Do Saliency Models Detect Odd-One-Out Targets? New Datasets and Evaluations

Iuliia Kotseruba
yulia_k@eecs.yorku.ca
Calden Wloka
cwloka@eecs.yorku.ca
Amir Rasouli
aras@eecs.yorku.ca
John K. Tsotsos
tsotsos@eecs.yorku.ca

Department of Electrical Engineering and Computer Science
York University
Toronto, Canada

Abstract

Recent advances in the field of saliency have concentrated on fixation prediction, with benchmarks reaching saturation. However, there is an extensive body of works in psychology and neuroscience that describe aspects of human visual attention that might not be adequately captured by current approaches. Here, we investigate singleton detection, which can be thought of as a canonical example of salience. We introduce two novel datasets, one with psychophysical patterns and one with natural odd-one-out stimuli. Using these datasets we demonstrate through extensive experimentation that nearly all saliency algorithms do not adequately respond to singleton targets in synthetic and natural images. Furthermore, we investigate the effect of training state-of-the-art CNN-based saliency models on these types of stimuli and conclude that the additional training data does not lead to a significant improvement of their ability to find odd-one-out targets.

1 Introduction

The human visual system processes vast amounts of incoming information and performs multiple complex visual tasks with perceived efficacy and ease. However, it is well known that not only are humans capacity-limited perceivers [11], but also that the vision problem in general is computationally intractable [56]. Instead of inspecting every element of a scene, humans use a set of attention mechanisms to prioritize and filter stimuli from the early visual processing stages through to the higher visual areas. The ability to use perceived saliency of objects for efficient scanning of the scene is one of the fundamental attention mechanisms.

Computational approaches to predicting human judgements of saliency have resulted in an extensive corpus of saliency models. The evaluation for these models in recent years has been conducted using a number of metrics for fixation prediction [7], however, such heavy focus on gaze prediction fails to address extensive research in psychology and neuroscience which has catalogued and quantified many aspects of human visual attention [15].

One area well studied in psychology is visual search, frequently associated with Feature Integration Theory (FIT) [54]. FIT posits that items that differ significantly in one feature

dimension are processed in a parallel manner, whereas an item uniquely defined by a combination of features requires a serial search through items. Wolfe [66, 70], on the other hand, notes that the parallel/serial dichotomy of FIT is better represented as a continuum of search efficiencies. In other words, if the difference between the target and surrounding distractors is sufficiently large, the target will be immediately detected ("pop-out"), otherwise as the target-distractor difference decreases search becomes increasingly inefficient [62, 67].

Visual search was a major guide in the early formulation of saliency models (*e.g.* [32, 37, 44]). Thus, given its conceptual importance to the topic of saliency, here we propose two new datasets with synthetic and natural odd-one-out targets and a large-scale comparative study that systematically evaluates the responses of classical and deep-learning saliency models to these types of stimuli. These results not only identify and quantify remaining gaps between human and algorithmic performance in visual search, but also have significant practical ramifications. Currently, saliency models are being applied in commercial areas, such as the design of marketing materials (*e.g.* [1, 42, 45, 57, 60]), where strong discrimination of specific information or products is a clearly stated goal. Thus the inadequate performance of saliency models on a range of realistic odd-one-out targets revealed by our evaluations highlights urgent new research avenues for these application areas.

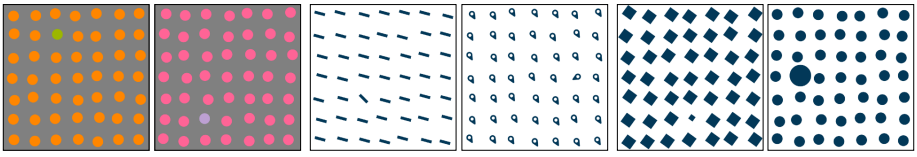
1.1 Relevant Works

Psychophysical Evaluation of Saliency Models. Although a number of saliency models present qualitative results on a small set of synthetic stimuli, such as color/orientation singletons, Q/O and visual search asymmetries [14, 16, 20, 24, 27, 29, 30, 38, 40, 43, 46, 58, 74], only a few have been systematically evaluated on the psychophysical patterns used in human experiments. For instance, IKN has been evaluated on color/orientation pop-out search and conjunctive search [31], AWS and Discriminant Saliency were tested on non-linearity of pop-out for orientation and various search asymmetries [22, 23].

Bruce et al. [15] discussed a range of psyhological factors influencing visual saliency and their modeling implications. Additionally, several studies systematically compared the performance of multiple saliency algorithms against one another and human gaze data. Borji et al. [10] conducted a quantitative analysis of 35 classical saliency models over 54 synthetic patterns from the CAT2000 dataset [9] and compared them to human fixation data. Wloka et al. [61] compared human response time data against a set of classical saliency models on singleton search with oriented bars. More recently, Berga et al. [7] evaluated a large set of models (both classical and deep) on CAT2000/Pattern image set and synthetic patterns [8].

