

Efficient Zero-shot Visual Search via Target and Context-aware Transformer

Zhiwei Ding^{1,2,*}, Xuezhe Ren^{3,*}, Erwan David⁴, Melissa Vo⁴, Gabriel Kreiman^{5,6}, Mengmi Zhang³

* Equal contribution

¹ Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA,

² Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, TX, USA,

³ CFAR and I2R, Agency for Science, Technology and Research, Singapore,

⁴ Department of Psychology, Goethe University Frankfurt, Frankfurt, Germany,

⁵ Children’s Hospital, Harvard Medical School, Boston, MA, USA,

⁶ Center for Brains, Minds and Machines, Cambridge, MA, USA,

Address correspondence to mengmi@i2r.a-star.edu.sg

Abstract

*Visual search is a ubiquitous challenge in natural vision, including daily tasks such as finding a friend in a crowd or searching for a car in a parking lot. Humans rely heavily on relevant target features to perform goal-directed visual search. Meanwhile, context is of critical importance for locating a target object in complex scenes as it helps narrow down the search area and makes the search process more efficient. However, few works have combined both target and context information in visual search computational models. Here we propose a zero-shot deep learning architecture, **TCT** (Target and Context-aware Transformer), that modulates self attention in the Vision Transformer with target and contextual relevant information to enable human-like zero-shot visual search performance. Target modulation is computed as patch-wise local relevance between the target and search images, whereas contextual modulation is applied in a global fashion. We conduct visual search experiments on TCT and other competitive visual search models on three natural scene datasets with varying levels of difficulty. TCT demonstrates human-like performance in terms of search efficiency and beats the SOTA models in challenging visual search tasks. Importantly, TCT generalizes well across datasets with novel objects without retraining or fine-tuning. Furthermore, we also introduce a new dataset to benchmark models for invariant visual search under incongruent contexts. TCT manages to search flexibly via target and context modulation, even under incongruent contexts.*

1. Introduction

When searching for toilet paper, we usually expect a bathroom scene and will likely search through regions around the toilet or the sink for a white and short cylinder-shaped object. If provided with a kitchen scene that is irrelevant to our object of interest, we might feel lost about where to focus on and simply search randomly throughout the scene; nonetheless, our expectation of toilet paper’s visual appearance will still lead us to find it without too much difficulty (**Fig. 1**). Humans rely heavily on knowledge about the target object and the surrounding context in daily visual search tasks. Target information provides reference of specific visual features to guide attention during the search, whereas context information helps narrow down the search area and guide attention towards more relevant locations. Both sources of information work coherently to make visual search more efficient.

Visual search constitutes a ubiquitous challenge in daily life and localizing a target object in a complex scene is important in many applications. Visual search must fulfill four key properties: (1) selectivity (to distinguish the target from distractors in a cluttered scene), (2) invariance (to localize the target despite differences in its appearance or in cases where the target is partially occluded), (3) efficiency (to localize the target with a few eye movements instead of exhaustive sampling) and (4) zero-shot (to generalize to novel target objects despite minimal or zero prior exposure to them). Given a search image and a target object, it remains unclear how to build a computational model that capitalizes on both target and contextual cues to efficiently localize the target object and generalize to various search conditions without additional training. Furthermore, investigation on the impact of

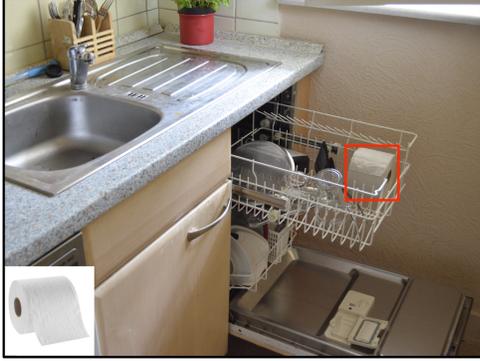


Figure 1. **Searching for targets in semantically incongruent scenes is inefficient.** The image describes a kitchen scene photograph with a red bounding box indicating the location of target object (toilet paper, shown also in the bottom left corner). The image has a semantic object-scene violation: the target object does not fit into the global meaning of the scene (i.e. toilet paper inside a dishwasher in kitchen).

incongruent context on invariant visual search performance is limited, largely due to the lack of such dataset. Synthetic out-of-context images have been introduced previously [7], but those images are not realistic.

