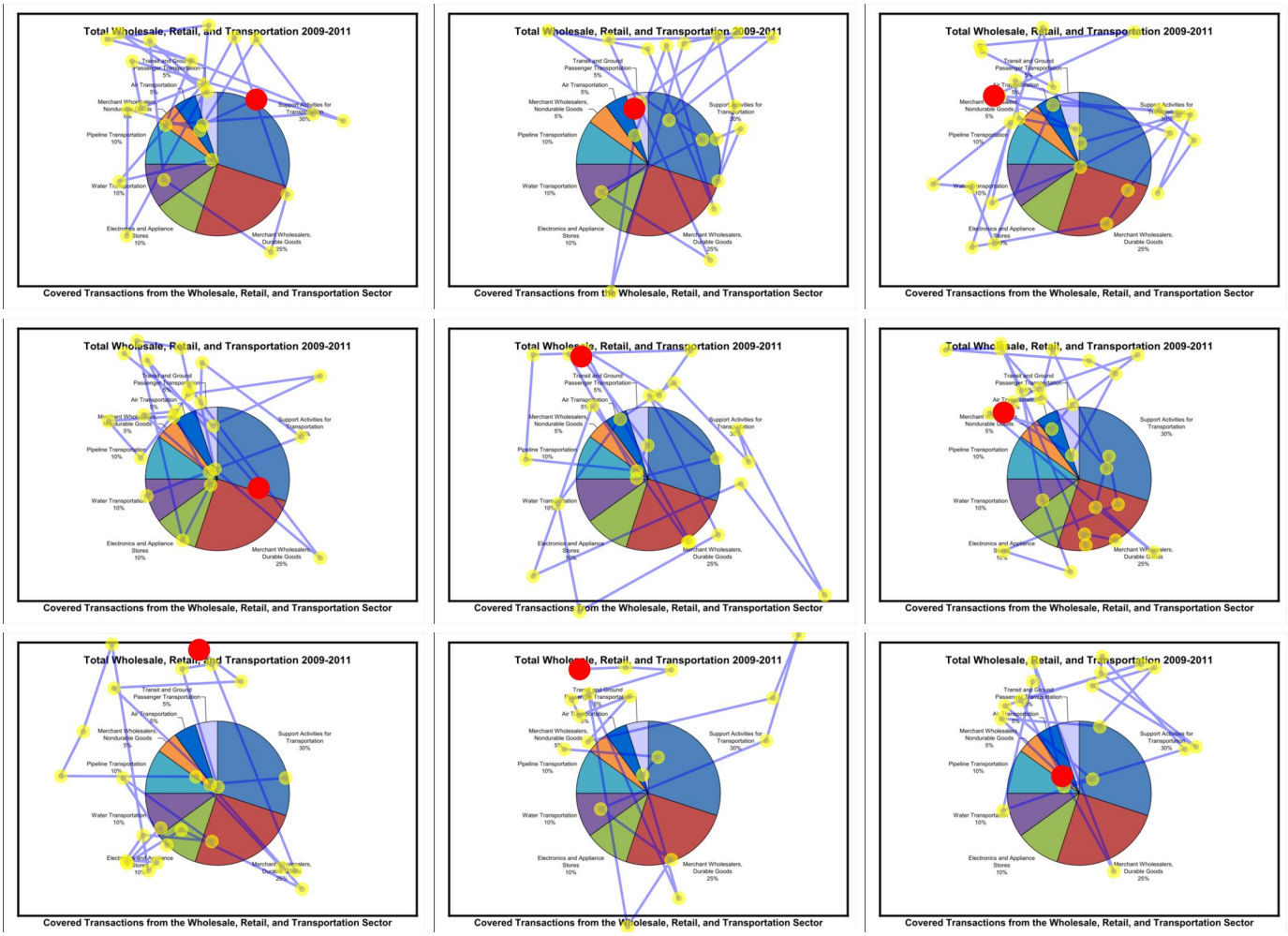


Fig. 7: Scanpath predictions using UMSS (ours) on a sample visualisation from the MASSVIS dataset.

