Analysis of scores, datasets, and models in visual saliency prediction Supplementary Material

Ali Borji[†] Hamed R. Tavakoli[+] Dicky N. Sihite[†] Laurent Itti[†]
[†] Department of Computer Science, University of Southern California, Los Angeles
[+]Center for Machine Vision Research, University of Oulu, Finland

This supplementary material accompanies the paper "Analysis of scores, datasets, and models in visual saliency prediction" at ICCV 2013. It provides more detail on model scoring as well as some illustrative figures.

1. Saliency Model Comparison

We compare 32 saliency models. Details of our model comparison setup and compared models are provided in this supplement. We report eye movement prediction accuracy of models using three scores: Shuffled AUC (sAUC), Correlation Coefficient (CC), and Normalized Scanpath Saliency (NSS). Additionally, we show results of statistical significance (statistical tests) of models using t-test. Results are reported over four widely-used and benchmark eye-tracking datasets which are publicly available. In the following, we explain statistics of employed datasets.

2. Datasets

The main reason behind employing several datasets is that current datasets have different image and feature statistics, stimulus variety, biases (e.g., center-bias), and eye tracking parameters. Hence, it is necessary to employ several datasets as models leverage different features that their distribution varies across datasets. Some datasets are publicly available[1]. The MIT [2] dataset is the largest

2.1. Static image datasets

TORONTO[2] **[1].** This is the most widely-used dataset for saliency model comparison. It contains 120 color images with the resolution of 511×681 pixels from indoor and outdoor environments. Images are presented at random to 20 subjects for 4 seconds with 2 seconds of delay (a gray mask) in between.

MIT[3] **[2].** This is the largest dataset containing 1003 images (resolution from 405×1024 to 1024×1024 pixels) collected from Flicker and LabelMe datasets. There are 779 landscape and 224 portrait images. Fifteen subjects freely viewed images for 3 sec. with 1 sec. delay in between.

NUSEF[4] **[3].** This dataset includes 758 images containing affective stimuli such as expressive faces, nudes, unpleasant concepts, and semantic stimuli such as interactive actions. In total, 75 subjects free-viewed part of the image set for 5 seconds each (on average 25 subjects per image).

KOOTSTRA[5] **[5].** This dataset contains 101 images from 5 different categories: 12 animals, 12 automan, 16 buildings, 20 flowers, and 41 natural scenes. Images are observed by 31 subjects in the age range of 17 to 32 for 3 seconds. Image resolution is 768×1024 pixels. This dataset is specially challenging because there are not explicit objects or salient regions within many of the images.

3. Evaluation Metrics

The most widely used score for saliency model evaluation is the Area Under the ROC Curve (**AUC**) [1]. In AUC, human fixations (heatmap) are considered as the positive set and some points from the image are randomly chosen (uniformly) as the negative set. The saliency map is then treated as a binary classifier to separate the positive samples from negatives. By thresholding over this map and plotting true positive rate vs. false positive rate an ROC curve is achieved and its underneath area is calculated. A concern about the AUC (classic AUC) score is that it generates a large value for a central Gaussian kernel [12]. To tackle center bias, Zhang *et al.* [6] introduced **shuffled AUC** score (sAUC) with the only difference with AUC being that instead of selecting negative points randomly, all fixations (except the positive set) are used as the negative set.

[1] **Toronto [1]**: http://www-sop.inria.fr/members/Neil.Bruce/
NUSEF [3]: http://mmas.comp.nus.edu.sg/NUSEF.html
MIT [2]: http://people.csail.mit.edu/tjudd//WherePeopleLook/index.html
Le Meur [?]: http://www.irisa.fr/temics/staff/lemeur/visualAttention/
Kootstra [5]: http://www.csc.kth.se/∼kootstra/
Tatler [?]: http://www.activevisionlab.org/
DOVES: http://live.ece.utexas.edu/research/doves/
FIFA: http://www.ffadb.com/ - **Reinagel**: http://zadorlab.cshl.edu/

[2]This dataset is available at www-sop.inria.fr/members/Neil.Bruce

[3]This dataset is available at http://people.csail.mit.edu/tjudd/

[4]This dataset is available at http://mmas.comp.nus.edu.sg/NUSEF.html

[5]This dataset is available at http://www.csc.kth.se/∼kootstra/

The next score we use is the **Normalized Scanpath Saliency (NSS)** [7]. NSS is the average of response values at human eye positions (x_H^i, y_H^i) in a model's saliency map (S) that has been normalized to have zero mean and unit standard deviation:

$$NSS = \frac{1}{N}\sum_{i=1}^{N}\frac{S(x_H^i, y_H^i) - \mu_S}{\sigma_s} \quad (1)$$

where N is the number of fixation for each image. $NSS = 1$ indicates that the subjects' eye positions fall in a region whose predicted density is one standard deviation above average. Meanwhile $NSS \leq 0$ indicates that the model performs no better than picking a random position.

The third score is **Correlation Coefficient (CC)** between human saliency map H (a map with frequency of saccades at each location which is usually convolved with a small Gaussian kernel) and a model's saliency map S:

$$CC(H,S) = \frac{\sum_{xy}(H(x,y) - \mu_H).(S(x,y) - \mu_S)}{\sqrt{\sigma_H^2.\sigma_S^2}} \quad (2)$$

In above formula, μ and σ^2 are the mean and the variance of the values in these maps and $\sum_{xy}$ is the covariance matrix. An advantage of CC is that it has the well-defined upper-bound of 1.

Note that CC, NSS, and classic AUC scores are all affected by center-bias. Here, we emphasize more on shuffled AUC score which is becoming a standard for saliency model evaluation [21][6].

We smoothed saliency maps of each model by convolving them with a Gaussian kernel. We plotted the sAUC of each model over the range of standard deviations of the Gaussian kernel in image width (from 0.01 to 0.13 in steps of 0.01).

In the following, we mention four types of AUC scores used in previous works:

Area Under Curve (AUC). The most widely used score for saliency model evaluation is the Area Under the ROC Curve [18]. Having its roots in signal detection theory, AUC measures the ability of a saliency map in separating fixated locations from non-fixated locations. Thus far three different variations of AUC exist in the literature:

AUC Type 1. First, the prediction map is resized to the image size where fixations have been recorded. Then, human fixations are considered as the positive set and some points from the image are randomly sampled (using a 2D uniform distribution) as the negative set. To form the negative set, some researchers [15, 17] use the non-fixated locations while some take random sample from the entire image. The saliency map S is then treated as a binary classifier to separate the positive samples from the negatives. By thresholding over the saliency map, the *true positive rate* is the proportion of fixations above a threshold while the *false positive rate* is the proportion of random points above that same threshold. Then, ROC is plotted by sweeping a threshold from 0 to 1 (on a normalized map) and then AUC is calculated. Perfect prediction corresponds to a score of 1 while a score of 0.5 indicates chance level. This definition has been used in [19, 8, 17].

AUC Type 2. Here, instead of using random points, *true positive rate* (number of fixations falling on the thresholded saliency map) is plotted against the normalized saliency map area above a threshold. This score has been used by [2, 22]. To tackle center-bias, [22] proposed a control strategy (called "cross-image control"): For each saliency map, instead of using fixations for that image, they used fixations from another randomly selected image (to see how much just viewing strategy scores).

AUC Type 3. Above AUC definitions receive many true positives for a trivial central Gaussian model since majority of fixations happen in the center [20]. To tackle the center-bias issue and handle the spatial priors in viewing (compensation), [16] and [18] suggested to draw random locations from the distribution of eye fixations. Here, we use the AUC Type-1, but instead of uniform random points for an image, we draw negative points from fixations of other observers over other images. This way, central fixations receive less credit compared with off-center (non-trivial) saccades. This score is also known as the Shuffled AUC score and has been used in [6, 18, 21]. [18] argue that it is better to draw the negative samples from the fixations of the same observer on different images to account for any *individual* biases.

AUC Type 4. This type of AUC is basically the same as type 3 but instead of drawing random points from fixations of other subjects over all other images, random points are same fixation locations as the current image but salinecy values are taken from other images. While AUC type three gives score about 0.5, this type of AUC leads to score of exactly 0.5.

4. Fixation Location Prediction Results

Fig. 1 illustrates smoothing process for an image from Toronto [1] dataset along with smoothed prediction map of the AWS [9] model. Second raw in this figure shows the size of Gaussian kernels.