Here we propose a zero-shot deep learning architecture, the Target and Context-aware Transformer (TCT), that integrates both target and contextual information to modulate self-attention via a multi-head transformer encoder as in the Vision Transformer (ViT) [14]. Target modulation is computed as the patch-wise local relevance between the target and search images, whereas contextual modulation is applied in a global fashion. Unlike the existing visual search models that learn to predict fixation sequences based on human eye tracking data [44], our model is zero-shot where it directly transfers knowledge from contextual reasoning in object recognition tasks to visual search.

We test our model on three image datasets spanning various levels of difficulty. TCT beats the SOTA models in highly difficult datasets, handling complex scenes with clutter distractors and incongruent contexts without task-specific fine-tuning. With built-in target and context modulation, TCT can flexibly use both sources of information to strategize for high search efficiency.

In summary, our key contributions include:

[model] We propose a biologically inspired search model with a hierarchical deep architecture based on the pre-trained ViT. Without additional training, the model can efficiently find novel target objects in cluttered natural scenes, despite changes in the target’s color, scale, rotation or even when the search scene contains a different exemplar belonging to the same category as the target.

[dataset] We introduce a new image dataset to benchmark invariant visual search under natural incongruent contexts.

[zero-shot and human-like] Our model generalizes to search for targets efficiently in various context conditions, beats baseline models by a large margin, and demonstrates human-like visual search performance.

2. Related Works

Visual search in daily life is particularly challenging since the instance of the target object in the search scene could vary dramatically due to unexpected transformations like rotation, scale, color, occlusion, etc. Previously, a large body of behavioral [32, 36, 37, 42] and neurophysiological [4, 9, 12, 38] studies have focused on *identical target* visual search, deviating largely from the complexities of real-world scenes. Recent works extend these efforts to establish human visual search benchmarks in approximately naturalistic conditions, where the search scenes were highly cluttered and the target information were provided as the text of the object category [44] or an image instance different from the one in the search scene [48].

2.1. Bottom-up Saliency Prediction

Certain parts of an image automatically attract attention due to color, contrast, or spatiotemporal changes. Several models have been proposed to capture such bottom-up saliency [8, 16, 19, 21, 25, 45, 46], recently using deep convolution networks [20, 27, 28, 30]. These models do not include information about the sought target.

2.2. Object Detection

Many works [15, 34] use object detection algorithms [18] to perform visual search. Object detection models generate bounding boxes around several objects of interest. Generally, models for object detection include a convolutional neural network backbone with additional engineered components that use the output of the network to produce candidate bounding boxes. Compared to human visual search, object detection models employ brute-force strategies which are not driven by top-down cues. Critically, object detection algorithms are not able to generalize to search for *novel* objects, but rather require extensive training with the target objects.

2.3. Goal-directed Invariant Visual Search

Humans rely heavily on the provided target as the task goal during visual search; such target-dependent modulation is likely originated from the prefrontal cortex [4, 33] and projected onto lower-level visual cortical structures [5, 31]. There have been only a few attempts in using deep learning models to predict human fixation sequences in visual search tasks [1, 41, 44, 48], all of which were inspired by top-down modulation in humans. [48] developed a convolutional neural network (CNN) that

processes search and target image independently, where the attention map for visual search is computed by convolving the top-level search and target feature representation. While this model demonstrates high resemblance to human behavior in terms of search efficiency and scanpath history, it solely focuses on target-based modulation without considering other sources of information present in the search scene such as the surrounding context. Another work [44] uses Inverse Reinforcement Learning (IRL) to learn human scanpaths by treating each fixation as a potential source of reward and trains a fixation sequence generator to maximize the reward. This model, however, requires supervised learning with human eye tracking data, which is labor-intensive and economically costly to obtain. Additionally, since the model was trained to perform the task using 18 pre-defined object categories, it could hardly generalize to all visual search conditions, such as novel categories outside of the training domain or objects without category labels.