Together these studies suggest that no existing saliency model accounts for most of the identified saliency phenomena, although there is evidence that models based on the FIT framework [54] may perform better on synthetic images [10]. While pointing to the same conclusion, these studies are difficult to compare directly as they all use different metrics, sets of algorithms and datasets for evaluation.

Psychophysical Patterns and Odd-One-Out Images in Datasets. Datasets typically used in saliency research include only a handful of synthetic psychophysical patterns, if any. For instance, the Toronto dataset [13] includes 16 such cases out of 136 images, MIT1003 [34] has 4 patterns out of 1003 and CAT2000/Patterns image set [9] has 71 synthetic patterns out of 2000. The recently proposed SID4VAM dataset [8] provides human fixation data for 230 synthetic patterns covering 15 different stimuli types. However, since human participants were given task instructions, this data is not directly comparable with the free-viewing conditions under which saliency models are more typically tested.



(a) (b) (c)

Figure 1: Sample images from P^3 with (a) color, (b) orientation and (c) size singletons.

Odd-one-out images, which can be thought of as realistic counterparts to psychophysical patterns (see Figure 2), are also rarely found in fixation datasets. Out of 9 datasets that we examined none contained a significant portion of these types of stimuli, *e.g.* only 3 such images in the MIT1003 dataset (1003 images), 19 in DUT-OMRON (5168 images) [71] and 33 in SALICON (20K images)[33] (see the supplementary material for a complete list).

Contributions. We address the research challenges described above by providing the four following contributions: 1) a large dataset of synthetic search arrays, analysis and evaluation of saliency models; 2) a large annotated dataset of images with odd-one-out objects captured "in the wild"; 3) extensive analysis of saliency algorithms with respect to several features known to guide human attention; 4) experiments showing the impact of augmenting training data with synthetic and/or real images on the performance of the CNN-based models on odd-one-out task and fixation prediction.

2 Psychophysical Patterns (P^3) Dataset

We propose a large dataset of psychophysical patterns (P^3)[1] for evaluating the ability of saliency algorithms to find singleton targets. We focus on color, orientation, and size: three features recognized as undoubted attention guiding attributes [70] (see Figure 1)

All visual search arrays in the P^3 dataset are on a regular 7×7 grid with distractor size 75 px ($\approx 2^\circ$) and target size between 18 px (0.5°) and 140 px (4°). The target location in the grid is selected at random. Jitter of 15 px is applied to each element (as recommended in [2] to avoid perceptual grouping effects). The size of images (1024×1024) matches the setup of MIT300 [17] ($\approx$35 px per degree of visual angle and image sizes of 768×1024). Each image has a corresponding ground truth mask for the target and distractors.

There are a total of 2589 images in P^3 including 885 color, 864 orientation and 840 size search arrays. Further details on how the images are generated can be found in the supplementary material.

3 Odd-One-Out (O^3) Dataset

In addition to the psychophysical arrays, we propose a dataset of realistic odd-one-out stimuli gathered "in the wild". Each image in the novel Odd-One-Out (O^3) dataset[1] depicts a scene with multiple objects similar to each other in appearance (distractors) and a singleton (target) which usually belongs to the same general object category as distractors but stands out with

[1] http://data.nvision2.eecs.yorku.ca/P3O3/

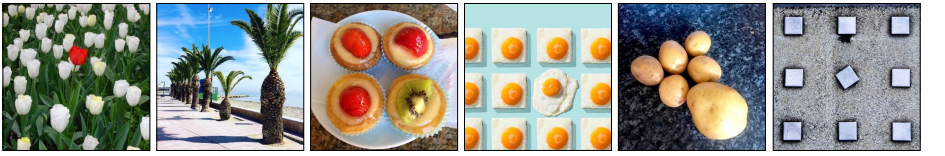


Figure 2: Sample images from the O^3 dataset with singletons in various feature dimensions. From left to right: color, size, color/texture, shape, size, orientation.

respect to one or more feature dimensions (e.g. color, shape, size) according to the evaluation of three human annotators (see samples in Figure 2).

The O^3 dataset consists of 2001 images with the larger image dimension set to 1024 to match the setup of MIT300, as most saliency models used for evaluation are optimized for it. Each image is annotated with segmentation masks for targets and distractors as well as text labels for the following: type of the target object, number of distractors and pop-out features (color, pattern/texture, shape, size, orientation, focus and location).

Targets in the O^3 dataset represent nearly 400 common object types such as *flowers*, *sweets*, *chicken eggs*, *leaves*, *tiles* and *birds*. Due to the importance of color as an attention guiding feature [68] the dataset contains many color singletons. Specifically, 37% of all targets differ from the distractors by color alone and 47% differ by color and one or more additional feature dimension. In addition, 33% of the targets are distinct in texture, 26% in shape, 19% in size and 8% in orientation. The targets vary considerably in size: approximately 80% of the targets occupy between 7% and 20% of the image by area, 30% of the targets are larger and few take up to 80% of the image. All targets are accompanied by at least 2 distractors, with half of the targets surrounded by at least 10 distractors and 10% by over 50 distractors.