Tables 1, 2, and 3 show shuffled AUC, CC, and NSS scores of all compared models over datasets mentioned above. They also show performances of a central Gaussian blob and the human inter-observer model. Size of the convolved Gaussian blob where each model achieves its maximum performance is also listed.

Fig. 2 shows shuffled AUC scores of models over the range of smoothness of saliency maps for four datasets. Note that in the paper (main text), we reported the maximum value of this range. Accuracies here are slightly above those reported in the paper using our own code. All trends are the same. The difference (with classic AUC) is the way

that negative set is calculated. We used the fixations over other images as the negative set, while [21] uses the fixations over other images and those fixations over the current tested image which are not part of the positive set.

Figs. 3, 4, and 5 show results of statistical significance test of compared models (each pair) using t-test. Results are reported for shuffled AUC, CC, and NSS scores over Toronto, NUSEF, and MIT datasets. Left-hand columns show the sorted accuracies (maxmia over range of Gaussians) and right-hand columns show the t-test values of models (sorted according to the left-hand columns).

Figs. 6 shows the sAUC, CC, NSS of models over five categories of Kootstra [5] dataset. Fig. 7 shows shuffled AUC scores of models over categories of NUSEF dataset.

Figs. 8, 9, 10, 11, and 12 show some samples from each category of NUSEF [3] dataset along with predictions of compared saliency models.

Fig. 13 shows histogram of emotional valence over the emotional images of the NUSEF dataset.

5. Sequence Comparison

Band-width: Mean shift is a non-parametric clustering technique with no assumption considering shape and number of clusters. It is robust toward outliers. Mean shift approximates maxima of a density function from discrete samples of that function. Assuming that we have a set of points x_i, $i = 1 \dots n$ in a d-dimensional space R^d, mean shift is defined as an iterative multivariate kernel density estimator [13]. The density function can be defined in terms of kernel $K(x)$ with bandwidth h as follows [14] :

$$\hat{f}_{h,K}(x) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{n} K\left(\| \frac{x - x_i}{h} \|^2 \right). \qquad (3)$$

where $c_{k,d}$ is a constant. The bandwidth parameter h influences mean shift performance and results in different clusters. In order to select appropriate bandwidth, we compute fixation transition between clusters. The bandwidth which provides us a clustering with maximum transition between fixation clusters is chosen.

Fig. 14 illustrates scanpath coding and model scanpath prediction evaluation for an example image from the Toronto dataset. See also http://people.irisa.fr/Olivier.Le_Meur/publi/2012_BRM/index2.html for another way of scanpath evaluation.

6. Task Decoding

The NUSEF dataset contains 5 categories over total of 409 images (Event : 36, Face: 52, Nude: 20, Portrait: 123, Others: 178) .

Features from statistics of fixation includes, histogram of observers' saccade velocity (50 bins), saccade orientation (36 bins), saccade length (50 bins), saccade orientation (30 bins), saccade duration (60 bins) and fixation duration (60 bins). In order to compute the histograms for a given image, we initially compute corresponding feature (e.g. saccade velocity, etc.) for each observer and quantize the values into several bins. Later, the histogram of saccade statistic for an image is computed from all the observers and L1 normalized. We also made a fiction histogram by dividing the image into a grid pattern and counting number of fixations in each grid. Saliency based features include histogram of saliency values at fixation points (10 bins), vectorized saliency map of size 20×15 (a vector of size 1×300) and coordinates of ten most salient point coordinates obtained by applying IOR to the saliency map (a vector of size 1×20).

Classification was done using Multi-class support vector machine (SVM) using binary C-SVC support vector classification with RBF kernel. Initially, we pick exactly 20 images from each category and optimized the SVM parameters using a 5-fold cross validation scheme. This process was repeated 100 times by reselecting another random set of images to guaranty the best parameters for the classifier. Later, We utilize all the data available to evaluate the performance of the classifier. In order to cope with the imbalanced data , we do oversampling to change the training samples distribution by duplicating higher-cost training samples until the appearance of different training samples are proportional to their costs. Later, we categorized the images by a 5-fold cross validation scheme. This process was repeated 10 times. For measuring chance, we shuffled the class labels and trained the SVM classifier again.

In order to evaluate saliency models, we extract the saliency features from that model. Tables. 4, 5, 6, and 7 show confusion matrices of stimuli decoding using four saliency models.

6.1. Evaluating Feature Performance

To check performance of different features, we study performance of each individual feature in classification of class categories. Saliency specific features were extracted from human average fixation density map. The following table summarizes the classification accuracy of each feature. Table 8 shows importance (i.e., decoding accuracy) of individual feature channels.

7. Other Saliency Benchmarks

In addition to our model comparison, at least four other efforts exist for saliency model evaluation:

1. http://people.csail.mit.edu/tjudd/SaliencyBenchmark/,
2. http://www.tcts.fpms.ac.be/attention/?article38/saliency-benchmark,
3. http://www.cim.mcgill.ca/∼lijian/database.htm,

4. http://people.irisa.fr/Olivier.Le_Meur/publi/2012_BRM/index2.html

Hopefully ongoing efforts will lead to a unified consensus in comparing visual saliency models. Note that here were focused on models that aim fixation prediction and not salient region detection.

Results (code and data) of our model comparison will be available in our online challenge website at: https://sites.google.com/site/saliencyevaluation/.

References

[1] N.D.B. Bruce and J.K. Tsotsos. Saliency based on information maximization. *NIPS*, 2005.

[2] T. Judd, K. Ehinger, F. Durand and, A. Torralba. Learning to predict where humans look. *ICCV*, 2009.

[3] R. Subramanian, H. Katti, N. Sebe, M. Kankanhalli, and T.S. Chua. An eye fixation database for saliency detection in images. *ECCV*, 2010.

[4] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *NIPS*, 2005.

[5] G. Kootstra, A. Nederveen, and B. de Boer. Paying attention to symmetry. *BMVC*, 2008.

[6] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, SUN: A Bayesian framework for saliency using natural statistics. *JOV*, 2008.

[7] R. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Res.*, 45, 2005.

[8] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. *NIPS*, 2007.

[9] A.G. Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Decorrelation and distinctiveness provide with human-like saliency. *ACIVS*, 2009.

[10] S. Marat, T. Ho-Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué. Modeling spatio-temporal saliency to predict gaze direction for short videos. *IJCV*, 2009.

[11] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE PAMI*, 2006.

[12] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11(3), 2011.

[13] K. Fukunaga and L. Hostetler, The estimation of the gradient of a density function, with applications in pattern recognition, *IEEE Transactions on Information Theory*, vol. 21, pp. 32 – 40, 1975.

[14] D. Comaniciu and P. Meer, Mean shift: a robust approach toward feature space analysis, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, pp. 603 –619, may 2002.

[15] L., Itti, Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Visual Cognition., 12, 2005.

[16] D., Parkhurst and E., Niebur. Scene content selected by active vision. Spatial Vision., 16 , 2003.

[17] Berg, D., Boehnke, S., Marino, R., Munoz, D., and Itti, L. Free viewing of dynamic stimuli by humans and monkeys. Journal of Vision., 9 , 2009.

[18] Tatler, B. W., Baddeley, R. J., and Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. Vision Research, 45 , 643, 2005.

[19] N.D.B. Bruce and J.K. Tsotsos. Saliency based on information maximization. *NIPS*, 2005.

[20] B.W. Tatler. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor bases and image feature distributions. *J. Vision*, 14(7): 2007.

[21] Hou, X., Harel, J., and Koch, C. Image signature: Highlighting sparse salient regions. IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)., 34 , 2012.

[22] Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., and Oliva, A. Modeling search for people in 900 scenes: A combined source model of eye guidance. Visual Cognition., 17 , 2009.

		Predicted Value				
		Event	Face	Nude	Portrait	Other
Actual Value	Event	0.3011	0.1475	0.1214	0.0933	0.3367
	Face	0.1367	0.3473	0.0913	0.2108	0.2138
	Nude	0.1900	0.1130	0.2625	0.2455	0.1890
	Portrait	0.1498	0.2028	0.1486	0.3683	0.1306
	Other	0.2098	0.1411	0.1134	0.1210	0.4147

Table 4. GBVS confusion matrix for stimuli decoding.

		Predicted Value				
		Event	Face	Nude	Portrait	Other
Actual Value	Event	0.2639	0.1167	0.0750	0.0889	0.4556
	Face	0.1827	0.3077	0.0596	0.2404	0.2096
	Nude	0.2200	0.1950	0.2050	0.2000	0.1800
	Portrait	0.0846	0.1740	0.0699	0.5309	0.1407
	Other	0.1955	0.1472	0.0713	0.1140	0.4719

Table 5. AIM confusion matrix for stimuli decoding.

