A Taxonomy of Computational Models of Covert Attention during Visual Search

Ansh Soni[1], Sudhanshu Srivastava[2], Craig K. Abbey[1], and Miguel P. Eckstein[1,3,4,5]

[1]Department of Psychological and Brain Sciences, University of California, Santa Barbara

[2]Graduate Program in Dynamical Neuroscience, University of California, Santa Barbara

[3]Department of Electrical and Computer Engineering, University of California, Santa Barbara

[4]Department of Computer Science, University of California, Santa Barbara

[5]Institute for Collaborative Biotechnologies, University of California, Santa Barbara

Abstract

The performance degradation with increased distractors (the set size effect) during visual search has often been used to investigate the properties of covert attention. This approach has led to the development of various theories and models to understand covert attention in these contexts. Recently, attention has shifted towards exploring new machine learning models (Srivastava et al., 2021; Nicholson and Prinz, 2022), which have mainly been studied in isolation from classical models. Our work aims to unify these models to allow a direct comparison between old and new. We present Python implementations of seven models based on perceptual accuracy in covert attention during visual searches. These models are tested on two feature searches (angle or luminance) followed by a conjunction task combining both features. Among these, four models process image data directly, including a Bayesian ideal observer, a suboptimal Bayesian model, a compact 5-layer convolutional neural network (CNN), and a VGG-16 network adapted to the task through a transfer learning process. Additionally, we developed three models that process extracted features, assuming a normal distribution of activation from the target and each distractor. These models are a traditional signal detection theory model, a signal detection model with capacity limits, and a performance-based version of guided search (Wolfe, 2021). Our findings reveal that, like most models, both the CNN and VGG models demonstrate a convex set-size effect, with guided search being the sole model predicting non-convex set-size effects. When aligning feature performance at the smallest set size across all models, the set size effect in the CNN is less pronounced than in the VGG model, and both have a larger effect than the suboptimal Bayesian model while having a smaller effect than the traditional signal detection theory model. Integrating cutting-edge machine learning models with traditional ones in a common framework allows for broader comparisons, enhancing our understanding of covert attention mechanisms.

1 Introduction

Covert visual attention refers to the process of selecting regions of the visual field without moving the eyes. Human visual search for a target among distractors has traditionally been used as an experimental tool to make inferences about the mechanisms mediating covert attention. When the target and distractors are visually similar observers are slower and less accurate at detecting the target as the number of distracting objects in the scene increases. This is referred to as the set-size effect. Various theories and models have aimed to explain the set-size effect in terms of mechanisms of covert attention: limited resources spotlight (Posner, 1980, Luck and Vogel, 1997), a temporal serial processor (Treisman and Gelade, 1980, etc), or simply increasing probability that the target will be confused with a distractor (Swensson and Judy, 1981; Palmer et al., 2000; Palmer, 1994; Eckstein, 1998; Eckstein, Thomas, et al., 2000).

We begin by studying the earlier non-image-computable models that do not act directly on the images. These models all build off of a traditional signal detection theory (SDT) model (Green, Swets, et al., 1966), which assumes that each element in the display elicits an internal response within the observer's brain. The are normally distributed and are evoked in a parallel fashion, with independent processing of each feature, which is combined later for conjunction searches. This model is implemented in 3 different ways. First, a purely SDT model, where the cause of the set size effect is the increasing number of samples from a distractor distribution as more distractors are present, and the task difficulty stays constant (Figure 3.1, Eckstein, 1998). For this model, conjunction search has a higher set-size effect as it is inherently a more difficult task. Another variation follows the same steps as the pure SDT model but introduces a change in task difficulty where the set-size effect is further increased due to an increase in task difficulty with larger set-sizes (on top of the increasing distractor samplings). This increase in difficulty is interpreted as a capacity limitation and is expected to exist only in conjunction search (Figure 3.2, Põder and Kosiło, 2019). For this model, combining the extra capacity limits for conjunction along with increased task difficulty leads to the higher set size effect for conjunction. Last is an implementation of guided search (traditionally a model for reaction times, Wolfe, 2021) in the performance regime (Figure 3.3, Eckstein, Beutter, et al., 2000). This model incorporates a serial stage of attention following the initial parallel stage. In this stage, all items that provided an activation above a threshold, or are suspicious, are processed serially, in the order of suspicion (higher activation first). If the target is processed or all suspicious items are processed during a limited search time, the participant is always correct. Otherwise, they have to make an educated guess (based on the number of suspicious lines, number of lines left to search, etc.). The set size effect for this model is determined by a combination of the time it takes to process each item, the limited search time, and the threshold of suspicion, on top of the task difficulty from the parallel processing stage. For this model, the increased task difficulty and the lower threshold for suspicion for conjunction search leads to its increased set size effect.