4 Experiments

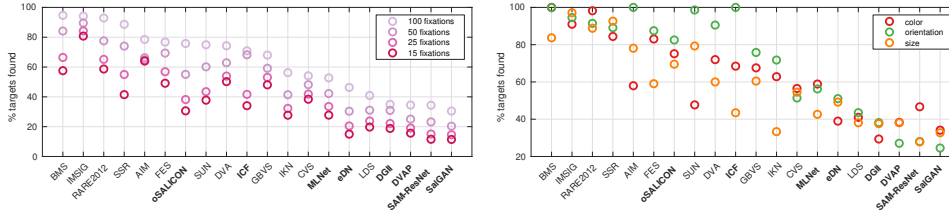
4.1 Metrics

Reaction time (RT) is a primary metric of human performance reported in the psychophysical literature (*e.g.* [2, 65, 66]). Although RT cannot be directly measured from saliency maps, there is a link between reaction time and the relative salience of targets and distractors [53, 72]. We therefore use several proxy measures: the number of fixations before reaching the target, global saliency index (GSI), and the saliency ratio (SR).

Number of fixations. To produce fixations from a saliency map we iterate through maxima and suppress attended locations with a circular mask (as in [53]) until a location near the target is reached or a maximum number of fixations is exceeded. The circular mask has a 1° radius which corresponds to the size of distractors in all images.

There is no definition of how close to the target the fixation should land in order to be considered as a hit. Values reported in the literature vary from 1° [3, 4, 5] to 2° [55] or up to 2.35° [6] radius around the center of the target. In our experiments we set the radius to 1° around the center of the target and up to 2° for larger stimuli in *size* patterns.

Global Saliency Index. The metric is defined as $GSI = \frac{S_{target} - S_{distr}}{S_{target} + S_{distr}}$, where S_{target} and S_{distr} are the average saliency values within the target and distractor masks respectively [48, 49].



(a) % of detected targets vs number of fixations (b) % of detected targets vs singleton feature

Figure 3: a) Number of fixations vs % of targets detected. b) Performance on color, orientation and size singletons at maximum of 100 fixations. Models are sorted by the % of targets detected at maximum of 100 fixations. Labels for deep models are shown in bold.

GSI measures how well the target is distinguished from the distractors and varies between -1 (target saliency is 0, only distractors are highlighted) to 1 (only the target is highlighted).

Saliency Ratio. Due to the highly varying target-distractor sizes and configurations in real pop-out images in the O^3 dataset, both the GSI and fixations-to-target metrics become challenging to apply. We, therefore, use a more straightforward ratio of maximum saliency of the target vs maximum saliency of the distractors as in [61] and the same for the background vs target, referred to in the text as MSR_{targ} and MSR_{bg} respectively. Since most practical applications require locations of maxima within saliency maps, these two metrics help determine whether the target is more salient than the surrounding distractors and background.

4.2 Saliency Models and Datasets

For all experiments that follow we use 12 classical saliency models, i.e. theory-driven models with clear semantic interpretation (AIM [12], BMS [73], CVS [20], DVA [28], FES [61], GBVS [26], IKN [32], IMSIG [29], LDS [21], RARE2012 [43], SSR [46] and SUN [74]), and 8 deep learning models, which are primarily data-driven (DGII [39], DVAP [59], eDN [58], ICF [40], MLNet [18], oSALICON [30, 52], SalGAN [41] and SAM-ResNet [19]).

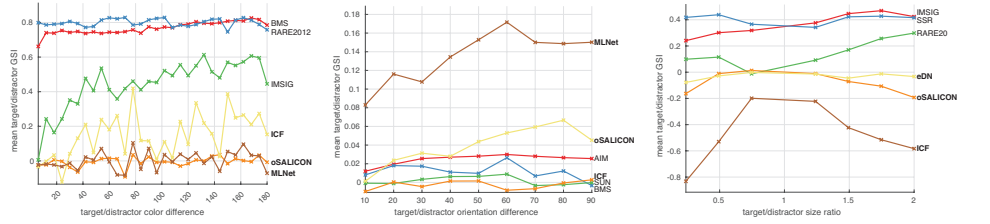
We use the SMILER framework [64] to run all models without center bias on P^3 and with center bias on O^3. Default values for other parameters (including smoothing) were kept. We evaluate all models on P^3, O^3 and CAT2000/Pattern.

4.3 Evaluation of Saliency Models

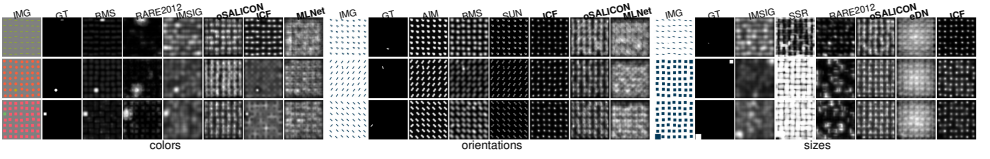
4.3.1 Psychophysical Evaluation of Saliency Models

Evaluating models on psychophysical search arrays allows us to measure their response to basic attention guiding features and determine whether they capture known properties of the human visual system.