		Predicted Value				
		Event	Face	Nude	Portrait	Other
Actual Value	Event	0.2756	0.0812	0.2197	0.1009	0.3226
	Face	0.1204	0.2713	0.1320	0.2503	0.2260
	Nude	0.2438	0.2054	0.1938	0.1262	0.2308
	Portrait	0.1351	0.2161	0.1422	0.3664	0.1402
	Other	0.2168	0.1703	0.1504	0.1087	0.3538

Table 6. AWS confusion matrix for stimuli decoding.

		Predicted Value				
		Event	Face	Nude	Portrait	Other
Actual Value	Event	0.2350	0.0836	0.0311	0.1714	0.4789
	Face	0.1242	0.2429	0.0221	0.2344	0.3763
	Nude	0.1215	0.0805	0.0295	0.4545	0.3140
	Portrait	0.0584	0.0960	0.0419	0.6190	0.1847
	Other	0.0920	0.0785	0.0037	0.1977	0.6280

Table 7. ITTI model confusion matrix for stimuli decoding.

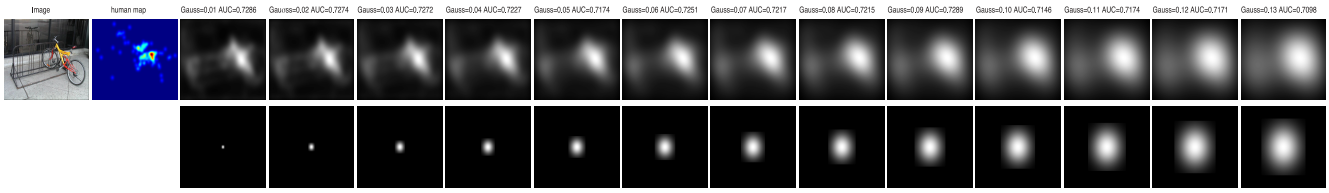

Figure 1. A sample saliency map (from AWS model) smoothed by convolving with a variable-size Gaussian kernel for an image taken from Toronto dataset.

Model	KOOTSTRA	MIT	NUSEF	Toronto
AIM	0.5918 - 1	0.6797 - 3	0.6379 - 4	0.6897 - 2
AWS	0.6250 -1	0.6936 - 1	0.6408 - 2	0.7209 - 2
E-Saliency	0.5647 - 1	0.6182 - 1	-	0.6551 - 3
Bian	0.5745 - 2	0.6064 - 1	0.6286 - 2	0.6083 - 3
Entropy	0.5689 - 3	0.6360 - 4	0.6150 - 5	0.6552 - 4
GBVS	0.5650 - 1	0.6411 - 1	0.5959 - 1	0.6462 - 1
Kootstra	0.5583 - 2	0.6024 - 2	-	0.6128 - 3
Tavakoli	-	0.6508 - 1	0.5645 - 1	0.6434 - 1
HouCVPR	0.5862 - 3	0.6537 - 4	0.6260 - 5	0.6922 - 4
HouNIPS	0.5949 - 1	0.6676 - 3	0.6181 - 4	0.6935 - 1
Itti	0.5797 - 4	0.6216 - 4	0.5721 - 7	0.6279 - 4
Jia Li	0.5566 - 1	0.5038 -13	-	0.6673 - 1
Judd	0.5926 - 1	0.6606 - 3	0.6109 - 4	0.6824 - 3
Le Meur	0.5658 - 1	0.5726 - 1	-	0.6565 - 1
Marat	0.5436 - 2	0.6235 - 3	-	0.6379 - 3
PQFT	0.5770 - 3	0.6618 - 3	0.6104 - 4	0.6769 - 3
Rarity-G	0.6079 - 3	0.6739 - 4	0.6249 - 5	0.6892 - 3
Rarity-L	0.5782 - 3	0.6293 - 5	0.5965 - 5	0.6533 - 3
SDSR	0.5978 - 3	0.6479 - 7	0.6068 - 6	0.6913 - 4
SUN	0.5611 - 2	0.6516 - 4	0.6089 - 5	0.6664 - 3
Surprise	0.5813 - 4	0.6276 - 4	0.5878 - 4	0.6268 - 3
Torralba	0.5892 - 3	0.6712 - 4	0.6292 - 5	0.6892 - 4
Variance	0.5794 - 3	0.6510 - 4	0.6174 - 5	0.6659 - 4
VOCUS	0.5950 - 3	0.6455 - 5	-	0.6614 - 4
STB	0.5568 - 8	0.5858 - 5	0.5570 -6	0.6216 - 4
Yan	0.5806 - 1	0.6377 - 3	-	0.6834 - 2
Yin Li	0.5927 - 3	0.6489 - 2	-	0.6877 - 4
Gaussian	0.5	0.50	0.49	0.50
Human	**0.62**	**0.75**	**0.66**	**0.73**

Table 1. Eye movement prediction accuracy of compared saliency models using shuffled AUC score over four datasets. Numbers after dash shows size of the Gaussian kernel where a models take its maximum values (from 0.01 to 0.13 in steps of 0.01). For example, 4 for 0.04. Please note since NUSEF datasets contain stimuli that are not publicly and easily accessible due to copyright reasons, we avoid using them. Thus, results are reported over 414 images from NUSEF dataset.

Model	MIT	NUSEF	Toronto
AIM	0.2329 - 13	0.3168 - 13	0.3334 - 13
AWS	0.2455 - 5	0.3231 - 13	0.3899 - 3
E-Saliency	0.1984 - 13	-	0.3074 - 9
Bian	0.1931 - 13	0.3262 - 13	0.2628 - 13
Entropy	0.2079 - 13	0.14 - 2	0.3024 - 13
GBVS	0.2853 - 2	0.3750 - 4	0.4565 - 1
Kootstra	0.2666 - 4	-	0.4134 - 5
Tavakoli	0.2298 - 13	0.3556 - 8	0.5005 - 2
HouCVPR	0.2160 - 13	0.3095 - 13	0.3237 - 7
HouNIPS	0.2450 - 13	0.3244 - 13	0.3699 - 4
Itti	0.1881 - 13	0.2184 - 13	0.2756 - 13
Jia Li	0.1657 - 13	-	0.4406 - 1
Judd	0.2630 - 5	0.3412 - 12	0.4008 - 4
Le Meur	0.2216 - 13	-	0.3065 - 13
Marat	0.2448 - 3	-	0.3490 - 13
PQFT	0.2194 - 6	0.2771 - 13	0.2840 - 6
Rarity-G	0.2253 - 13	0.3046 - 13	0.3162 - 13
Rarity-L	0.2192 - 13	0.2903 - 13	0.3188 - 13
SDSR	0.2252 - 13	0.2876 - 13	0.3568 - 6
SUN	0.2192 - 13	0.2906 - 13	0.3027 - 13
Surprise	0.2019 - 13	0.2603 - 13	0.2708 - 13
Torralba	0.2314 - 13	0.3126 - 13	0.3298 - 13
Variance	0.2209 - 13	0.2946 - 13	0.3066 - 13
VOCUS	0.2262 - 13	-	0.3258 - 13
STB	0.1585 - 13	0.2012 - 13	0.2685 - 13
Yan	0.2544 - 12	-	0.4090 - 4
Yin Li	0.2290 - 13	-	0.3878 - 4
Gaussian	0.275	0.38	0.412

Table 2. Eye movement prediction accuracy of compared saliency models using Correlation Coefficient (CC) score over three datasets.