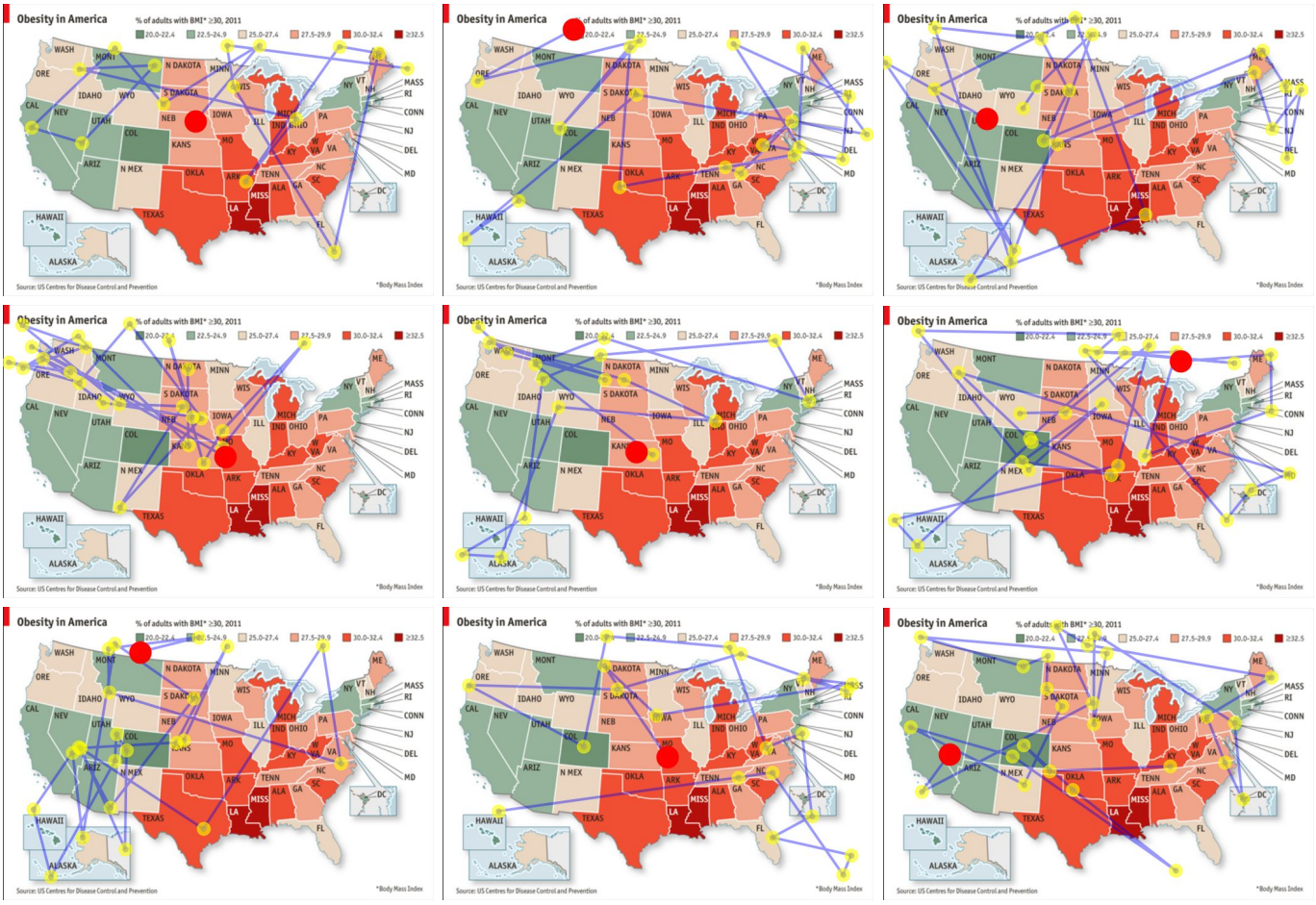


Fig. 8: Scanpath predictions using UMSS (ours) on a sample visualisation from the MASSVIS dataset.