Target detection rate. To measure the strength of attention guidance we count the total number of targets detected and the average number of fixations required to do so. The relationship between these quantities is shown Figure 3(a). When the maximum number of fixations is set to 100 (or 2× the number of elements in the grid), several classical algorithms (BMS, IMSIG, and RARE2012) achieve success rate above 90% and only one deep-learning model (oSALICON) comes close to the 75% mark. Note that the performance of nearly all models degrades quickly when the allowed number of fixations is reduced.



(a) color difference (b) orientation difference (c) size ratio



(d) saliency maps

Figure 4: The discriminative ability (GSI score) of the top-3 classical and deep models for a range of TD differences in color (a), orientation (b) and size (c) feature dimensions. The models are selected based on the total number of targets found within 100 fixations in each dimension. d) Sample saliency maps for each model. TD difference for color and orientation targets increases from top to bottom row. Size targets range from the smallest to largest.

The same evaluation applied to human fixation maps for 51 singleton patterns from the CAT2000/Pattern yields 98% target detection rate and 13 fixations on average. Note that the CAT2000 data is not directly comparable with P^3 in image size and stimuli used. However, even when the maximum number of fixations is set to 25 (equivalent to examining half of the elements in the grid), most saliency algorithms miss $>$ 50% of the targets in the P^3 dataset and only one model, IMSIG, detects slightly more than 80% of singletons.

Response to simple features. As shown in Figure 3(b), the performance of most models varies significantly depending on the type of stimuli they are applied to. There are, however, several exceptions, namely IMSIG, SSR, CVS and LDS, which perform equally well or equally poorly on all three types of singletons. In general, orientation and color targets appear to be easier to detect for many of the models and especially for those models that process these features explicitly, *e.g.* RARE2012, BMS, AIM. Size targets are missed more often, even by the models that do multi-scale processing (*e.g.* RARE2012, IKN, MLNet).

Sensitivity to target-distractor similarity. It has been established that visual search efficiency improves as the target-distractor (TD) difference increases [59]. Above a certain value, the pop-out effect can be observed, i.e. the target is found in nearly constant time (or only few fixations) regardless of the number of the distractors.

The plots in Figure 4 visualize the discriminative ability of saliency algorithms for a range of TD differences in color, orientation and size. Due to the space limitations, we only show and discuss the 3 best classical and deep models for each feature type. Top models are selected based on the % of the detected targets of each feature type. The plots with GSI measures and fixation counts for all models can be found in the supplementary material.

Color singletons. Among the classical models, only IMSIG has a resemblance to human behavior on color stimuli (Figure 4(a)). It has more difficulty distinguishing similar colors and gradually improves as the TD difference increases. However, the behavior of IMSIG is not consistent as indicated by the dips in the GSI plot. Two other classical algorithms, BMS

and RARE2012, nearly perfectly find even the smallest differences in hue and detect targets in fewer than 10 fixations on average. Note also very clear discrimination of the targets in the saliency maps (see Figure 4(d)). ICF, the only CNN-based model that explicitly includes color contrast features, shows a tendency of improving on larger TD differences but is less consistent (note the spikes around 80° and 150°). oSALICON and MLNet both have low GSI scores for all TD differences and need more than 50 fixations on average to find even the most salient targets.

Orientation singletons. The GSI scores for orientation targets are nearly an order of magnitude lower than scores for color targets as shown in Figure 4(b). Here, the top-3 classical models, AIM, SUN and BMS, do not discriminate the targets well. Only AIM has a GSI score consistently above 0 and takes only a few fixations on average to reach the target. SUN and BMS have less consistent GSI scores and perfom $\approx$ 30 fixations to find targets regardless of TD difference. Deep models MLNet and oSALICON detect more distinct targets better (with spikes around 60° and 70°) but require at least 40 fixations on average to reach the target.

Size singletons. Similar to color and orientation features, we expect that the targets closest to the size of the distractors would be more difficult to find, while very small and very large targets should 'pop' easily. Unlike orientation and color, most classical algorithms do not encode size explicitly as a feature but may handle it indirectly via multi-scale processing. Here, IMSIG, SSR and RARE2012 exhibit anticipated behavior to some extent, having higher GSI for larger TD differences (Figure 4(c)). Interestingly, deep learning models eDN and oSALICON demonstrate almost the opposite of the human behavior as their GSI scores drop for larger targets. Another deep model, ICF, has negative GSI scores for the whole range of TD differences, meaning that the target is consistently less salient than the distractors on average, as can be seen in the sample saliency maps shown in Figure 4(d).

Based on the evidence presented we conclude that the majority of saliency models are not guided by attributes such as color, orientation and size which have an undoubted impact on human attention [70]. Several classical algorithms (*e.g.* IMSIG, RARE2012 and AIM) are guided by some of the basic features, but none of them respond to all three features considered. In comparison to classical models, deep algorithms have lower discriminative ability and are less consistent in finding singletons.