Model	Toronto	NUSEF	MIT
AIM	1.0664 - 13	1.0286 - 13	1.0710 - 13
AWS	1.3230 - 2	1.0589 - 13	1.1832 - 3
E-Saliency	1.0061 - 7	-	0.9307 - 13
Bian	0.8398 - 13	1.0706 - 12	0.9061 - 13
Entropy	0.9721 - 13	0.9622 - 13	0.9577 - 13
GBVS	1.5276 - 1	1.2369 - 4	1.3602 - 1
Kootstra	1.3470 - 4	-	1.2533 - 3
Tavakoli	1.6729 - 2	1.1731 - 8	1.0644 - 13
HouCVPR	1.0858 - 6	1.0185 - 13	1.0139 - 13
HouNIPS	1.2113 - 4	1.0980 - 13	1.1340 - 13
Itti	0.9075 - 13	0.7255 - 13	0.8953 - 13
Jia Li	1.4704 - 1	-	0.7620 - 13
Judd	1.3315 - 2	1.1168 - 10	1.2425 - 4
Le Meur	1.0114 - 1	-	1.0344 - 13
Marat	1.1539 - 4	-	1.1346 - 13
PQFT	0.9704 - 4	0.9130 - 13	1.0425 - 5
Rarity-G	1.0539 - 3	0.9940 - 13	1.0426 - 13
Rarity-L	1.0256 - 13	0.9504 - 13	1.0174 - 13
SDSR	1.1957 - 5	0.9453 - 13	1.0547 - 13
SUN	0.9684 - 13	0.9464 - 13	1.0094 - 13
Surprise	0.8787 - 13	0.8541- 13	0.9510 - 13
Torralba	1.0567 - 13	1.0172 - 13	1.0653 - 13
Variance	0.9864 - 13	0.9613 - 13	1.0209 - 13
VOCUS	1.0494 - 13	-	1.0516 - 13
STB	0.8899 - 13	0.6617 - 13	0.7601 - 13
Yan	1.3597 - 3	-	1.1934 - 8
Yin Li	1.3134 - 4	-	1.0687 - 13
Gaussian	1.6	1.15	1.3
Human	**2.4**	**1.92**	**3.1**

Table 3. Eye movement prediction accuracy of compared saliency models using Normalized Scanpath Saliency (NSS) over three datasets.

Feature	Saccade Length Histogram	Saccade Orientation Histogram	Saccade Duration Histogram	Saccade Velocity Histogram	Saccade Slope Histogram	Saliency Map	Fixation Duration Histogram	Fixation Histogram	Saliency Histogram at Fixation Locations	10 Top Salient Locations	All the Features
Acc	0.3430	0.2550	0.3320	0.3174	0.3064	0.3863	0.3377	0.4005	0.2161	0.2553	0.4322
Std	0.0260	0.0289	0.0319	0.0291	0.0340	0.0371	0.0283	0.0310	0.0209	0.0354	0.0161

Table 8. Performance of different features for stimuli decoding.

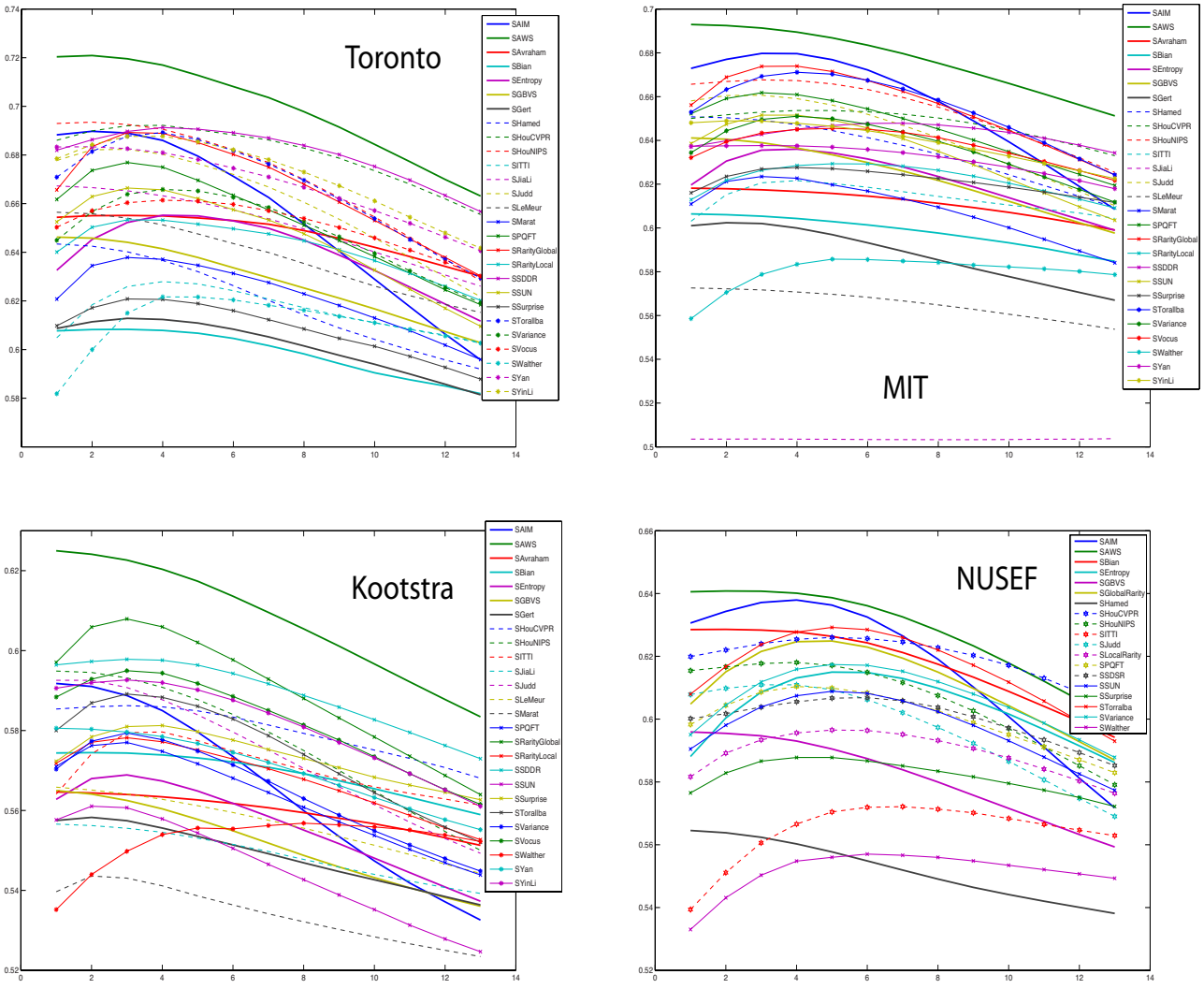

Figure 2. Shuffled AUC scores of saliency models over the range of smoothing over four eye-tracking datasets from 0.01 to 0.13 in steps of 0.01.