We also study image-computable models. This is important as earlier models worked only on hypothetical internal responses and therefore did not model any of the early processing stages. This leads to a more difficult interpretation of how various features are processed and combined. We first implement 2 models based on the Bayes theorem. The first is a traditional Bayesian Ideal Observer (BIO) (Figure 3.4, Geisler, 2011) that compares the current trial against every possible trial combination (templates). The set size effect occurs due to an increasing amount of uncertainty since there are more possible templates as the set size increases. Conjunction searches also have significantly more uncertainty as having multiple distractor types leads to many more distractor combinations, causing an increased set size effect. Another model is suboptimal and compares each location within the trial separately rather than comparing the global trial (3.5, Ma et al., 2011). A final decision is made for each trial after combining across locations. The set size effect occurs due to the model combining information over an increased number of locations as the set size increases. Furthermore, Conjunction search has a higher set size effect due to an increased amount of uncertainty at each location, comparing with 3 templates (2 distractors and the target) as opposed to 2.

Finally, we study a second type of image-computable model, convolutional neural networks (CNN). Unlike the previous image-computable models, a CNN learns from the images, optimizing for performance. While harder to interpret due to the large number of hidden parameters, these models have served as some of the best general models of the human visual system (Yamins et al., 2014, Konkle and Alvarez, 2022). We implement two types of CNNs. The first is a small 5-layer CNN that is fully trained only on visual search (3.6, Srivastava et al., 2021). The second utilizes a pre-trained CNN (Simonyan and Zisserman, 2014), trained on the ImageNet database (Deng et al., 2009). The convolution layers of this model are assumed to extract features similar to those humans use. Using this model, we complete a transfer learning process, training only the 2 dense layers and output layer, maintaining the features from the image-net-trained convolution layers (Nicholson and Prinz, 2022). Due to the black-box nature of these models, it is hard to pinpoint the reason for the set-size effect, although it seems to come from combining across an increasing number of locations when the set size increases. Conjunction search seems to have a higher set-size effect due to an increased number of possible training and testing samples.

We explore these various models of visual search on a simple search task with feature and conjunction variants. The task is a target present/absent task with varying set-size (N) where 50% of trials have a target and N-1 distractors, and the other 50% have N distractors. For the feature variant, all of these distractors are the same, varying from the target in only one feature dimension (angle or luminance). For the conjunction variant, both of these distractors (angle or luminance) could be present in the display, with all present distractors being sampled independently (the number of each distractor does not need to be balanced).