4.3.2 Evaluation of Saliency Models on Odd-One-Out Data

Unlike the psychophysical evaluation that focused on how well the saliency models capture fundamental properties of the human visual system, testing models on the O^3 data aims to examine their behavior in more realistic scenarios. To the best of our knowledge, ours is the first work that addresses detection of singletons by saliency algorithms in realistic stimuli.

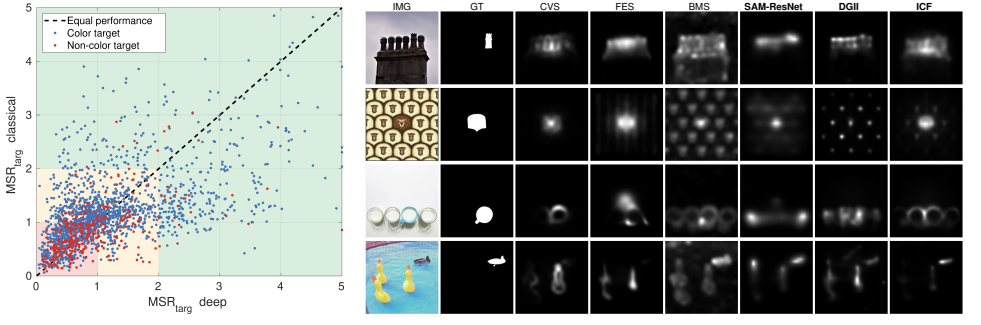
Due to space limitations, we report results only for the top-3 classical and deep models (Table 1) and present complete results in the supplementary material. Given that the majority of targets in the O^3 dataset differ in several feature dimensions we do not list all possible combinations and discuss only subsets which do or do not differ in color since the presence of a color feature dominates other features during visual search [35].

All models discriminate color targets better than targets that differ in other feature dimensions, as indicated by low MSR_{targ} values for non-color singletons in Table 1 (see sample images and saliency maps in Figure 5(b)). As shown in Figure 5(a) the majority (80%) of non-color targets have an average $MSR_{targ} < 1$ compared to only 30% of color targets.

Classical and deep models discriminate 55% and 45% of all targets respectively. Deep

Model	Color targets		Non-color targets		All targets	
	MSR_{targ}	MSR_{bg}	MSR_{targ}	MSR_{bg}	MSR_{targ}	MSR_{bg}
SAM-ResNet	1.47	1.46	1.04	1.84	1.40	1.52
CVS	1.43	2.43	0.91	4.26	1.34	2.72
DGII	1.32	1.55	0.94	1.95	1.26	1.62
FES	1.34	2.53	0.81	5.93	1.26	3.08
ICF	1.30	2.00	0.84	2.03	1.23	2.01
BMS	1.29	0.97	0.87	1.59	1.22	1.07

Table 1: The top-3 classical and deep models (shown in bold) evaluated on the O^3 dataset.



(a) (b)

Figure 5: a) The mean MSR_{targ} of the top-3 classical and deep models for the color and non-color targets in O^3 (shown as blue and red dots respectively). Along the dashed line the performance of the models is equal. The red, yellow and green colors show areas where targets are not discriminated ($MSR_{targ} < 1$), somewhat discriminated ($1 \leq MSR_{targ} \leq 2$) and strongly discriminated ($MSR_{targ} > 2$). b) Sample images and corresponding saliency maps. From top to bottom row: hard for both, easy for both, classical models perform better, deep models perform better.

models strongly discriminate only 15% of the images (MSR_{targ} is > 2) compared to 10% for classical models. However, for all models the average MSR_{bg} scores are higher than the MSR_{targ} scores (Table 1), i.e. saliency maxima in the background are higher than the maxima within the target mask meaning that these models get distracted by the background.

We also observe that most classical models perform better on P^3 with few exceptions such as CVS and IMSIG. For instance, BMS, which by design handles color features well, excells on P^3 and also benefits from an abundance of color targets in O^3. For deep learning models the opposite is true: models in general show much better results on O^3. In fact, SAM-ResNet which ranks first on O^3 (see Table 1) ranks second to last on P^3 (see Figure 3(a)). One possible reason for such discrepancy is that SAM by design learns a strong center bias prior which works against it on P^3, where target distribution is uniform, but is an advantage on natural image datasets like O^3 which are usually more center-biased. Another consideration is that SAM and most other state-of-the-art saliency algorithms rely on features produced by a backbone CNN pre-trained on object classification task. While deep learning models capture high-level features more often associated with human gaze locations (e.g., faces, text), it may be the case that training data does not capture some of the fundamental properties exemplified in psychophysical stimuli. On the other hand, many classical models explicitly define salient regions as different from their surroundings based on a set of features