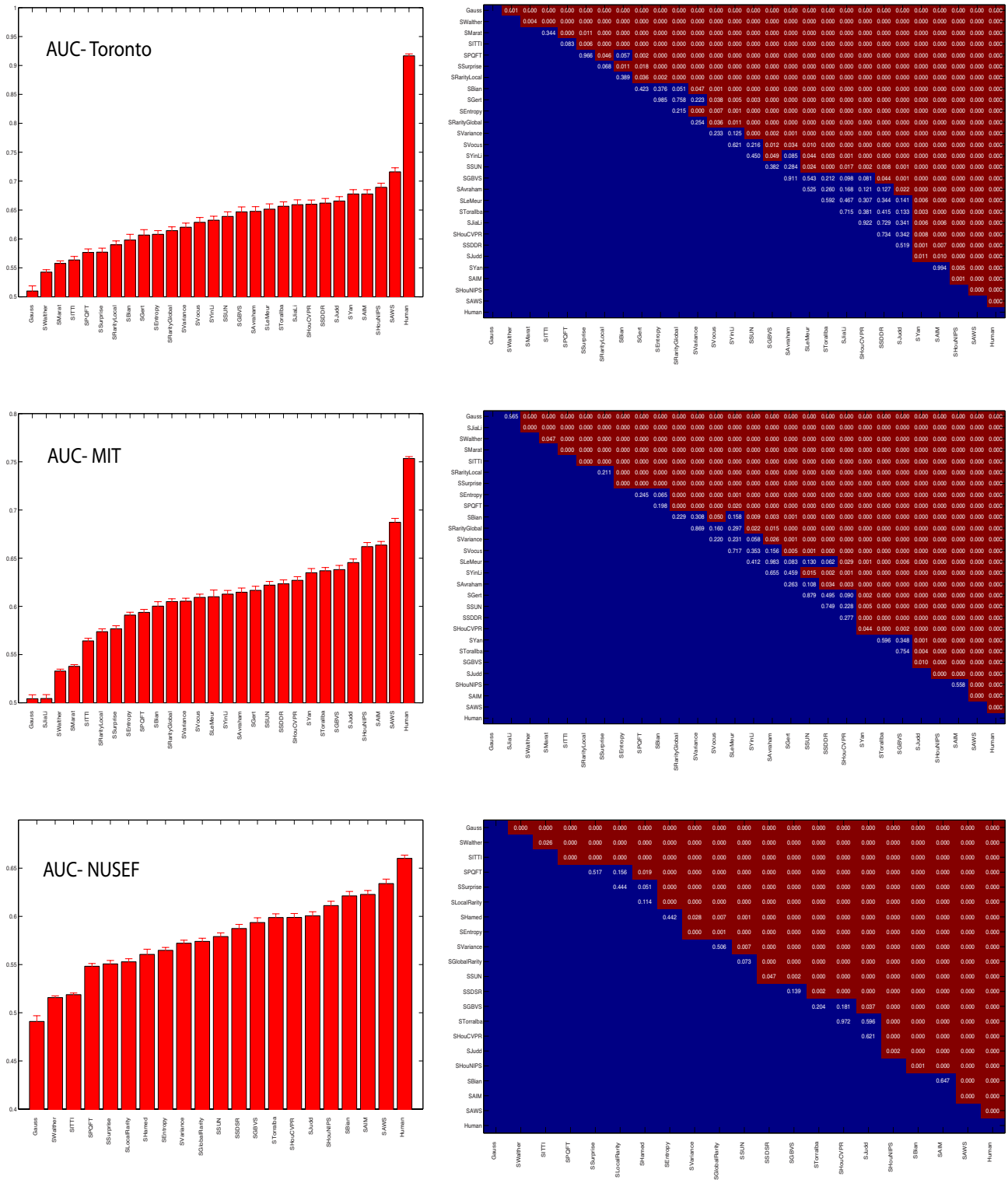

Figure 3. Left: sorted shuffled AUC scores of models over three datasets. Right: results t-test (95%, $p \leq 0.05$) comparison of models for all pairs (sorted). When reading this figure, please ignore the lower part. On the top-part red areas mean statistically significant while blue areas are not statistically significant. Only diagonal is reported in the main text.

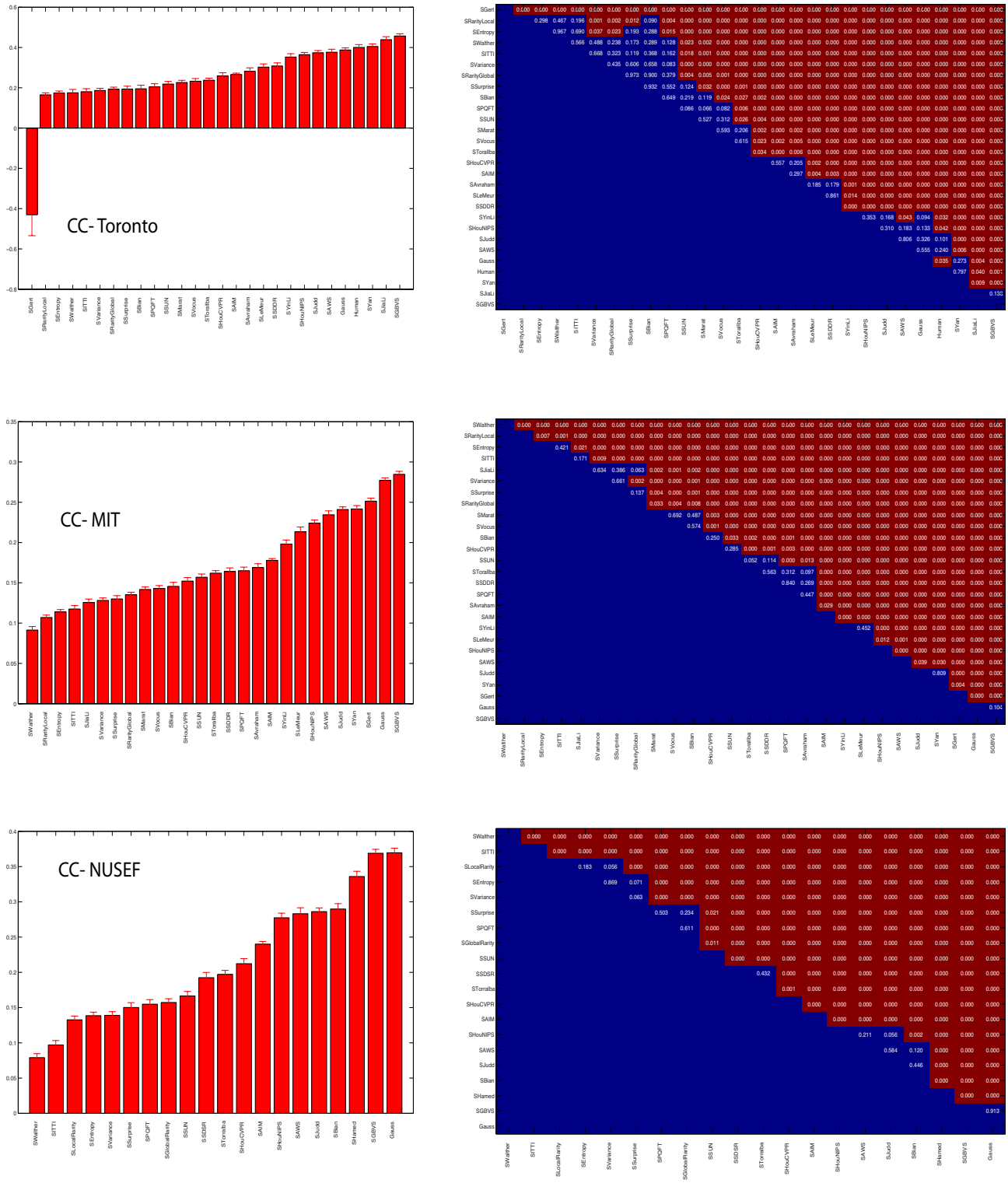

Figure 4. Left: sorted Correlation Coefficient (CC) scores of models over three datasets. Right: results of t-test (95%, $p \leq 0.05$) comparison of models for all pairs (sorted).