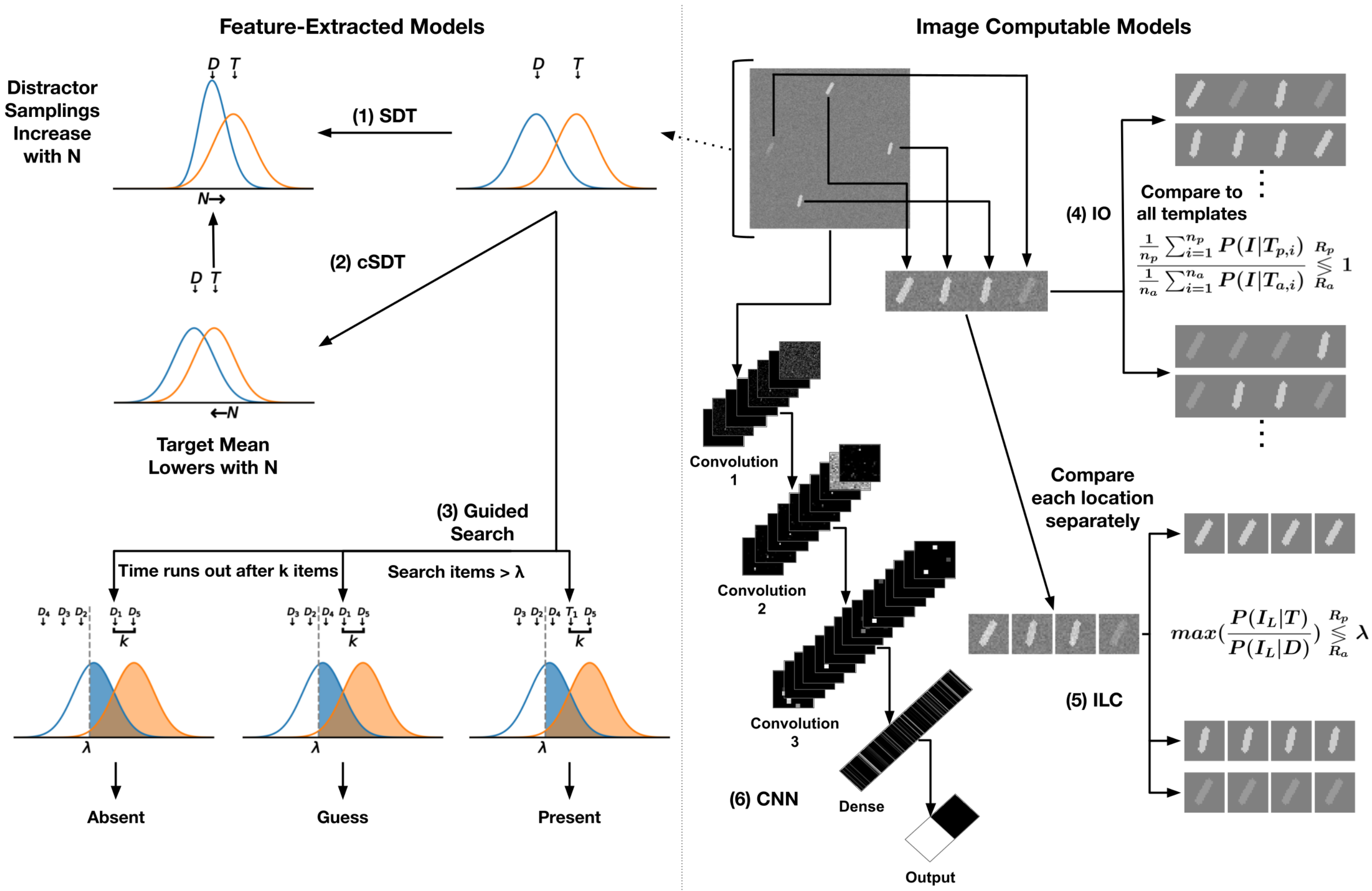

Figure 1

Visual depiction of all models. All models originate from either the raw image or extracted features and follow the visualized steps to make a decision.

2 Methods

Table 1

Overview of Models

Model	Image-Computable	Limited-Resources	Fitting Parameters	Citation
Ideal Observer (IO)	Yes	No	σ	Geisler, 2011
Individual-Location-Comparison (ILC)	Yes	No	σ	Ma et al., 2011
Signal-Detection-Theory (SDT)	No	No	d'	Green, Swets, et al., 1966, Eckstein, 1998
Signal-Detection-Theory with Capacity (cSDT)	No	Yes	d', b_c, b_f	Põder and Kosiło, 2019
Guided Search (GS)	No	Yes	d', k, λ_c, λ_f, g_c, g_f	Wolfe, 2021
Fully Trained Convolutional Neural Network (CNN)	Yes	No	σ	Srivastava et al., 2021
Pre-Trained Convolutional Neural Network (VGG)	Yes	No	σ	Simonyan and Zisserman, 2014, Nicholson and Prinz, 2022

2.1 Signal Detection Theory (SDT)

This model assumes different response distributions to the target and the distractor. These distributions come from the observer's internal noise and are assumed to be normal distributions with a standard deviation of 1, with the mean of the distractor distribution set to 0. The overlap of these distributions determines the task's difficulty and is characterized by the index of detectability (d') or the mean of the target response distribution. The distractor distribution is sampled once for each distractor. The maximum activation from all items is taken as the overall activation for the trial, which is then compared to a threshold (λ) with the model predicting target present if the trial activation is above the threshold and target-absent otherwise. For this model, we describe a hit-rate (HR) and false-positive rate (FP) dependent on a criterion (λ), or the threshold of determining target present or absent. The hit-rate and false-positive rate are then combined to find an overall performance (PC):

$$HR(\lambda, N, d') = 1 - \Phi(\lambda - d')\Phi(\lambda)^{N-1} \tag{1}$$

$$FP(\lambda, N) = 1 - \Phi(\lambda)^{N} \tag{2}$$

$$PC(\lambda, N, d') = \frac{1}{2} * (1 - FP(\lambda, N, d') + HR(\lambda, N, d')) \tag{3}$$

To find a final performance of this model for a specific set-size (N) and task difficulty (d') we use the criterion that maximizes the PC. We can find the optimal λ by finding where $\frac{\partial}{\partial\lambda}PC(\lambda, N, d')$ equals 0:

$$\frac{\partial}{\partial\lambda}PC(\lambda, N, d') = \frac{\partial}{\partial\lambda}(\frac{1}{2} * (\Phi(\lambda)^{N} + 1 - \Phi(\lambda - d') * \Phi(\lambda)^{N-1})) = 0 \tag{4}$$

$$=> N\varphi(\lambda)\Phi(\lambda)^{N-1} - \varphi(\lambda - d')\Phi(\lambda)^{N-1} - (N-1)\Phi(\lambda - d')\varphi(\lambda)\Phi(\lambda)^{N-2} = 0 \tag{5}$$

Divide by $\Phi(\lambda)^{N-1}\varphi(\lambda)$:

$$=> N - \frac{\varphi(\lambda - d')}{\varphi(\lambda)} - (N-1)\frac{\Phi(\lambda - d')}{\Phi(\lambda)} = 0 \tag{6}$$

Eq. 6 can then be solved numerically to find the optimal λ, which is then inputted into eq. 3 along with n and d' to find the performance.