While CNNs have been widely adopted in computer vision algorithms, Transformer-based architectures – particularly, the Vision Transformer (ViT) [14] – has recently outperformed CNNs in object classification. Importantly, while ViT has not been constructed with explicit invariance properties like CNNs (with built-in translation invariance), it demonstrates stronger robustness against input perturbations including natural corruptions, real-world distribution shifts, and natural adversarial examples [3] when trained on sufficient amount of data. Such robustness against a broad range of perturbations is essential for invariant visual search.

The proposed TCT is more biologically plausible than other search models, as it integrates two essential properties that guide human searching behavior: target and context. These human-inspired attention modulators are incorporated into the pre-trained ViT architecture. Moreover, TCT does not require additional training or fine-tuning on human eye tracking data or ground truth labels for visual search tasks.

2.4. Contextual Modulation

A few psychophysics experiments have been carried out to establish human benchmarks in in- and out-of-context object recognition and showed that incongruent contextual cues impair recognition performance [7, 49]. Computer vision models in object recognition can learn co-occurrence statistics between an object’s label and visual appearance but also its label and surrounding context [2, 13, 39]. The multi-head self-attention blocks in ViT [14] compute long-range interactions between local image patches, by definition carrying statistical information between the object of interest and the surrounding context. A recent work [7] introduced a Context-aware Recognition

Transformer Network (CRTNet) that processes object and context information in two independent streams before integrating them in subsequent decoder modules. With an additional confidence estimator that balances the weights of the two streams, CRTNet could recognize out-of-context objects robustly. However, we still lack such computational models for visual search tasks.

The aforementioned IRL model [44] attempts to integrate contextual information in visual search by including features of the target as well as other “anchor objects” [6, 40] and background scenes to model the internal belief state. The belief state representation is computed based on “contextual beliefs” of a limited number of object and scene categories extracted from a panoptic segmentation model [22, 23] and gets updated dynamically throughout the search process. The segmentation model, however, requires additional supervised training if the search task is to be generalized to novel object categories. Overall, such an approach assumes that humans have an internal scene parser that segments the search scene into object and scene classes. Our model, on the other hand, does not make any presumption upon the mechanism of the contextual reasoning process and does not require segmentation. Instead, TCT transfers the contextual knowledge of the search scene obtained implicitly during object recognition tasks to visual search tasks.

3. Our TCT Model

We propose a Target and Context-aware Transformer (TCT) to perform naturalistic visual search tasks with high efficiency. TCT takes in a target object \mathbf{I}_T and a search scene \mathbf{I}_S and extracts feature representation of the target and the context independently. The extracted features are then applied onto each individual transformer self-attention layer l in the form of target modulation $\mathbf{M}_{T,l}$ and context modulation $\mathbf{M}_{C,l}$ to guide self-attention map \mathbf{SA}_l . We call each of these modulated self-attention layers a Target and Context-aware Attention Block (TCAB). The visual search task is performed on the attention map extracted from the final TCAB.

We design the TCT architecture closely following that of the original Vision Transformer (ViT). This setup allows us to borrow the pre-trained ViT directly and easily re-purpose it into a visual search model without additional training. A schematic of TCT is shown in Fig. 2. Each of the self-attention layers in the original ViT is converted into a TCAB, while the remaining architecture stays completely intact and thus is not elaborated here.

3.1. Feature Extraction for Target

We take the Vision Transformer (ViT) pre-trained on ImageNet-21K and fine-tuned on ImageNet-1K [11] with the official checkpoints from Google. We further fine-tune