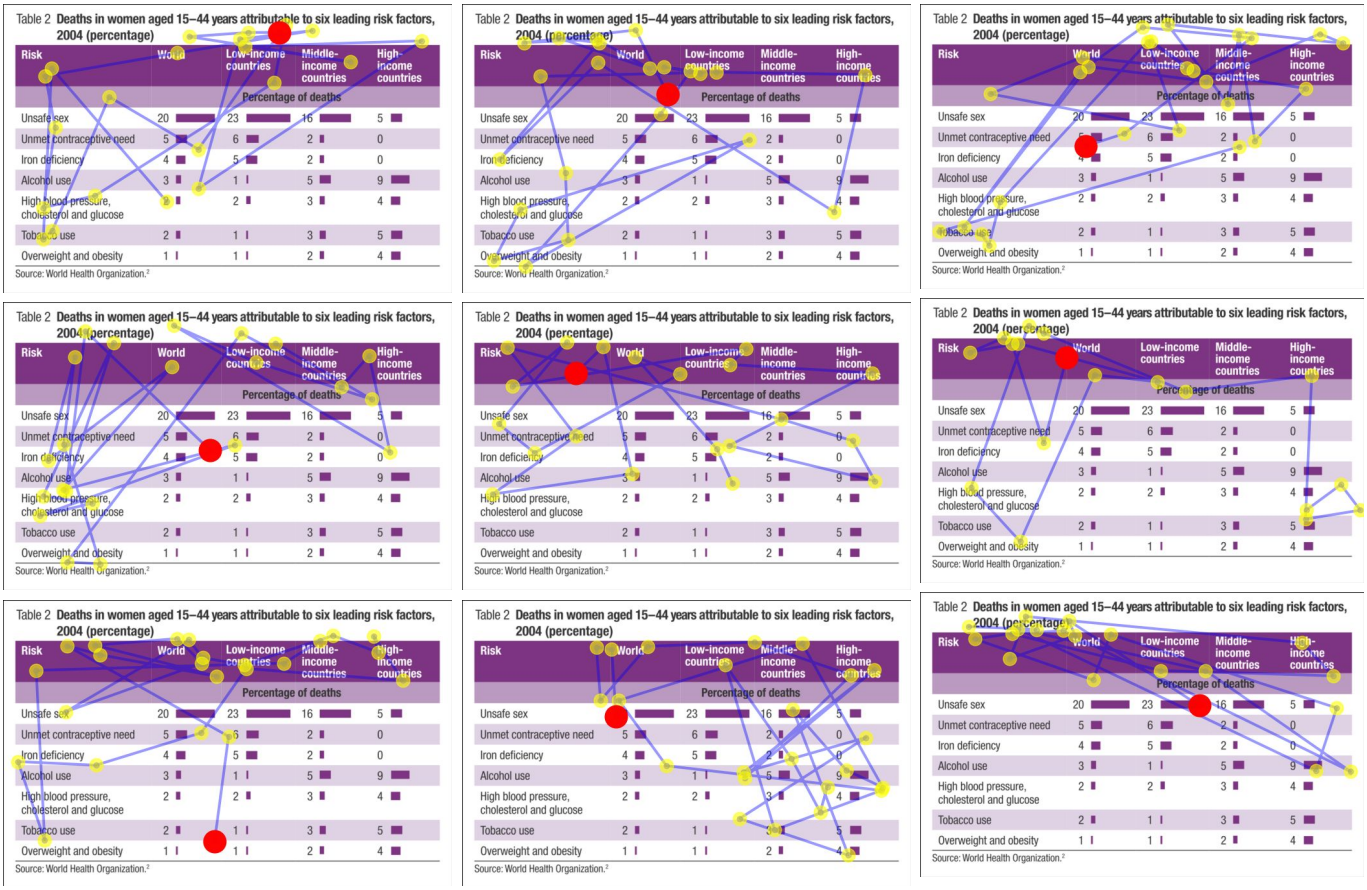


Fig. 9: Scanpath predictions using UMSS (ours) on a sample visualisation from the MASSVIS dataset.

