Objects in Focus: Predicting Object-Based Attention from Spatial Features

Elizabeth Hall
Meta
Redmond, Washington
elizabethhall@meta.com

Zoe Loh
University of California, Merced
Merced, California
zloh@ucmerced.edu

Abstract

We tackle the problem of predicting object-based attention in cluttered real-world scenes. The availability of open-source eye-movement data combined with semantic segmentations has advanced this research, yet existing models struggle to predict fixation distributions across space and depth. We created a new approach to predict how attention is distributed across objects, grounded in an cognitively inspired object-based attention paradigm. To predict the total number of fixations on each object, we factor in its position in the scene—considering eccentricity from the center, depth, pixel size, and salience. We tested our model using both Objects in Focus (OiF) dataset objects and the widely-used COCO-Freeview fixation set, along with the MS-COCO segmentations. When evaluated on the OiF dataset, the object-based attention model explained approximately 82% of the variance in cumulative object fixations, and accounted for 66% of the variance when tested on the COCO-Freeview dataset. Our results highlight the critical role of object-based parameters in predicting attention distribution across complex scenes.

1. Introduction

What constitutes an ”object”? In the context of human vision, an object is generally defined as a discrete, bounded entity that can be visually distinguished from its background or surrounding elements based on its features [14, 42, 45]. Objects are perceived as coherent units within the visual field and are processed as such by the brain [40]. These object-based representations effectively segment the scene along its physical boundaries. The transformation of millions of retinal signals into a limited number of behaviorally relevant objects requires prior knowledge of the physical world and the selection of information that is relevant to current behavioral goals [43].

The processes most related to human-like ’object attention’ fall under other terms in computer vision. Channel attention describes the “what” of what is being attended [23], while spatial attention refers to “where” the model should attend to in order to find the most relevant information [20]. Attention mechanisms have been integrated into systems trained to perform a variety of tasks including image classification [23], object detection [12], semantic segmentation [50], face recognition [47] and image generation [19].

A frequent bottom-up approach to modeling human attention in visual scenes is using low-level features to identify where stimuli are most salient [25]. These salience maps are calculated per pixel to highlight regions of interest in a scene [25]. In contrast, ”object-based attention” is the selective allocation of cognitive resources to specific objects within a scene (defined by an object boundary) [37]. In real-world settings, attention is typically focused on objects relevant to current goals or interests, even when they occupy only a small part of the visual field [14, 29].

We developed a paradigm to predict the distribution of human attention across objects. We calculate the sum of total fixations landing on an object, based on where the object lies in the scene (defined as the object’s eccentricity from center, its depth in the scene, its summed pixel size, and salience). We test this model on both the fixations and the Objects in Focus dataset objects [11, 21, 35, 39], as well the publicly available COCO-Freeview fixation set [10, 49] and accompanying MS-COCO segmentations [8]. Our contributions can be summarized as follows:

1. We introduce an object-based attention task, which focuses on understanding how human attention is distributed across objects in well-cluttered scenes.
2. We introduce an innovative ”object-based” training model to tackle this challenge, leveraging spatial and salience features of objects to predict the log sum of fixations on each object. This approach enables the model to seamlessly integrate object data across different scene images and perspectives. We present the results of this analysis using both the OiF and COCO-Freeview datasets, emphasizing the success of our model within each dataset, as well as its limitations in generalizing across different datasets.

Figure 1. Examples of indoor and outdoor scenes with their corresponding object masks for the Objects in Focus dataset (top) and the COCO-Freeview dataset (bottom). Background objects were excluded.

2. Related work

Most publicly shared eye-movement datasets have been collected under the Saliency Benchmark challenge [7, 28, 31, 32]. The Saliency Benchmark evaluates how different models can best predict patterns of fixation density on the MIT300 [28], CAT2000 [3], and COCO-Freeview datasets [10, 49]. Other datasets have been included under the benchmark, including OSIE [46], Toronto [4], FIGRIM [6], and EMOd [15]. Interest in the benchmark has waned over the years, although there is still a substantial gap between what patterns of fixation can be explained by the best performing models and the lower bound of explainable information [32, 34]. The current highest performing models are built on learned features from a deep convolutional model trained on object recognition, with most using ImageNet as a pretest task [13, 26, 34]. DeepGazeIIE [34], the reigning champion on all three datasets in the challenge, uses an ensemble mixture of ShapeNet-C [16], EfficientNet-B5 [41], ResNext-50 [22] and DenseNet-201 [24], along with the relevant fixation training set.