Furthermore, this model is able to define a relationship between feature and conjunction by assuming the multiple features are extracted independently. When the index of detectability for the multiple features is matched, the combination of information across features reduces the index of detectability by a factor of $\sqrt{2}$ (Eckstein, Thomas, et al., 2000). Due to this, the model has only one shared free parameter, d'.

2.2 Limited Resources - SDT

This model is a variation of the SDT models that also incorporates capacity limits. This model predicts that d' will lower as a function of set size as attention is being spread over a larger number of items. One way to implement this is to define a capacity variable (b) that can control how much d' changes as a function of set size:

$$d'_N = \frac{d'_1}{N^{\frac{b}{2}}} \tag{7}$$

This d'_N replaces d' in eq. 3, which is then used to find the performance. While the relationship in d'_1 between feature and conjunction search is still defined the same as SDT, the capacity variable is not. This leads to two free parameters in the model, with only one shared value between the search types.

2.3 Limited Resources - Guided Search

This model incorporates multiple stages of processing. There is an initial parallel stage where every element in a display elicits a noisy response along each feature dimension. In the next stage, up to k elements (the number of items that can be processed during a limited presentation time) that had an activation above a threshold (λ) are searched through serially, in order of the activation they produced in the parallel stage. If there are k or fewer items above the threshold, the model responds with target present if one of these items is the target, and target absent otherwise. If there are more than k items above the threshold, the model responds with target present if the target is within the top k activations. Otherwise, you guess target present or absent based on the guess rate of target present (g).

Similar to SDT, we split the performance into a hit rate (HR) and a false-positive rate (FP). False positives are made when the model guesses target present, with guess rate g, on a target-absent display, which occurs when more than k distractors produce an activation above the threshold. This means that the model is unable to process all of the items with an activation above the threshold and is, therefore, forced to guess:

$$FP(\lambda, g, k, N) = g \sum_{i=k+1}^{N} \binom{N}{i} (1 - \Phi(\lambda))^i \Phi(\lambda)^{N-i} \tag{8}$$

Note that if N is less than k, this is an empty summation, and the false-positive rate is 0.

The hit rate can further be split into two terms, probability of responding target present (R_p) when the target activation (T) is above the threshold and another for when it is below:

$$HR(\lambda, g, k, N, d) = \Phi(\lambda - d)P(R_p|T < \lambda) + (1 - \Phi(\lambda - d))P(R_p|T > \lambda) \tag{9}$$

Starting with the first term in eq. 9, or when the target activation is below the threshold, the model can guess target present through a similar process as the false positive rate, with N-1 distractors instead of N since the target is the remaining item:

$$\Phi(\lambda - d)P(R_p|T < \lambda) = g \cdot \Phi(\lambda - d) \sum_{i=k+1}^{N-1} \binom{N-1}{i} (1 - \Phi(\lambda))^i \Phi(\lambda)^{N-1-i} \tag{10}$$

For the second term in eq. 9, or when the target activation is above the threshold, the model responds with target present when the target activation is within the k highest activations, meaning the model has processed the target. If the target activation is below the k highest activations, the model is forced to guess and could respond with target present based on the guess rate. To account for both of these conditions, we start by defining $P(R_p|T > \lambda)$:

$$(1 - \Phi(\lambda - d))P(R_p|T > \lambda) = (1 - \Phi(\lambda - d)) \int_{-\infty}^{\infty} P(R_p|T = x)P(T = x|T > \lambda)\, dx \quad (11)$$

We can then define the second term within the integral by making the probability of the target activation under λ equal 0:

$$P(T = x|T > \lambda) = \begin{cases} 0, & x \leq \lambda \\ \frac{\varphi(x-d)}{1-\Phi(\lambda-d)}, & x > \lambda \end{cases} \quad (12)$$

We can then write the other term, $P(R_p|T = x)$ in terms of the probability at least k distractors are over the target activation, or the probability the model is forced to guess (the model is correct when it does not guess):

$$P(R_p|T = x) = 1 + (g - 1) \sum_{i=k}^{N-1} \binom{N-1}{i} (1 - \Phi(x))^i \Phi(x)^{N-1-i} \quad (13)$$