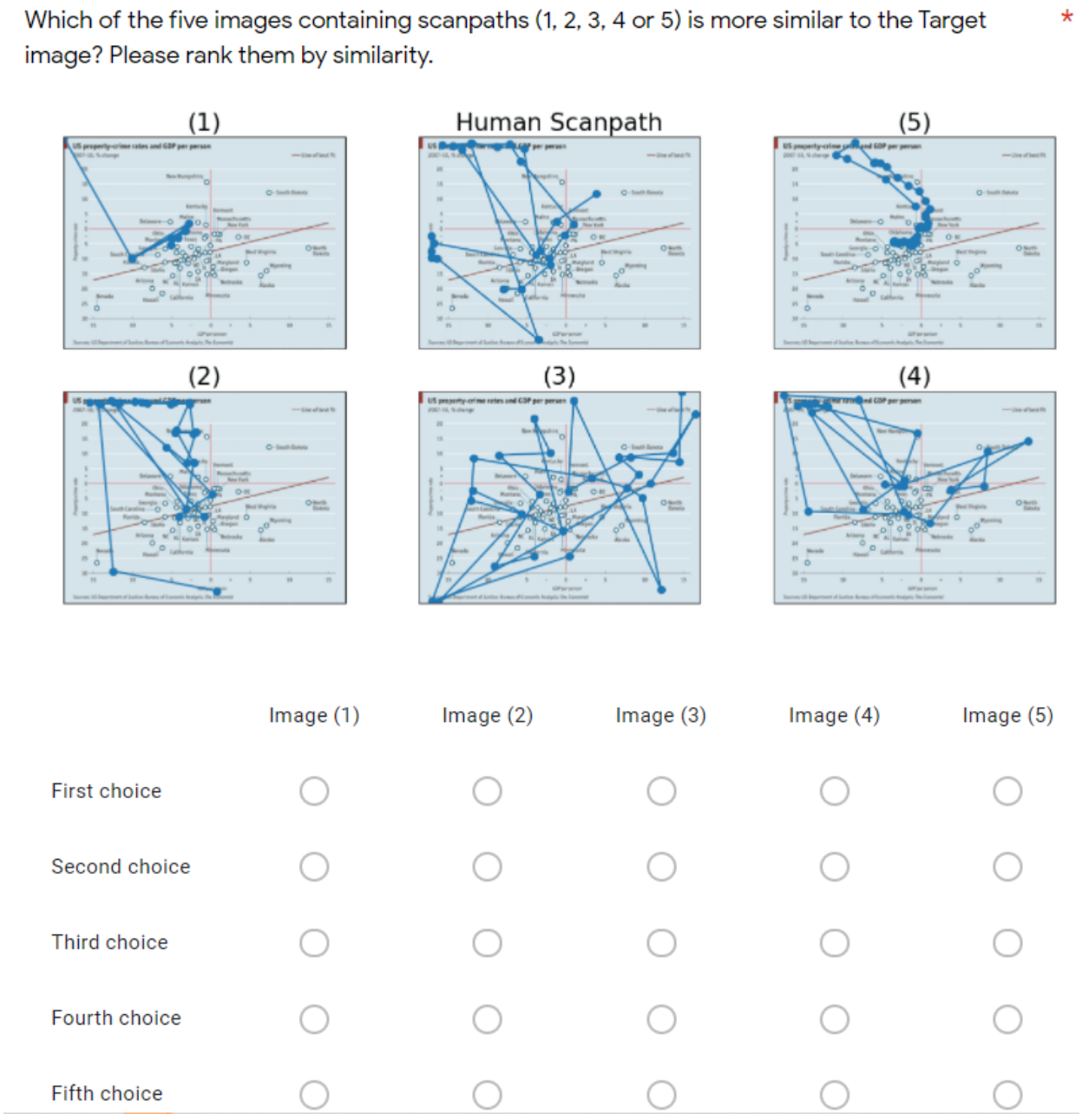


Fig. 10: An example questionnaire interface of one trial (out of 40) from our user study with visualisation experts. Scanpaths were shown to the study participants as GIFs. Fixations and saccades were drawn sequentially on the image. At the end of one loop, the visualisation paused for a short period of time until a new loop started to allow subjects to compare all the scanpaths. Study participants had to rank the five options in order of their similarity when compared to one ground-truth, human scanpath. The presentation order of the baselines (1, 2, 3, 4, and 5) was counterbalanced according to a latin-square study design.

REFERENCES

[1] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva, "Beyond memorability: Visualization recognition and recall," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 519–528, 2015.

[2] C. Fosco, A. Newman, P. Sukhum, Y. B. Zhang, N. Zhao, A. Oliva, and Z. Bylinskii, "How much time do you have? modeling multi-duration saliency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4473–4482.

[3] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[4] T. Jost, N. Ouerhani, R. Von Wartburg, R. Müri, and H. Hügli, "Assessing the contribution of color in visual attention," *Computer Vision and Image Understanding*, vol. 100, no. 1-2, pp. 107–123, 2005.

[5] M. Kummerer, T. S. Wallis, and M. Bethge, "Saliency benchmarking made easy: Separating models, maps and metrics," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 770–787.

[6] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[7] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1072–1080.

[8] C. Fosco, V. Casser, A. K. Bedi, P. O'Donovan, A. Hertzmann, and Z. Bylinskii, "Predicting visual importance across graphic design types," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 2020, pp. 249–260.

[9] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister, "What makes a visualization memorable?" *IEEE transactions on visualization and computer graphics*, vol. 19, no. 12, pp. 2306–2315, 2013.

[10] M. Assens, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Pathgan: visual scanpath prediction with generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 406–422.

[11] W. Bao and Z. Chen, "Human scanpath prediction based on deep convolutional saccadic model," *Neurocomputing*, vol. 404, pp. 154–164, 2020.

[12] M. Assens Reina, X. Giro-i Nieto, K. McGuinness, and N. 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*, 2017, pp. 2331–2338.

[13] L. E. Matzen, M. J. Haass, K. M. Divis, Z. Wang, and A. T. Wilson, "Data visualization saliency model: A tool for evaluating abstract data visualizations," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 563–573, 2017.