3. Stimuli

Objects in Focus (OiF) The dataset consists of 100 real-world scenes, each with all objects fully segmented (Figure 1). Of the 100 scenes, 50 depict outdoor environments, while the remaining 50 represent indoor environments. All images were segmented at their full resolution (1024 $\times$ 768 pixels). Objects of the same category were uniformly labeled. Segmentations represented 360 semantically-unique object labels found in the ConceptNet architecture. One hundred participants viewed each scene for 12 seconds. The eye data was analyzed with courtesy to the UC Davis Visual Cognition Lab [11, 21, 35, 39].

COCO-Freeview. Participants in COCO-Freeview [10, 49] viewed the same scenes as those collected for the prediction of target-based attention in COCO-Search [9, 48]. This dataset includes 6202 images sourced from Microsoft COCO, Common Objects in Context (MS-COCO) [33]. The dataset consists of images with varying resolutions, and each image was scaled to a height of 1050 pixels while maintaining its aspect ratio. From this dataset, we selected a subset of 2243 images.

The COCO-Stuff annotation set includes semantic segmentations for things (objects with a well-defined shape, e.g. car, person) and stuff (amorphous background regions, e.g. grass, sky) [8]. As the images were originally selected for a target-search task, a majority of the images contain one of the 18 MS-COCO object categories [33]. Although there are more objects overall in the COCO-Freeview annotation set than in OiF, they represent fewer semantically unique object labels (168 labels vs. OiF's 360 labels). To calculate object-level attention, we separated masks for objects sharing the same ID. We filtered objects occupying 0.05 percent of image area, downscaled arrays to 1/30th original size, applied connected component analysis to identify separate objects per ID, then upscaled and calculated values for each unique mask label.

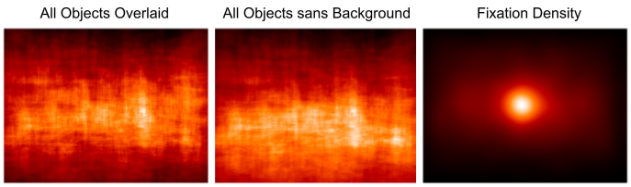

Figure 2. Density of all objects, all non-background objects, and fixations across scenes from the OiF dataset.

4. Object-based attention (OBA) model

We hypothesized that probable distributions for eye movements across objects can be predicted from a combination of objects' size and spatial locations. Object size was calculated as the sum of the pixels in the object mask. For eccentricity, we first found the center of mass for each binary object mask (1 for object pixels, 0 for background). Eccentricity was then defined as the Euclidean distance from the center of mass of an object to the center of an image array. Per-pixel depth maps for each image were created using the PyTorch implementation of MONODEPTH2's [17] self-supervised monocular depth estimation model. Depth for each object was calculated as the depth value at the centroid point of the object mask. We also include the object's relative salience in the image. Salience maps were constructed using DeepGaze IIE [34] with a uniform center bias array map. The salience feature value was indexed in the same manner as the depth value, by extracting the value at the center of mass for each object mask.

The distributions for the cumulative sum of fixations on each object, the object's size and the object's depth were all positively-skewed. These parameters were logarithmically transformed before fitting the full model to account for the heavy skew (Figure 2). We then fit the full multiple linear regression model, where the outcome $\log(S_x)$ is the log of the sum of cumulative fixations on object mask x, predicted by the log transformation of object depth, the log transformation of object size, the z-scored object eccentricity, and the z-scored object salience.