Substituting eq. 12 and eq. 14 into eq. 11 and simplifying, we get:

$$P(R_p|T > \lambda) = 1 - \Phi(\lambda - d) + (g - 1) * \sum_{i=k}^{N-1} \binom{N-1}{i} \int_{\lambda}^{\infty} \varphi(x - d)(1 - \Phi(x))^i \Phi(x)^{N-1-i}\, dx \quad (14)$$

Finally, we can combine eq. 11 and eq. 10 to find the hit rate (eq. 9) before combining the hit rate and the false positive rate (eq. 8) to get the performance:

$$PC(\lambda, g, k, n, d) = \frac{1}{2}(HR(\lambda, g, k, N, d) + 1 - FP(\lambda, g, k, N)) \quad (15)$$

The relationship in d' between feature and conjunction search is still defined the same as SDT and k is also shared as the number of items you can process in a given time is consistent. However, g and λ are independent on each search, as they depend on the observer's uncertainty for each task. This leads to 4 free parameters in the model, with two shared values between the search types.

2.4 Bayesian Ideal Observer (BIO)

This Bayesian Ideal Observer uses all of the available information and provides the optimal performance at a certain task difficulty. This makes it useful as a benchmarking technique for other models/observers. The BIO comes from Bayes Rule, which describes the probability of something happening (A) based on a set of priors (B):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{16}$$

For our paradigm, we can define two events, the probability that a given image is a target present image (plus some noise) ($P(T_p|I)$) or a target absent image (plus some noise) ($P(T_a|I)$) for each given image. We can then define a likelihood ratio to determine the decision of target present(1) or absent(0):

$$\textbf{Decision} = \begin{cases} 0, & \frac{P(T_p|I)}{P(T_a|I)} < 1 \\ 1, & \frac{P(T_p|I)}{P(T_a|I)} > 1 \end{cases} \tag{17}$$

Using this ratio (eq. 17) along with Bayes Rule (eq. 16), we can create a ratio of definable probabilities that can be used to make the decision:

$$\frac{P(T_p|I)}{P(T_a|I)} = \frac{\frac{P(I|T_p)P(T_p)}{P(I)}}{\frac{P(I|T_a)P(T_a)}{P(I)}} = \frac{P(I|T_p)P(T_p)}{P(I|T_a)P(T_a)} \tag{18}$$

This is further reduced since $P(T_a) = P(T_p) = 0.5$ to get $\frac{P(I|T_p)}{P(I|T_a)}$. We now just need to define $P(I|T_p)$ and $P(I|T_a)$. Starting with the target present term, we first take the fact that there is a finite number of target present stimuli (templates), and therefore the image (I) must be a combination of one of these templates, say template k, and the noise added:

$$P(I|T_p) = \sum_{i=1}^{n_p} P(T_{p,k} + N(0,\sigma)|T_{p,i}) \tag{19}$$

Since the added noise is normally distributed, we can define $P(T_{p,k} + N(0,\sigma)|T_{p,i})$ as:

$$P(T_{p,k} + N(0,\sigma)|T_{p,i}) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{\mu^2}{2\sigma^2}} \tag{20}$$

Where μ, defined as the mean of the added noise, can be defined as:

$$\mu = ||T_{p,k} + N(0,\sigma) - T_{p,i}|| \tag{21}$$

We then combine eq.20 and eq.19 and repeat the process for $P(I|T_a)$. Combining these results, we get:

$$\frac{P(T_p|I)}{P(T_a|I)} = \frac{\frac{1}{n_p}\sum_{i=1}^{n_p} P(I|T_{p,i})}{\frac{1}{n_a}\sum_{i=1}^{n_a} P(I|T_{a,i})} = \frac{\frac{1}{n_p}\sum_{i=1}^{n_p} e^{-\frac{||I-T_{p,i}||^2}{2\sigma^2}}}{\frac{1}{n_a}\sum_{i=1}^{n_a} e^{-\frac{||I-T_{a,i}||^2}{2\sigma^2}}} \tag{22}$$

The number of templates is dependent on the set size (N), search type, and whether the target is present or absent:

Table 2

Number of Templates

Type	T_p	T_a
Feature	N	1
Conjunction	$N2^{N-1}$	2^N

To get a performance for this model, we take the model decisions from eq. 17 for 40000 trials For each trial, we randomly choose a template and add noise.

The relationship between feature and conjunction search is defined within this model as there is an increase in uncertainty in conjunction search due to the many more possible templates. Since this relationship is defined, there is only one fitting parameter, the task difficulty, which in this case is the standard deviation of the Gaussian white noise added to each trial.