Model	Training Data	MIT1003					P^3		O^3	
		AUC_Judd	CC	KLDiv	NSS	SIM	Avg. num. fix.	% found	MSR_{targ}	MSR_{bg}
MLNet	SALICON	0.82	0.47	1.3	1.7	0.35	40.97	0.46	0.92	0.98
	P^3	0.51	0.04	12.07	0.13	0.18	51.81	0.44	0.78	2.77
	O^3	0.74	0.29	1.83	1.01	0.3	41.32	0.46	1.01	0.91
	SALICON+P^3	0.82	0.47	1.36	1.69	0.36	40.01	0.45	0.93	0.97
	SALICON+O^3	0.82	0.46	1.33	1.65	0.34	40.87	0.46	0.97	0.91
	SALICON+P^3+O^3	0.82	0.46	1.36	1.64	0.35	42.00	0.44	0.96	0.91
SALICON	OSIE	0.86	0.6	0.92	2.12	0.48	57.12	0.68	0.88	1.16
	P^3	0.57	0.08	2	0.31	0.23	51.60	0.66	0.90	1.32
	O^3	0.77	0.32	1.52	1.08	0.34	45.37	0.70	1.05	2.12
	OSIE+P^3	0.83	0.52	1.1	1.87	0.42	50.23	0.64	0.87	1.27
	OSIE+O^3	0.85	0.58	1	1.98	0.45	45.90	0.72	0.91	1.10
	OSIE+P^3+O^3	0.83	0.51	1.12	1.84	0.41	49.37	0.65	0.90	1.26

Table 2: Results of training MLNet and SALICON on different data.

and/or heuristics (e.g., AIM, IKN, GBVS, FES, CAS, CAS, SSR), and so are able to more efficiently detect target/distractor differences in the simple P^3 stimuli. This suggests that a combination of both theory- and data-driven approaches should be explored, similar to the ICF model that includes explicit local intensity and local contrast features as well as readout network trained on human gaze data.

Overall, these results demonstrate that the inability of saliency models to adequately respond to singletons extends beyond synthetic stimuli and is also apparent in more realistic scenarios. This calls for changes in algorithm design and also has ramifications for the use of these computational models in practical applications mentioned in Section 1.

4.3.3 Training Saliency Models on P^3 and O^3 Data

As mentioned in Section 1.1, fixation datasets contain a negligibly small portion of synthetic and real odd-one-out stimuli. Only CAT2000 has ≈4% synthetic patterns but is rarely used for training. Specifically, only the SAM model is trained on CAT2000. Here, we conduct a simple test to determine whether the struggle of deep learning models to detect singletons in synthetic and real images can be corrected by augmenting the training data. We select MLNet and SALICON for further experimentation[2].

We follow the original training procedures and only modify the training data. For MLNet we add 1553 P^3 and/or 1200 O^3 images to its training set of 10K images from the SALICON dataset. We augment the smaller OSIE dataset (700 images) used for training the SALICON model by adding 100 and 76 images from P^3 and O^3 respectively, thus matching the proportion of extra P^3 and O^3 data of 16% and 12%, respectively, for both models. We also train both models only on P^3 and O^3 to test whether the task can be learned from these examples. We use target binary masks as a substitute for human fixation maps. We evaluate models on 518 P^3 images and 401 O^3 images not used for training and on the MIT1003 dataset.

Table 2 summarizes the results of training MLNet and SALICON on various combinations of the original training data and data from P^3 and O^3. According to the evaluation against human fixations for the MIT1003 dataset (which is a reasonable approximation of performance on the MIT300 benchmark), augmenting training data with P^3 and/or O^3 patterns does not have a significant effect on fixation prediction accuracy. Both models generally do not learn well from the P^3 data alone and results are the worst overall. The O^3 training

[2]MLNet, SAM and SALICON are the only models with publicly available training code, however SAM also requires fixation points for training which are not available for P^3 and O^3. We use our own implementation of SALICON (https://github.com/ykotseruba/SALICONtf) since open-source code [] did not reproduce published results (see the supplementary material).

data results in the best performance on the O^3 test data, which is expected, but also leads to relatively high performance on the MIT1003 data as well as P^3 patterns. One possible explanation for the poor performance from P^3 compared to O^3 is that synthetic stimuli considerably differ in appearance from natural images making it difficult for CNN-based models to effectively leverage object representaions learnt in object classification (*e.g.* VGG[47]).

The best results are achieved when O^3 is mixed with the original training set for both models. Note that training on the O^3 data leads to the highest MSR_{targ} and $MSR_{targ} > MSR_{bg}$ for both MLNet and SALICON models, meaning that they somewhat improved their ability to distinguish targets and are less distracted by the objects in the background.

Our experiment shows that only minor improvements can be achieved by augmenting the training datasets with odd-one-out data and suggests that other factors might be at play. This is in agreement with literature exploring the behavior of deep-learning models in a related category of tasks requiring comparisons between items (*e.g.* the *same-different* task to determine whether the presented objects are the same or not). These works explore a range of architectures and determine that even training on datasets several orders of magnitude larger than ours does not bring these models close to human-level [25, 36, 50]. In particular, in [36] the authors hypothesize that feedforward architectures lack mechanisms for binding features to individuated objects which makes them struggle with these types of tasks. However, further research is needed to determine whether these conclusions can be extended to the odd-one-out detection task.