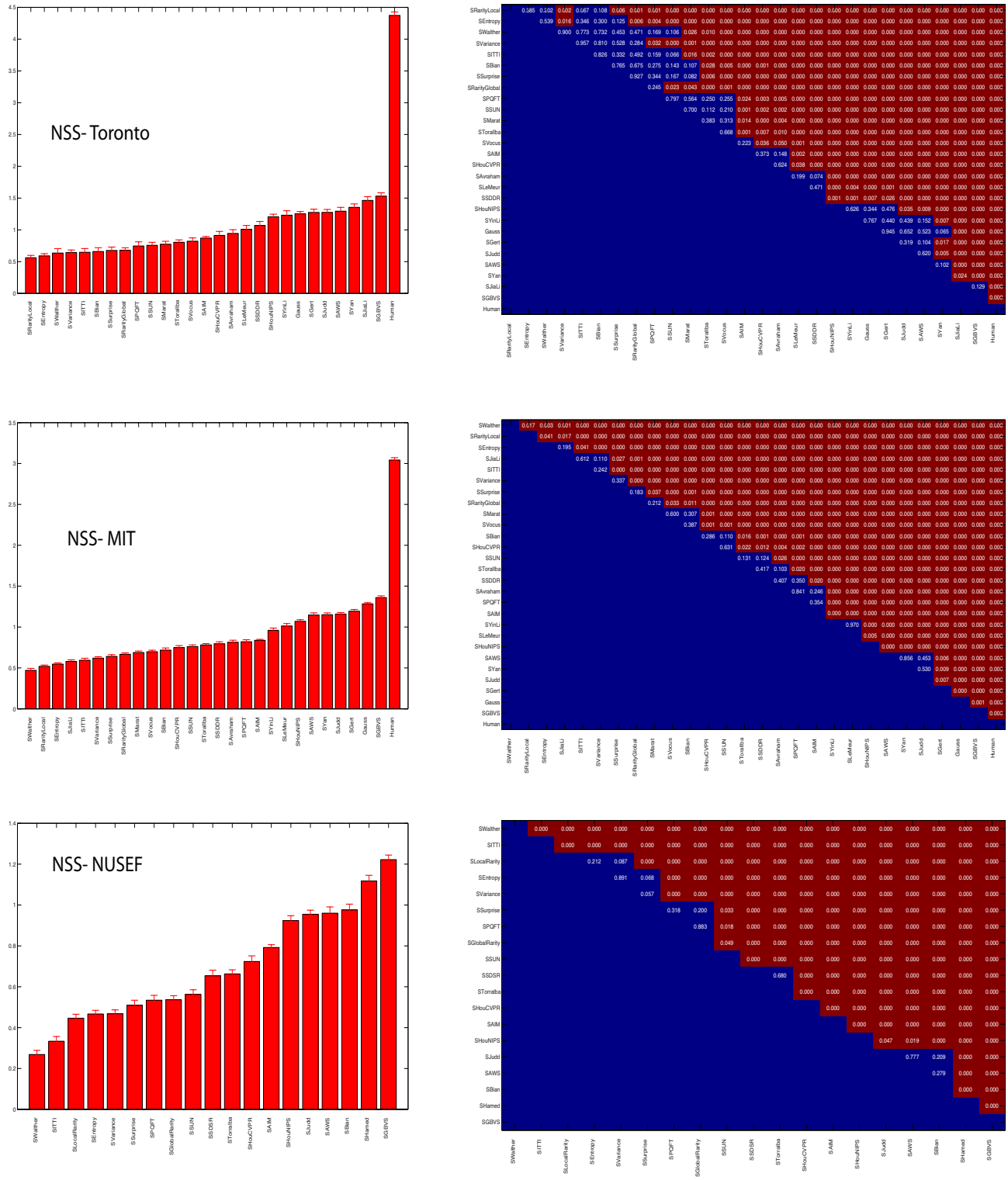

Figure 5. Left: sorted Normalized Scanpath Saliency (NSS) scores of models over three datasets. Right: results of t-test (95%, $p \leq 0.05$) comparison of models for all pairs (sorted).

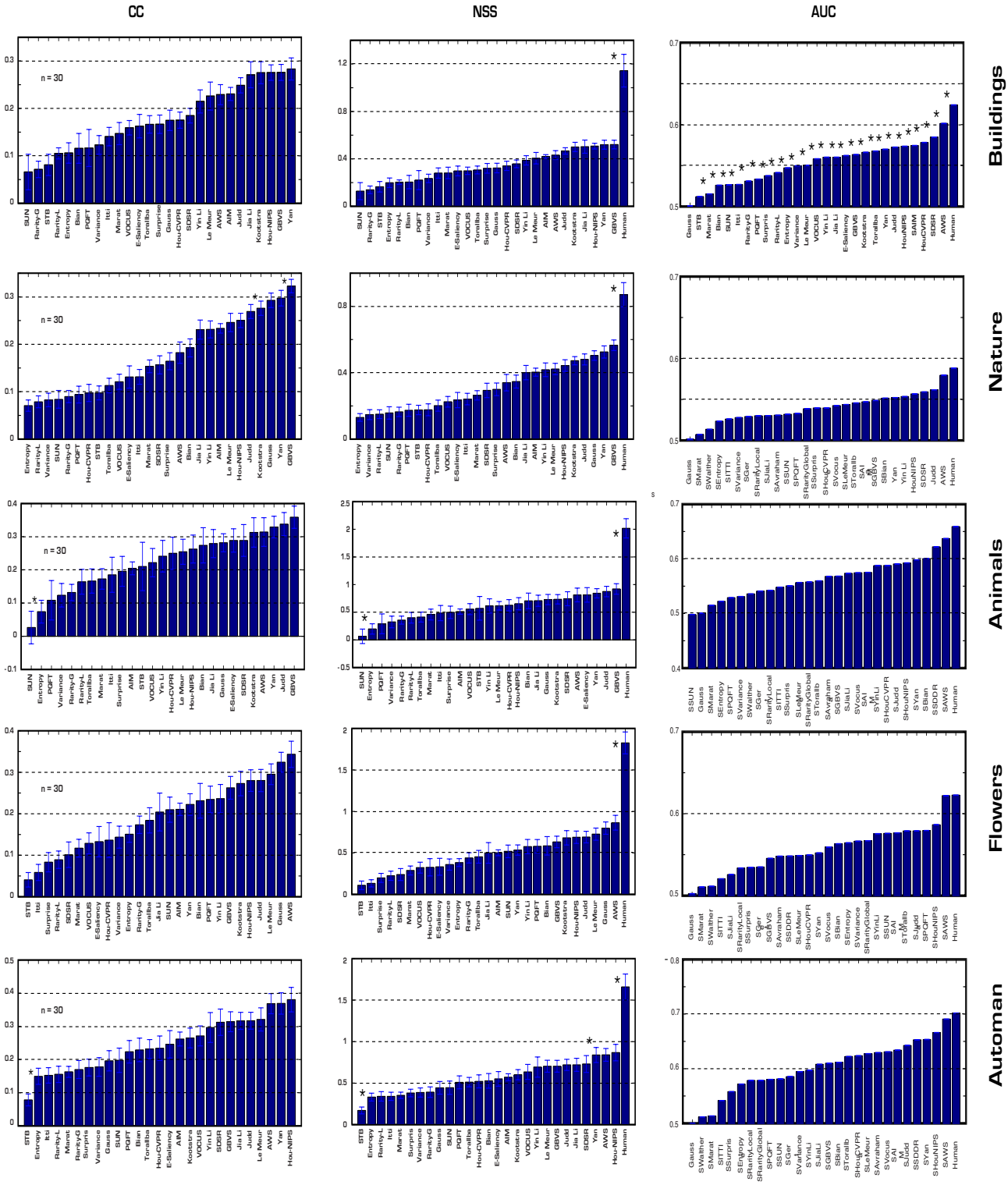

Figure 6. Model accuracies using CC, NSS, and sAUC scores over categories of Kootstra [5] dataset. N is equal 16, 41, 12, 20, and 12 for "buildings", "nature", "animals", "flowers", and "automan" categories, respectively.

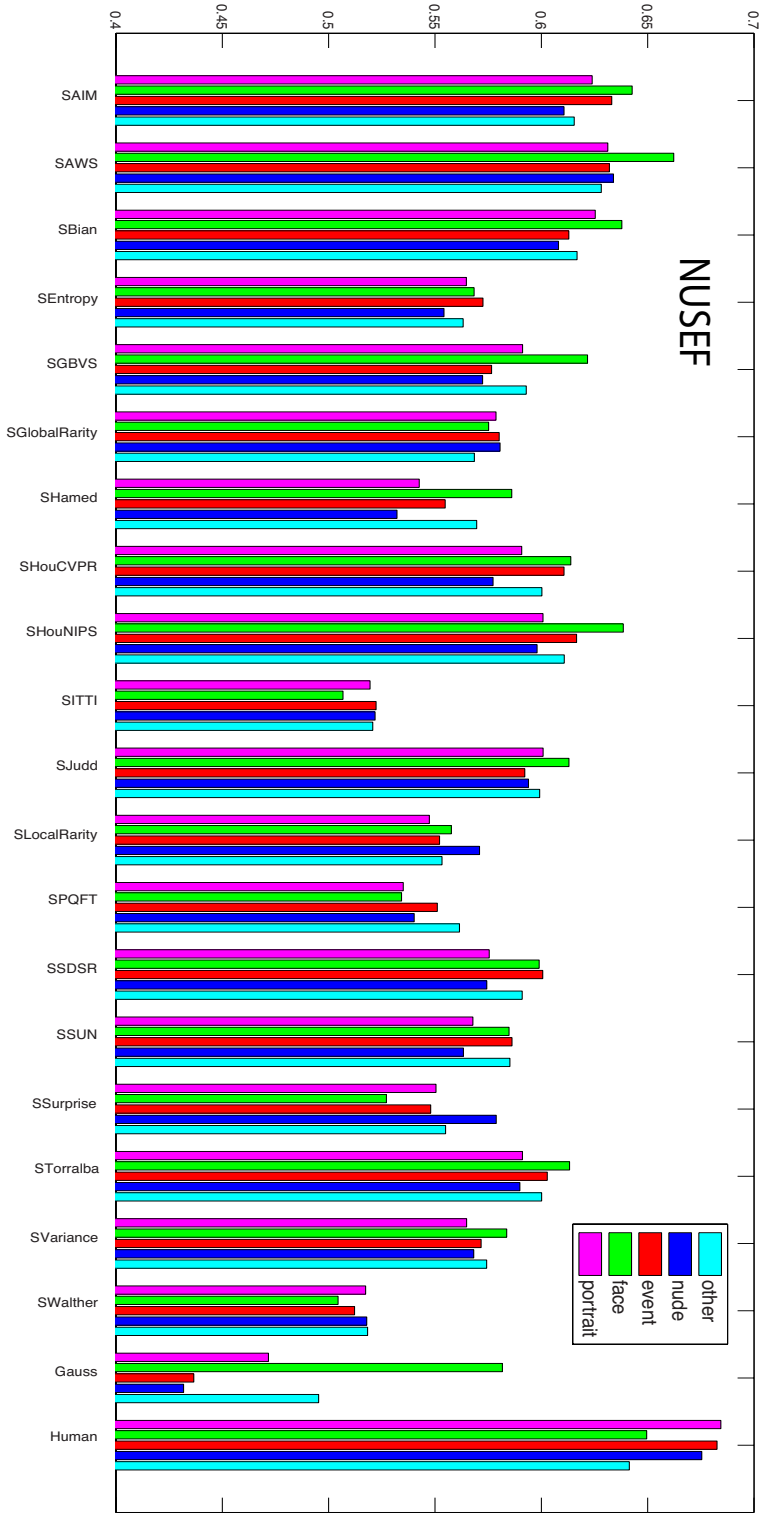

Figure 7. Model accuracies using sAUC score over categories of the NUSEF [3] dataset.