5. Results

We compare the results of our object-based attention model for the OiF and COCO-FV datasets (Table 1). We test the model on each dataset individually, and a set with all objects from OiF and COCO-FV combined. We find that a model trained on the object values collected from the OiF dataset performs best at predicting the sum of fixations on an object, with an R^2 of 0.82. Each object in the OiF dataset received 49.8 fixations on average (SD = 96.4, range = 0 - 1180), and the OiF-trained model had a mean absolute error of ± 35.4 fixations. Figure 3 shows examples for object-based model

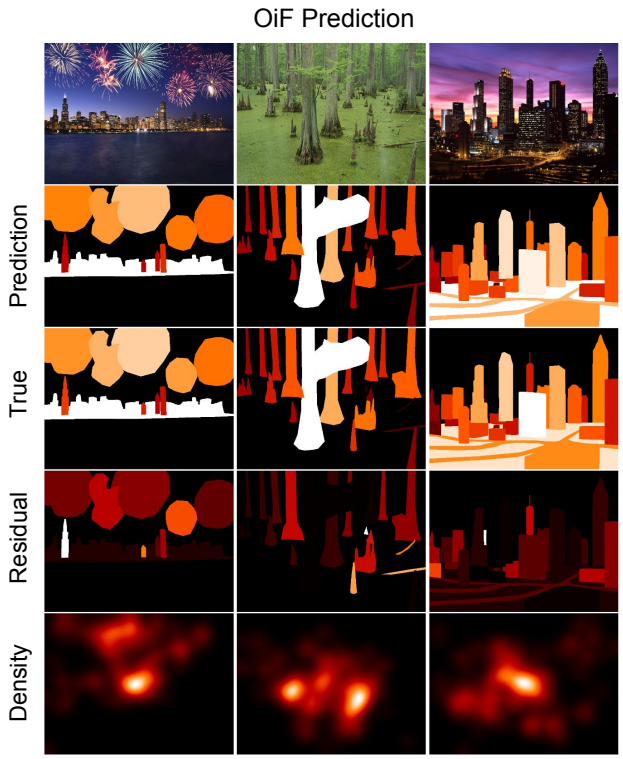

Figure 3. Visualization of the images with the best performance in predicting object attention for the OiF dataset. For the heatmaps, the first row depicts the model prediction for how attention will be distributed across objects. The second row depicts the true distribution, while the third row highlights the absolute residual for each object. The final row is the gaussian fixation density heatmap. Each heatmap was independently normalized, with black representing the minimum value and white representing the maximum value within that heatmap.

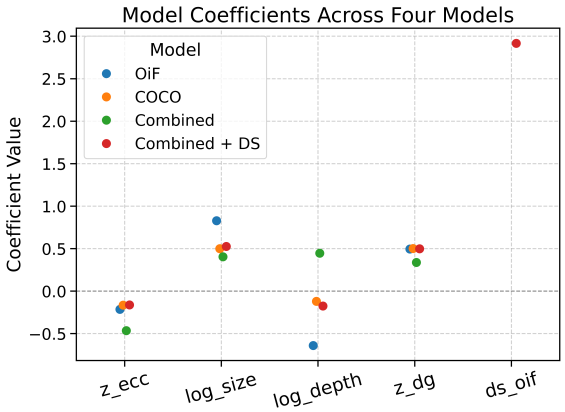

Figure 4. Coefficient values for eccentricity, size, depth, saliency across the four models.

Dataset	# Obj.	R^2	F	Log MAE Obj.	Log MAE Scene	MAE Obj.	MAE Scene
Oif	2493	0.82	2920.5	0.50	0.49	35.4	46.9
COCO	31138	0.66	1530.7	0.59	0.61	4.8	7.1
Combined	33631	0.49	7945.6	0.76	0.72	11.5	12
Combined + β DS	33631	0.73	18,330	0.59	0.61	8.6	10.1

Table 1. Model fit and prediction accuracy (mean absolute error) statistics for the OiF dataset, the COCO dataset, the two datasets combined, and the two datasets combined with the addition of a categorical dataset coefficient in the model.

trained on the OiF dataset. It shows for 3 example scenes, the true spread of attention across objects, the predicted attention and residual error per object, as well as a fixation density map. The fixation density map for the image was generated by convolving a Gaussian filter across fixation locations for all observers.

The object-based model performed less well on the COCO-FV dataset, with an R^2 of 0.66. Each object in this set received on average 7.6 fixations (SD = 18.1, range = 0 - 182), and the MAE for the COCO-FV - trained model was ± 4.8 fixations per object. Further, the object-based model was unsuccessful at generalizing across datasets. When trained on the combined OIF + COCO-FV objects, the R^2 dropped to 0.49, with a MAE of ± 11.5 fixations. The large degree of difference between the datasets in the number of viewers per image, the scene perspective, and the retinal object size can most likely explain why the model had difficulty in generalizing across sets. The inclusion of a categorical 'dataset' parameter to the model allowed it to account for the source of each objects' values, improving performance on the combined dataset to an R^2 of 0.73 (vs. 0.49) and a MAE per object of ± 8.6 fixations (vs. 11.5).