5 Conclusion

Despite saturated performance on common benchmarks, most classical and deep models still do not adequately model fundamental attention guiding mechanisms. Using two newly proposed datasets, P^3 and O^3, we systematically examined the performance of the saliency algorithms on synthetic stimuli similar to the ones used in human research and on a large collection of realistic odd-one-out images. We showed that several classical models discriminate targets in some of the feature dimensions considered and also perform better on average than the CNN-based models. However, a large gap remains between human and algorithm performance both quantitatively and qualitatively. We also showed that augmenting training data for two CNN-based saliency models or training them exclusively on P^3 and O^3 images does not significantly improve their ability to discriminate singletons and has only a minor positive effect on fixation prediction accuracy, which requires further investigation.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the NSERC Canadian Robotics Network (NCRN), the Air Force Office for Scientific Research (USA), and the Canada Research Chairs Program through grants to John K. Tsotsos.

References

[1] *3M White Van VAS Sample Report*. 3M Visual Attention Service, 2015. Version 5.2.

[2] SP Arun. Turning visual search time on its head. *Vision Research*, 74:86–92, 2012.

[3] Stefanie I Becker. The mechanism of priming: Episodic retrieval or priming of pop-out? *Acta Psychologica*, 127(2):324–339, 2008.

[4] Stefanie I Becker. The role of target–distractor relationships in guiding attention and the eyes in visual search. *Journal of Experimental Psychology: General*, 139(2):247, 2010.

[5] Stefanie I Becker. Simply shapely: Relative, not absolute shapes are primed in pop-out search. *Attention, Perception, & Psychophysics*, 75(5):845–861, 2013.

[6] Stefanie I Becker and Ulrich Ansorge. Higher set sizes in pop-out search displays do not eliminate priming or enhance target selection. *Vision Research*, 81:18–28, 2013.

[7] David Berga and Xavier Otazu. A Neurodynamical Model of Saliency Prediction in V1. *arXiv preprint arXiv:1811.06308*, 2018.

[8] David Berga, Xosé R Fdez-Vidal, Xavier Otazu, Víctor Leborán, and Xosé M Pardo. Psychophysical evaluation of individual low-level feature influences on visual attention. *Vision Research*, 154:60–79, 2019.

[9] Ali Borji and Laurent Itti. CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581.

[10] Ali Borji, Dicky N Sihite, and Laurent Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, 2013.

[11] Donald E. Broadbent. *Perception and Communication*. Pergamon Press, 1958.

[12] Neil Bruce and John Tsotsos. Saliency based on information maximization. In *NIPS*, pages 155–162, 2006.

[13] Neil Bruce and John Tsotsos. Attention based on information maximization. *Journal of Vision*, 7(9):950–950, 2007.

[14] Neil Bruce and John K Tsotsos. An information theoretic model of saliency and visual search. In *International Workshop on Attention in Cognitive Systems*, pages 171–183, 2007.

[15] Neil Bruce, Calden Wloka, Nick Frosst, Shafin Rahman, and John K Tsotsos. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research*, 116:95–112, 2015.

[16] Neil Bruce, Christopher Catton, and Sasa Janjic. A deeper look at saliency: Feature contrast, semantics, and beyond. In *CVPR*, pages 516–524, 2016.

[17] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. MIT Saliency Benchmark.

[18] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. In *ICPR*, pages 3488–3493, 2016.

[19] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.

[20] E. Erdem and A. Erdem. Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of Vision*, 13(4):11–20, 2013.

[21] Shu Fang, Jia Li, Yonghong Tian, Tiejun Huang, and Xiaowu Chen. Learning discriminative subspaces on random contrasts for image saliency analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 28(5):1095–1108, 2017.

[22] Dashan Gao, Vijay Mahadevan, and Nuno Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8(7): 1–18, 2008.

[23] Antón Garcia-Diaz, Xosé R Fdez-Vidal, Xosé M Pardo, and Raquel Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51–64, 2012.

[24] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *PAMI*, 34(10):1915–1926, 2012.

[25] ÇaÇğlar Gülçehre and Yoshua Bengio. Knowledge matters: Importance of prior information for optimization. *The Journal of Machine Learning Research*, 17(1):226–257, 2016.

[26] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2007.

[27] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *CVPR*, pages 1–8, 2007.

[28] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: Searching for coding length increments. In *NIPS*, pages 681–688, 2009.

[29] Xiaodi Hou, Jonathan Harel, and Christof Koch. Image signature: Highlighting sparse salient regions. *PAMI*, 34(1):194–201, 2012.

[30] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *CVPR*, pages 262–270, 2015.

[31] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, 2000.

[32] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, 1998.

[33] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, pages 1072–1080, 2015.

[34] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009.

[35] Nico A Kaptein, Jan Theeuwes, and AHC Van der Heijden. Search for a conjunctively defined target can be selectively limited to a color-defined subset of elements. *Journal of Experimental Psychology: Human Perception and Performance*, 21(5):1053–1069, 1995.

[36] Junkyung Kim, Matthew Ricci, and Thomas Serre. Not-So-CLEVR: learning same–different relations strains feedforward neural networks. *Interface focus*, 8(4):1–13, 2018.