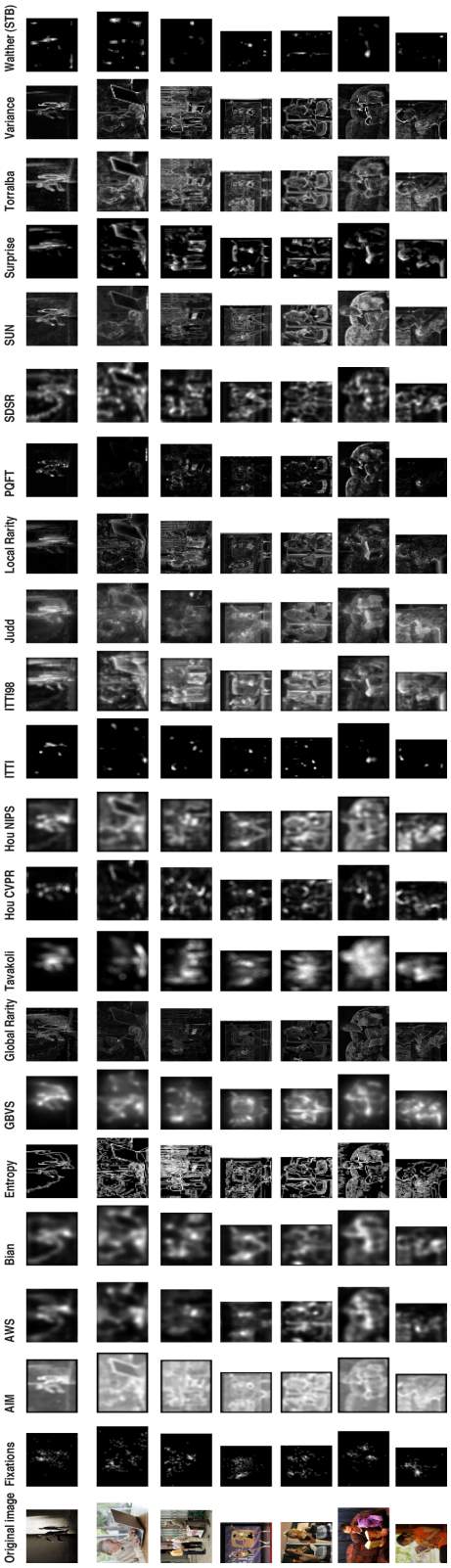

Figure 8. Sample images from “events“ category of NUSEF [3] dataset along with predictions of saliency models.

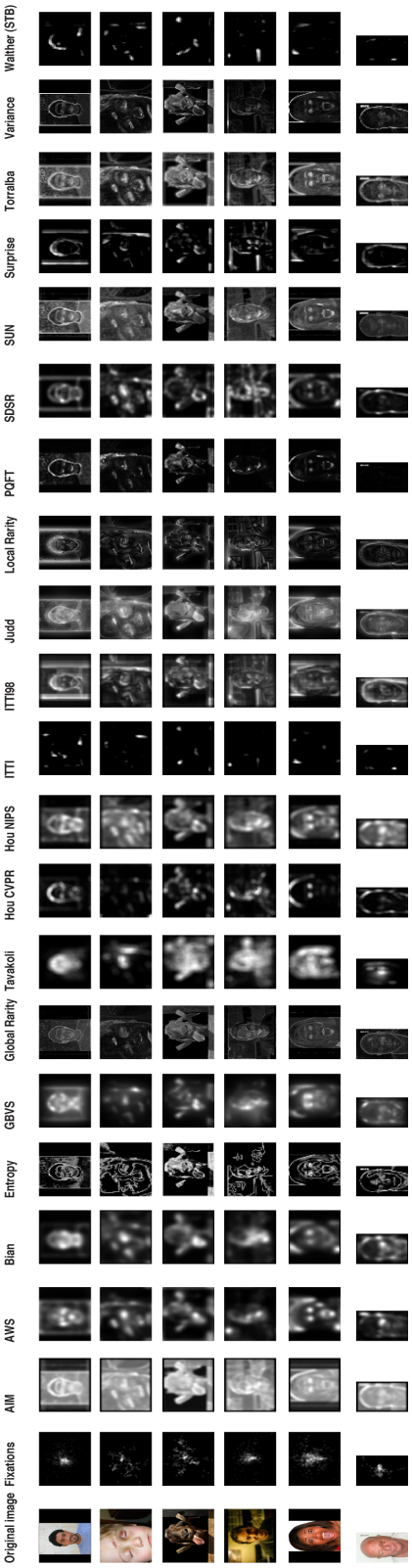

Figure 9. Sample images from “face“ category of NUSEF [3] dataset along with predictions of saliency models.

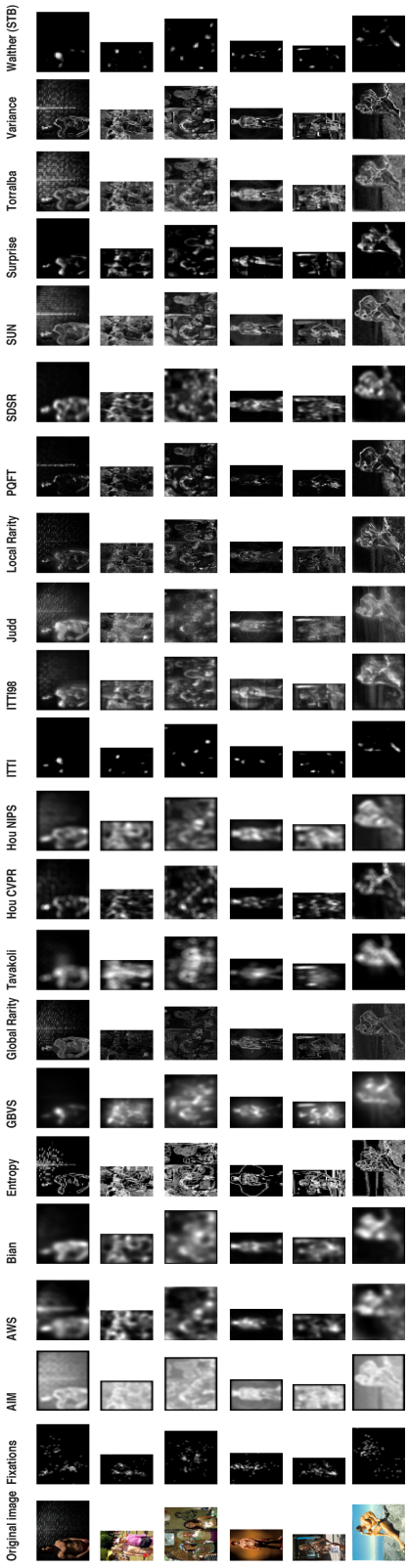

Figure 10. Sample images from “nude“ category of NUSEF [3] dataset along with predictions of saliency models.