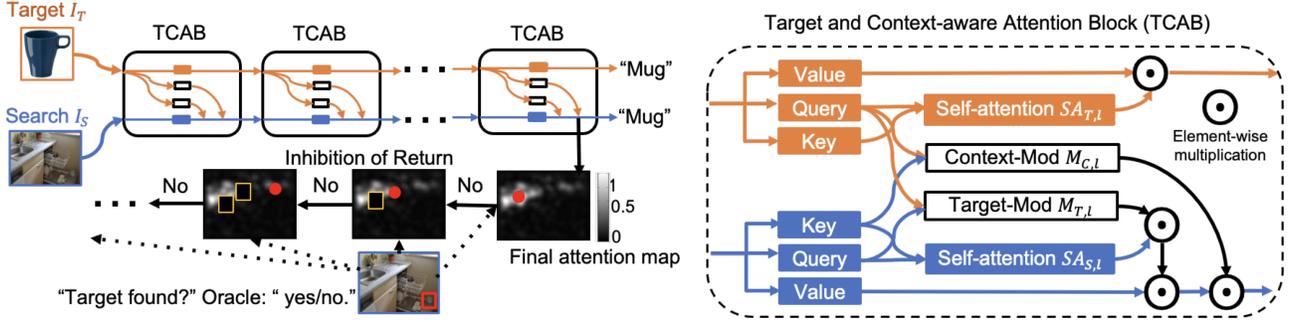


Figure 2. **Architecture for Target and Context-aware Transformer (TCT)**. TCT takes in a target object I_T and a search scene I_S and extracts feature representation of the target and the context independently. The extracted features are applied onto each Target and Context-aware Attention Block (TCAB) in the form of target modulation $M_{T,l}$ and context modulation $M_{C,l}$ to guide attention and in turn produce a final attention map. The model produces fixations by orderly selecting the maxima of the attention map. If the fixated area overlaps with the target bounding box, the search process ends. Otherwise, inhibition-of-return (IOR) is applied to the attention map by reducing the activation response within the fixated area to zero permanently in an area of pre-defined size. The process is repeated until the target is found. Red dots denote predicted eye fixations.

the ViT on CIFAR-100 [26] while explicitly removing the position embeddings in order to exclude the spatial information in feature representation that is irrelevant for visual search tasks.

The transformer reshapes an image of the target object $I_T \in \mathbb{R}^{C \times W \times H}$ into a sequence of image patches $\mathbf{p}_T \in \mathbb{R}^{C \times N \times P^2}$, where H and W is the resolution of the original image, C is the number of channels, N is the number of patches given a pre-defined patch size $P \times P$. The patches are then converted into a sequence of patch embeddings $\mathbf{h}_{T,1} \in \mathbb{R}^{N \times D}$, where D is the size of the hidden states in each transformer layer. In each layer l , the hidden states $\mathbf{h}_{T,l}$ are mapped through a linear projection $U_{QKV,l}$ into query $\mathbf{Q}_{T,l}$, key $\mathbf{K}_{T,l}$, and value $\mathbf{V}_{T,l}$ representations,

$$[\mathbf{Q}_{T,l}, \mathbf{K}_{T,l}, \mathbf{V}_{T,l}] = \mathbf{h}_T U_{QKV,l} \quad (1)$$

each one of which is then reshaped into $N \times D_h$ for computing multihead self-attention, where D_h is the hidden state size of each attention head. We retain the target query from each layer $\mathbf{Q}_{T,l}$ to represent the target feature and apply it in the corresponding layer to modulate attention when passing the search image I_S through the transformer (Sec. 3.3).

3.2. Feature Extraction for Context

We extract contextual feature representations from a modified version of the CRTNet [7], an object classification transformer that integrates object information and contextual reasoning with cross-attention mechanisms. The original CRTNet architecture remains intact except that the confidence weighting system is pruned away, enforcing the network to learn more context information that benefits object recognition. For each pair of context image I_S (i.e. the search image in visual search settings) and

target image I_T , the modified CRTNet extracts context feature embeddings from a stack of transformer decoder layers. Subsequently, a context classifier, which consists of fully-connected layers and a softmax layer, takes the feature embedding from the last decoder layer and outputs the predicted class label probabilities $\mathbf{y}_{T,C}$ among C object classes. The modified CRTNet is trained end-to-end using cross-entropy loss with respect to $\mathbf{y}_{T,C}$.

To further boost the contextual reasoning ability during object recognition, we introduced two changes for CRTNet training: (1) We expand the COCO-18 training dataset (Sec. 4.1) with additional images from corresponding 18 object categories in MSCOCO [29], generating a total of 201,383 target-context image pairs. (2) The original CRTNet was trained on target-context image pairs where the target image was simply cropped out from the context image. We perform 4 augmentations on the target-context pairs, combining 2 augmentations on target image I_T and 2 augmentations on context image I_S . Augmentations on target images I_T include: (1) the original target image I_T is replaced with a random target image belonging to the same target category; (2) the original target image I_T is blacked out with pixel value 0. Augmentations on context images I_S include: (1) the target image I_T within the original context image I_S is cropped out and replaced with a random target belonging to the same category as I_T ; and (2) the target image I_T within the original context image I_S is blacked out with pixel value 0.