[37] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[38] 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.

[39] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. DeepGaze II: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016.

[40] Matthias Kümmerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low- and high-level contributions to fixation prediction. In *ICCV*, pages 4789–4798, 2017.

[41] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. SalGAN: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[42] Rik Pieters, Michel Wedel, and Jie Zhang. Optimal feature advertising design under competitive clutter. *Management Science*, 53(11):1815–1828, November 2007.

[43] Nicolas Riche, Matei Mancas, Matthieu Duvinage, Makiese Mibulumukini, Bernard Gosselin, and Thierry Dutoit. RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, 28(6):642–658, 2013.

[44] Ruth Rosenholtz. A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, 39(19):3157 – 3163, 1999.

[45] Ruth Rosenholtz, Amal Dorai, and Rosalind Freeman. Do predictions of visual perception aid design? *ACM Transactions on Applied Perception*, 8(2):1–20, 2011.

[46] Hae Jong Seo and Peyman Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):1–27, 2009.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[48] Alireza Soltani and Christof Koch. Visual saliency computations: mechanisms, constraints, and the effect of feedback. *Journal of Neuroscience*, 30(38):12831–12843, 2010.

[49] Michael W Spratling. Predictive coding as a model of the V1 saliency map hypothesis. *Neural Networks*, 26:7–28, 2012.

[50] Sebastian Stabinger, Antonio Rodríguez-Sánchez, and Justus Piater. 25 years of cnns: Can we compare to human abstraction capabilities? In *International Conference on Artificial Neural Networks*, pages 380–387, 2016.

[51] Hamed Rezazadegan Tavakoli, Esa Rahtu, and Janne Heikkilä. Fast and efficient saliency detection using sparse sampling and kernel density estimation. In *Scandinavian Conference on Image Analysis*, pages 666–675. Springer, 2011.

[52] Christopher Thomas. OpenSALICON: An open source implementation of the salicon saliency model. *arXiv preprint arXiv:1606.00110*, 2016.

[53] Thomas Töllner, Michael Zehetleitner, Klaus Gramann, and Hermann J. Müller. Stimulus saliency modulates pre-attentive processing speed in human visual cortex. *PLoS ONE*, 6(1):1–8, 2011.

[54] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[55] Yuan-Chi Tseng, Joshua I Glaser, Eamon Caddigan, and Alejandro Lleras. Modeling the effect of selection history on pop-out visual search. *PLoS One*, 9(3):1–14, 2014.

[56] John K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13(3):423–445, 1990.

[57] Ralf van der Lans, Rik Pieters, and Michel Wedel. Research Note – Competitive Brand Salience. *Marketing Science*, 27(5):922–931, 2008.

[58] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *CVPR*, pages 2798–2805, 2014.

[59] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2018.

[60] Erik Wästlund, Poja Shams, and Tobias Otterbring. Unsold is unseen ... or is it? Examining the role of peripheral vision in the consumer choice process using eye-tracking methodology. *Appetite*, 120:49 – 56, 2018.

[61] Calden Wloka, Sang-Ah Yoo, Rakesh Sengupta, Toni Kunic, and John Tsotsos. Psychophysical evaluation of saliency algorithms. *Journal of Vision*, 16(12):1291–1291, 2016.

[62] Calden Wloka, Sang-Ah Yoo, Rakesh Sengupta, and John Tsotsos. The interaction of target-distractor similarity and visual search efficiency for basic features. *Journal of Vision*, 17(10):1130–1130, 2017.

[63] Calden Wloka, Iuliia Kotseruba, and John K. Tsotsos. Active fixation control to predict saccade sequences. In *CVPR*, pages 3184–3193, 2018.

[64] Calden Wloka, Toni Kunić, Iuliia Kotseruba, Ramin Fahimi, Nicholas Frosst, Neil DB Bruce, and John K Tsotsos. SMILER: Saliency Model Implementation Library for Experimental Research. *arXiv preprint arXiv:1812.08848*, 2018.

[65] Jeremy M Wolfe. Guided Search 2.0. A revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, 1994.

[66] Jeremy M. Wolfe. Visual search. In Harold Pashler, editor, *Attention*. Psychology Press, 1998.

[67] Jeremy M Wolfe. What can 1 million trials tell us about visual search? *Psychological Science*, 9(1):33–39, 1998.

[68] Jeremy M Wolfe. Visual search. *Current Biology*, 20(8):R346–R349, 2010.

[69] Jeremy M Wolfe and Todd S Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6):495–501, 2004.

[70] Jeremy M Wolfe and Todd S Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):0058, 2017.

[71] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173. IEEE, 2013.

[72] Michael Zehetleitner, Anja Isabel Koch, Harriet Goschy, and Hermann Joseph Müller. Salience-Based Selection: Attentional Capture by Distractors Less Salient Than the Target. *PLoS ONE*, 8(1):1–14, 2013.

[73] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *ICCV*, pages 153–160, 2013.

[74] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):1–20, 2008.