2.5 Individual Location Comparison (ILC)

This model follows a similar process as the ideal observer, with the key difference of comparing each item to a template of the target and distractors instead of comparing the whole image to all of the possible templates. Start by comparing each location to the target and distractors ($n_d = 1$ for feature search and $n_d = 2$ for conjunction search):

$$\frac{P(I_L|T)}{P(I_L|D)} = \frac{e^{-\frac{||I_{L,T}+N(0,\sigma)-T||^2}{2\sigma^2}}}{\frac{1}{n_d}\sum_{i=1}^{n_d} e^{-\frac{||I_{L,Di}+N(0,\sigma)-D_i||^2}{2\sigma^2}}} \tag{23}$$

After getting the likelihood ratios for every location, we combine across locations by taking the maximum likelihood ratio as the final activation for a single trial. We simulate 40000 trials, storing the activations. After simulating these trials, we compare the activations to a criterion (chosen to maximize model performance) to provide model predictions for every trial and an overall model performance.

Similar to the IO, the relationship between feature and conjunction search is defined within this model due to increased uncertainty from comparing with three templates at each location for conjunction search instead of only two for feature search. Therefore, we use the same fitting parameter as the IO, the standard deviation of the Gaussian white noise added to each trial.

2.6 Neural Network - Fully-Trained

We train small, 5-layer convolutional neural networks with three strided convolution layers, a fully connected layer, and an output layer, separately for each task (2 features and one conjunction). We follow an iterative training process where we start by training the network on stimuli with no noise, where it will reach ceiling performance across all set sizes. Here we train for 20 epochs each for 2 sets of 15000 randomly generated stimuli

(5000 each for set sizes 4,6, and 12). At the end of this initial training, we reached ceiling performance across all 12 set sizes. Afterward, we begin to add noise slowly, training for 5 epochs for 6 sets of 15000 randomly generated noisy stimuli (5000 each for set sizes 4, 6, and 12) at each noise level. The noise is increased 6 times before reaching the target noise level on the 7th iteration. We then repeat the 7th iteration another 5 times to ensure model convergence. Overall, the model sees $5.4 \cdot 10^5$ different noisy stimuli during the iterative increase in noise and another $5.4 \cdot 10^5$ different noisy stimuli at the target noise level. This training is done with the AdamW optimizer, with a batch size of 50 and a learning rate of 0.001, using a binary cross-entropy loss. We then obtain a final performance on 40000 noisy stimuli at the target noise level at every set size.

The relationship between feature and conjunction search is defined within this model as there are more possible training samples in conjunction search than in feature search. Since this relationship is defined, there is only one fitting parameter. Similar to the IO, the parameter is the added task difficulty or the standard deviation of the Gaussian white noise added to each trial.

2.7 Neural Network - Transfer Learning

For this model, we complete a transfer learning process utilizing a VGG-16 model pre-trained on ImageNet. We freeze the parameters in the convolution layers of this model and only train parameters in the final 2 dense layers and the output layer. The steps of this training and testing match the fully-trained networks, with the only difference being the architecture and trained parameters.

Similar to the fully-trained network, the relationship between conjunction and feature search is defined within the model. Therefore, we use the same single fitting parameter, the standard deviation of the Gaussian white noise added to each trial.

3 Results

We report the performance of 7 visual search models, 3 of which are image computable. For the image computable models, the target luminance and angle stayed constant across models (204 gray levels, 20° right of vertical), and the angle distractor's luminance and angle also stayed constant (204 gray levels, 10° right of vertical). The luminance distractors angle was also constant (20° right of vertical) but the luminance difference varied to match the performance of the two features (166 gray levels for BIO and ILC, 177 for the fully-trained CNN, and 186 for VGG). The background was always 128 gray levels. For all of the Single parameter models, we matched the feature performance at set size 1 to be approximately 98%. For cSDT, we used capacity values from Põder and Kosiło, 2019, and adjusted the d' until the performance for set size 1 was approximately 98%. For the guided search model, we set a search time of 5 objects, then fit the feature parameters to have approximately 98% performance at set-size 1, keeping the same parameters for conjunction search.

We first investigate feature search, investigating the set size effect and higher set sizes after only matching at set size 1:

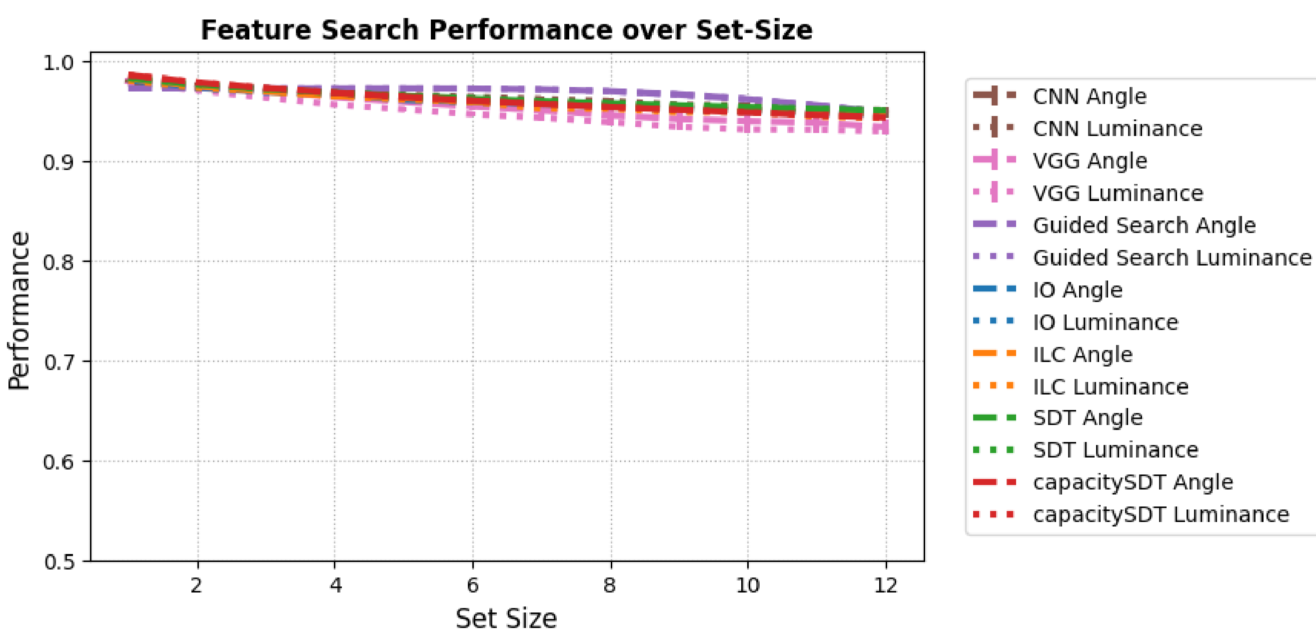

Figure 2

Performance for 7 models on the 2 feature search conditions, approximately matched at set-size 1

We see that the feature search predictions are extremely consistent across the models. Only the guided search and VGG models differ slightly from the rest at higher set-sizes. We report the parameters and set size effects (defined as the change in performance from set-size 1 to 12) in table 3.

Table 3

Model	Parameter Values	Feature Set Size Effect	Conjunction Set Size Effect
Ideal Observer (IO)	$\sigma = 85$ gray levels	3.4%	4.9%
Individual-Location-Comparison (ILC)	$\sigma = 85$ gray levels	3.5%	4.5%
Signal-Detection-Theory (SDT)	d' = 4.25	3.4%	10.2%
Signal-Detection-Theory with Capacity (cSDT)	d' = 4.4, $b_c = 0.8, b_f = 0.05$	4.6%	36.3%
Guided Search (GS)	d' = 2, k, $\lambda_c = \lambda_f = .4, g_c = g_f = .2$	2.4%	2.6%
Fully Trained Convolutional Neural Network (CNN)	$\sigma = 50$ gray levels	3.1%	6.8%
Pre-Trained Convolutional Neural Network (VGG)	$\sigma = 18$ gray levels	4.9%	9.5%

We see that the set size effect is also consistent across models for feature search. We do not see the same for conjunction search where the set size effects (table 3) and plots vary significantly:

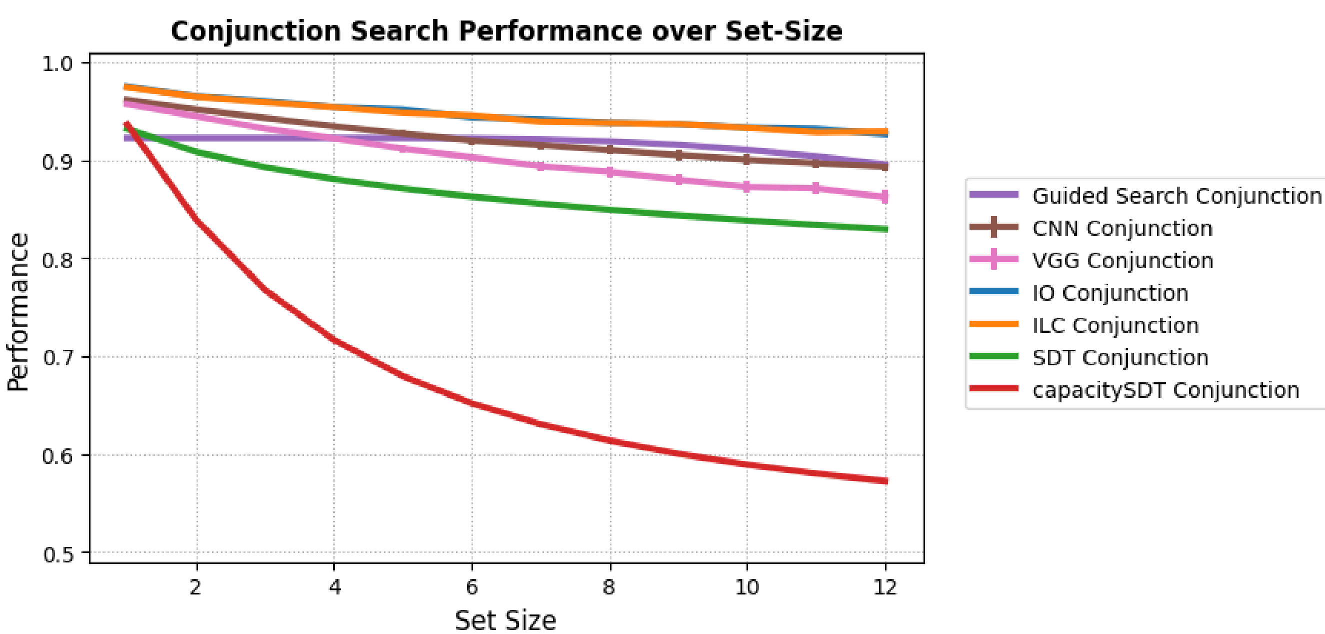

Figure 3

Performance for 7 models on conjunction search

4 Discussion

We implemented 7 different models of Visual search, 4 of which were image computable. The most unique of the 7 models is Guided Search. This is the only model that is able to have a non-convex prediction in its output. This is mainly due to its serial component which is not a part of any other model we explore. While the rest of the models are consistently convex, the CNN models have no mathematical requirement for this, and previous research has found some models that are concave and/or inverse set size effects (Nicholson and Prinz, 2022). Our models were consistently convex, and all 60 trained models had a normal set size effect. Out of the 4 models left, SDT with capacity is the only one that is able to modulate the set size effect independently for each search type. The remaining three models, the traditional SDT and the Bayesian models, all have a strict relationship between the set size effect of features and conjunctions are are unable to modulate the set size effect independently.

4.1 Image Computable vs Feature Extracted

While image computable models are able to model the processing stages, we see wildly different consequences of these stages from the varying luminance values for the distractor required to match the feature search performances. For the Ideal Observer models, differentiating angle is significantly easier than differentiating color. Since they are full-resolution models, they can process every large pixel difference that even a small shift in angle can cause. The CNN models are unable to process this small shift in angle as the image is significantly downscaled while it is being processed by the network. This is most apparent in the VGG model that not only downscales but is trained to process natural images which rarely require the angular resolution needed for this task.

4.2 Code Availibility

This work creates an important baseline on how testing a large number of models in visual search should be approached. Furthermore, we provide an implementation of the 7 models, either analytically or through simulation, which can be found on GitHub (https://github.com/anshksoni/Search).

References

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological science*, *9*(2), 111–118.

Eckstein, M. P., Beutter, B. R., & Stone, L. S. (2000). *Analytic guided-search model of human performance accuracy in target-localization search tasks* (tech. rep.).

Eckstein, M. P., Thomas, J. P., Palmer, J., & Shimozaki, S. S. (2000). A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Perception & psychophysics*, *62*, 425–451.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision research*, *51*(7), 771–781.

Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley New York.

Konkle, T., & Alvarez, G. A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, *13*(1), 491.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.

Ma, W. J., Navalpakkam, V., Beck, J. M., Berg, R. v. d., & Pouget, A. (2011). Behavior and neural basis of near-optimal visual search. *Nature neuroscience*, *14*(6), 783–790.

Nicholson, D. A., & Prinz, A. A. (2022). Could simplified stimuli change how the brain performs visual search tasks? a deep neural network study. *Journal of Vision*, *22*(7), 3–3.

Palmer, J. (1994). Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks. *Vision research*, *34*(13), 1703–1721.

Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision research*, *40*(10-12), 1227–1268.

Põder, E., & Kosiło, M. (2019). What limits search for conjunctions of simple visual features? *Journal of Vision*, *19*(7), 4–4.

Posner, M. I. (1980). Orienting of attention. *Quarterly journal of experimental psychology*, *32*(1), 3–25.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srivastava, S., Wang, W., & Eckstein, M. P. (2021). A feedforward convolutional neural network with a few million neurons learns from images to covertly attend to cues and context like humans and an optimal bayesian observer.

Swensson, R. G., & Judy, P. F. (1981). Detection of noisy visual targets: Models for the effects of spatial uncertainty and signal-to-noise ratio. *Perception & Psychophysics*, *29*(6), 521–534.

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

Wolfe, J. M. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, *28*(4), 1060–1092.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.