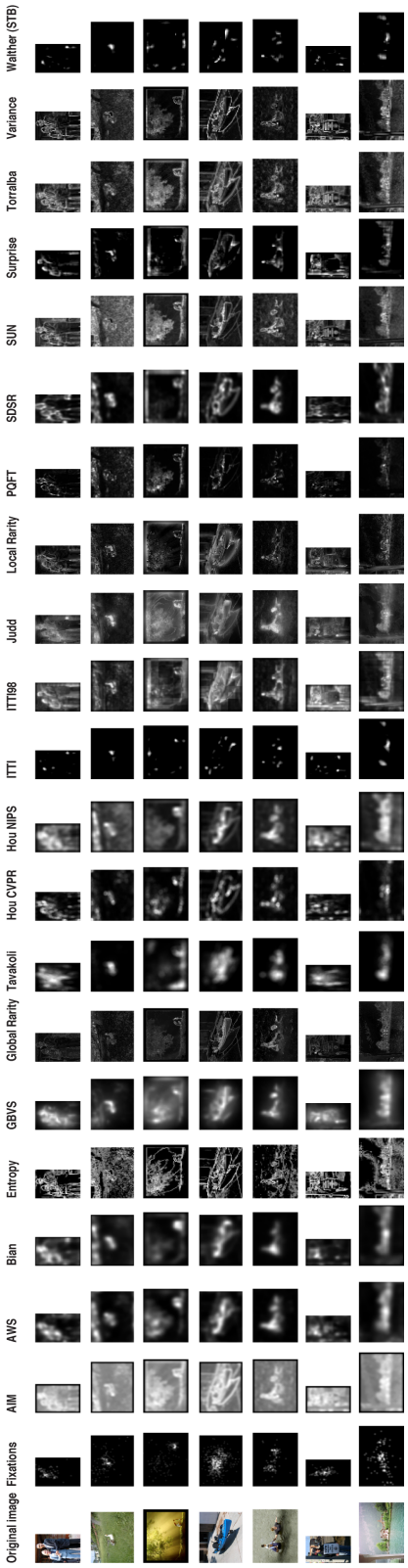

Figure 11. Sample images from “other“ category of NUSEF [3] dataset along with predictions of saliency models.

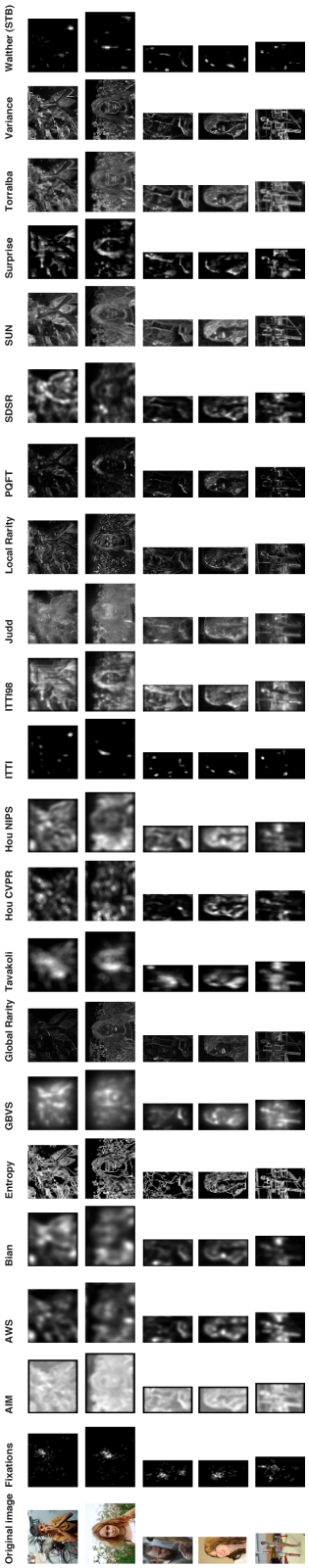

Figure 12. Sample images from “portrait“ category of NUSEF [3] dataset along with predictions of saliency models.

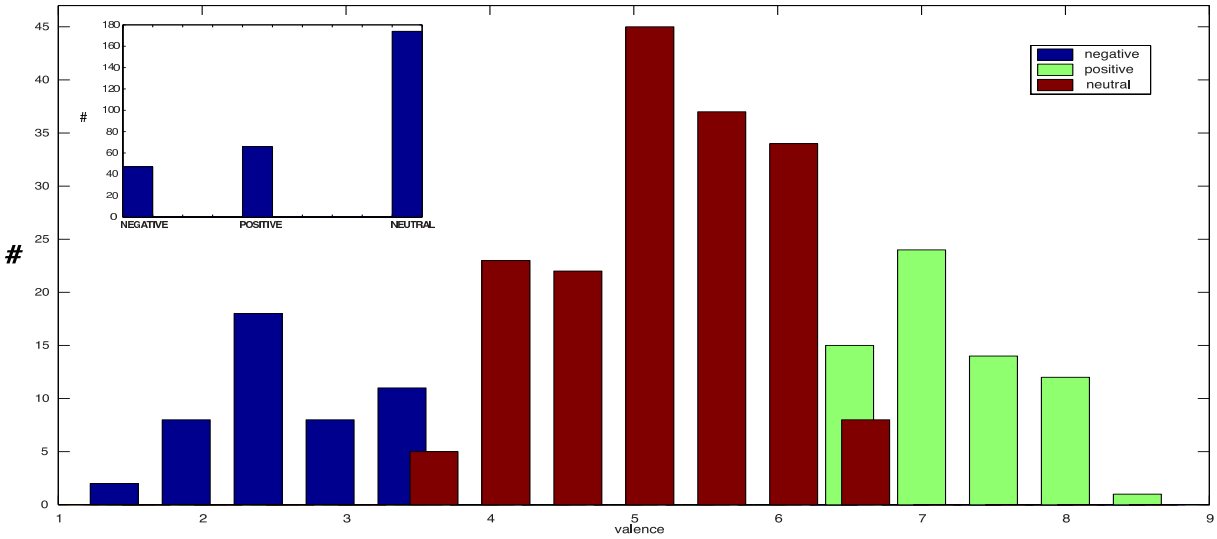

Figure 13. Histogram of emotional valence for affective stimuli from the NUSEF dataset. Inset shows the histogram of images labeled as *negative, positive, and neutral* in this dataset.

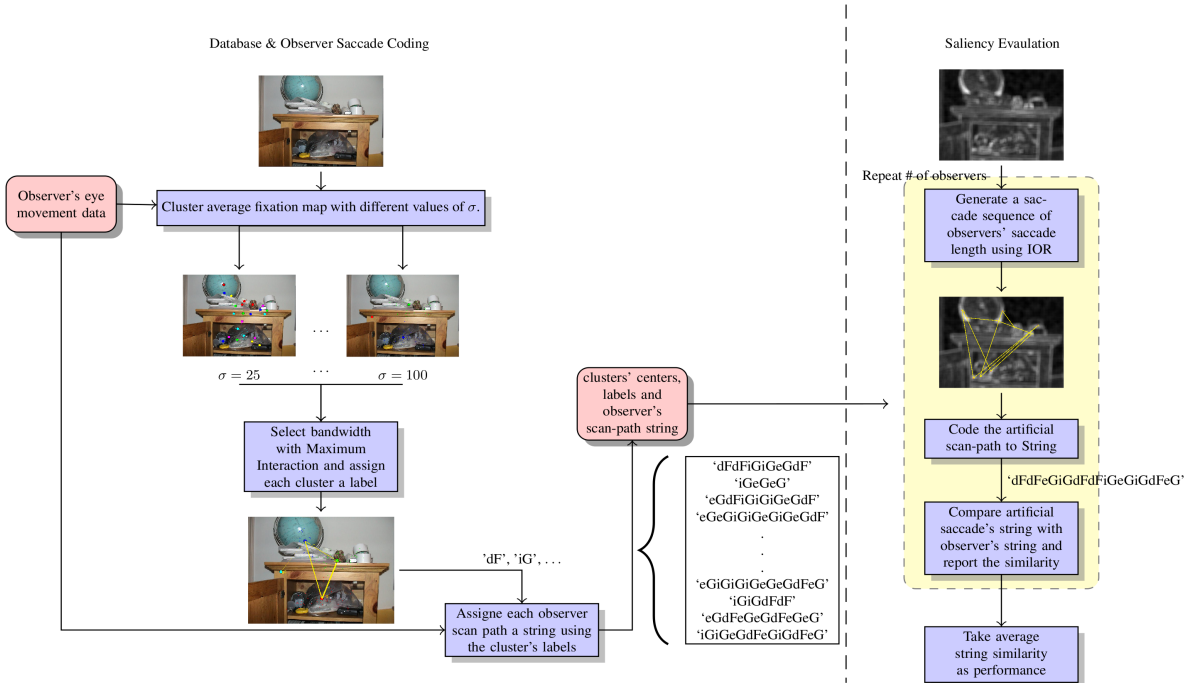

Figure 14. Illustration of the scanpath coding and evaluation for humans and models.