Aside from the categorical 'dataset' parameter, we find that the log size of the object accounts for the most variance in predicting the cumulative fixations on an object. It is particularly informative in the OiF-trained model, likely due to the fact that OiF images are largely scene-centric, have a greater average scene depth, and are well-cluttered with many objects. The coefficient values for log depth in the OiF-trained, COCO-trained, and Combined+DS models are all negative, likely representing the human bias to focus on objects presented closer in depth. However, the Combined model alone has a positive weight for the log depth coefficient, which may express the Combined model's difficulty in generalizing perspectives across datasets.

6. Discussion

We developed an object-based attention model and compared its performance in predicting fixations across the Objects in Focus dataset and the COCO-Freeview dataset. Object spatial probability plays a critical role in predicting patterns of fixations. DeepGazeIIE uses only 3-4 layers from the ultimate and penultimate layer spaces to train its model, and later layers of deep convolutional networks trained on ImageNet have been shown to correspond with feature patterns comprising entire objects [51]. These penultimate layers preferentially respond to the location and boundary of object shapes, in a way similar to the preferential response of object processing areas in the human brain [2, 30, 44]. Of note, there is substantial variability in how well different backbone models generalize to unseen fixation data [34], which may underscore how features relevant for network-based object recognition based on the ImageNet dataset [13] may fail to fully capture the relationship between object spatial probability and human perception. Model performance could be improved by including learned features from networks pretrained on other tasks such as object detection, by pre-training on larger sets of real and simulated fixations [27], or by incorporating more constraints relevant to human perception and attention.

Of note, there is a distinctive difference between the probable locations for objects, faces, and hands in images across large image datasets [13] and the location of those objects captured in stills from ego-centric video [1, 18, 36, 38]. Additionally, fixations collected from head-mounted eyetrackers in ambulatory conditions are less dispersed than those collected in desktop conditions, with shorter saccade latencies and a stronger bias to fixate towards the center of the frame of view [5].

Recent advancements in generative image models have integrated saliency-based image features to reliably guide viewers' attention to predefined regions within an image [52]. Future work could leverage object-level segmentations to direct attention toward specific objects, rather than regions, or to generate complex scenes populated with semantically related objects.

In conclusion, we found that the model performed worse when tested on the datasets combined compared to the individual datasets until dataset was included as a parameter. Our results reveal the importance of accounting for image viewpoint and its impact on object spatial parameters in object-based prediction models of gaze behavior. We believe this work represents a valuable resource for advancing research in both human and computer vision.

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 4

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017. 4

[3] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015. 2

[4] Neil Bruce and John Tsotsos. Attention based on information maximization. *Journal of Vision*, 7(9):950–950, 2007. 2

[5] Charlie S Burlingham, Naveen Sendhilnathan, Oleg Komogortsev, T Scott Murdison, and Michael J Proulx. Motor "laziness" constrains fixation selection in real-world tasks. *Proceedings of the National Academy of Sciences*, 121(12): e2302239121, 2024. 4

[6] Zoya Bylinskii, Phillip Isola, Constance Bainbridge, Antonio Torralba, and Aude Oliva. Intrinsic and extrinsic effects on image memorability. *Vision research*, 116:165–178, 2015. 2

[7] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*, 2016. 2

[8] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018. 1, 2

[9] Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1):1–11, 2021. 2

[10] Yupei Chen, Zhibo Yang, Souradeep Chakraborty, Sounak Mondal, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Characterizing target-absent human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 5031–5040, 2022. 1, 2

[11] Deborah A. Cronin, Elizabeth H. Hall, Jessica E. Goold, Taylor R. Hayes, and John M. Henderson. Eye movements in real-world scene photographs: General characteristics and effects of viewing task. *Frontiers in Psychology*, 10, 2020. 1, 2

[12] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 1

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 4

[14] J. Duncan. Selective attention and the organization of visual information. *Journal of experimental psychology. General*, 113(4):501–517, 1984. 1

[15] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L Koenig, Juan Xu, Mohan S Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 7521–7531, 2018. 2

[16] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2

[17] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation, 2019. 3

[18] Michelle R Greene, Benjamin J Balas, Mark D Lescroart, Paul R MacNeilage, Jennifer A Hart, Kamran Binaee, Peter A Hausamann, Ronald Mezile, Bharath Shankar, Christian B Sinnott, et al. The visual experience dataset: Over 200 recorded hours of integrated eye movement, odometry, and egocentric video. *arXiv preprint arXiv:2404.18934*, 2024. 4

[19] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International conference on machine learning*, pages 1462–1471. PMLR, 2015. 1

[20] Jingxia Guo, Nan Jia, and Jinniu Bai. Transformer based on channel-spatial attention for accurate classification of scenes in remote sensing image. *Scientific Reports*, 12(1):15473, 2022. 1

[21] Taylor R. Hayes and John M. Henderson. Looking for semantic similarity: What a vector-space model of semantics can tell us about attention in real-world scenes. *Psychological Science*, 32(8):1262–1270, 2021. PMID: 34252325. 1, 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[23] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 1

[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[25] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001. 1

[26] Sen Jia and Neil D. B. Bruce. Eml-net:an expandable multi-layer network for saliency prediction, 2019. 2

[27] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015. 4

[28] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012. 2

[29] Nancy Kanwisher and Jon Driver. Objects, attributes, and visual attention: Which, what, and where. *Current directions in psychological science*, 1(1):26–31, 1992. 1

[30] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10 (11):e1003915, 2014. 4

[31] Matthias Kümmerer, Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, Antonio Torralba, and Matthias Bethge. Mit/tübingen saliency benchmark. https://saliency.tuebingen.ai/, . 2

[32] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In *Computer Vision – ECCV 2018*, pages 798–814. Springer International Publishing, . 2

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[34] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12919–12928, 2021. 2, 3, 4

[35] Zoe Loh, Elizabeth H Hall, Deborah Cronin, and John M Henderson. Working memory control predicts fixation duration in scene-viewing. *Psychological research*, 87(4):1143–1154, 2023. 1, 2

[36] Zhaoyang Lv, Nicholas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, Kiran Somasundaram, Luis Pesqueira, Mark Schwesinger, Omkar Parkhi, Qiao Gu, Renzo De Nardi, Shangyi Cheng, Steve Saarinen, Vijay Baiyya, Yuyang Zou, Richard Newcombe, Jakob Julian Engel, Xiaqing Pan, and Carl Ren. Aria everyday activities dataset, 2024. 4

[37] George L. Malcolm and Sarah Shomstein. Object-based attention in real-world scenes. *Journal of experimental psychology. General*, 144(2):257–263, 2015. 1

[38] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng (Carl) Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20133–20143, 2023. 4

[39] Candace E. Peacock, Elizabeth H. Hall, and John M. Henderson. Objects are selected for attention based upon meaning during passive scene viewing. *Psychonomic Bulletin amp; Review*, 30(5):1874–1886, 2023. 1, 2

[40] Marius V. Peelen and Sabine Kastner. Attention in the real world: toward understanding its neural basis. *Trends in cognitive sciences*, 18(5):242–250, 2014. 1

[41] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2

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

[43] Thomas Tsao and Doris Y. Tsao. A topological solution to object segmentation and tracking. *Proceedings of the National Academy of Sciences of the United States of America*, 119(41):e2204248119, 2022. 1

[44] Haiguang Wen, Junxing Shi, Wei Chen, and Zhongming Liu. Deep residual network predicts cortical representation and organization of visual features for rapid categorization. *Scientific reports*, 8(1):3752, 2018. 4

[45] Jeremy M Wolfe, Kyle R Cave, and Susan L Franzel. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance*, 15(3):419, 1989. 1

[46] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014. 2

[47] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4362–4371, 2017. 1

[48] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[49] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Predicting human attention using computational attention. *arXiv preprint arXiv:2303.09383*, 2023. 1, 2

[50] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 701–710, 2018. 1

[51] MD Zeiler. Visualizing and understanding convolutional networks. In *European conference on computer vision/arXiv*, 2014. 4

[52] Yunxiang Zhang, Nan Wu, Connor Z Lin, Gordon Wetzstein, and Qi Sun. Gazefusion: Saliency-guided image generation. *ACM Transactions on Applied Perception*, 2024. 4