To explicitly test whether the contextual feature extracted from the modified CRTNet contains useful information for object recognition, we check the top-1 recognition accuracy for the augmented target-context pairs where both the target itself and the target within the context images are blacked out (Tab. S1). Although the average accuracy for such inputs is much lower compared

to the original intact target-context pairs (85% vs 99%), it is still significantly higher than chance level ($1/18 = 5.5\%$). This suggests that the surrounding context by itself carries statistical information that enables CRTNet to make reasonable inferences about the target object identity. Here we transfer such useful contextual information directly to our TCT model to aid visual search tasks.

3.3. TCTB

The transformer first takes in a search scene I_S and extracts its query $\mathbf{Q}_{S,l}$, key $\mathbf{K}_{S,l}$, and value $\mathbf{V}_{S,l}$ representations similarly as for the target image (Sec. 3.1, Eqn. 1). Within each TCAB, the unmodulated self-attention \mathbf{SA}_l is computed as in the original ViT:

$$\mathbf{SA}_l = \text{softmax}\left(\mathbf{Q}_{S,l}\mathbf{K}_{S,l}^\top/\sqrt{D_h}\right). \quad (2)$$

Intuitively, target modulation is computed as patch-wise local relevance between the target and search image, whereas contextual modulation is applied onto the attention output in a global fashion. For target modulation, we aim to locate the search patches that contain similar visual features as the target patches. This is implemented by first computing the cross attention between the target query $\mathbf{Q}_{T,l}$ and the search query $\mathbf{Q}_{S,l}$ and then voting for the most relevant search patch for each target patch \mathbf{p}_T . Voting is implemented as assigning the top one most relevant search patch with relevance 1 and all others with relevance 0 (denoted as $\text{TOP1}(\cdot)$).

$$\mathbf{M}_{T,l} = \text{TOP1}\left(\text{softmax}\left(\mathbf{Q}_{T,l}\mathbf{Q}_{S,l}^\top/\sqrt{D_h}\right)\right) \quad (3)$$

The resultant binary target-search relevance map serves as our target-driven attention modulator $\mathbf{M}_{T,l}$, which is applied onto the original search image self-attention weights \mathbf{SA}_l in an element-wise fashion (denoted as \odot) to achieve target-modulated attention weights $\mathbf{A}_{T,l}$.

$$\mathbf{A}_{T,l} = \mathbf{M}_{T,l} \odot \mathbf{SA}_l \quad (4)$$

The attention output computed based on target-modulated attention weights $\mathbf{A}_{T,l}$ and the search value $\mathbf{V}_{S,l}$ is then further modulated by the contextual modulator $\mathbf{M}_{C,l}$ (detailed in Sec. 3.2) to achieve the updated hidden states $\mathbf{h}_{S,l+1}$ that serve as the input for the next TCAB.

$$\mathbf{h}_{S,l+1} = (\mathbf{M}_{C,l} + \mathbf{1}) \odot (\mathbf{h}_{S,l} + \mathbf{A}_{T,l}\mathbf{V}_{S,l}) \quad (5)$$

While both target and contextual modulation happen within each TCAB, their influence propagates through the entire multi-block hierarchy of the transformer. Importantly, the class token is exempt from the target and contextual modulation operations within each block but

does inherit the modulation effects from previous blocks. We use the class token averaged across all attention heads from the last block as the final attention map to perform visual search tasks.

3.4. Fixation Sequence Generation

The ground truth location is defined as the bounding box of the target object, i.e., the smallest rectangle encompassing all pixels of the object. A fixation sequence is generated by iteratively selecting the maximal point in the attention map as the fixation center. For each fixation, we check whether the fixated area – a rectangle of pre-defined size centered at the fixation center – overlaps with the target bounding box. If so, the search process ends; otherwise, inhibition-of-return (IOR) [24] is applied to the attention map by reducing the activation response within the fixated area to zero. This reduction is permanent, in other words, infinite memory is assumed for inhibition of return. The search continues until the target is found. The oracle approach of judging whether the target has been located is a simplification; in the future, the model could include a recognition module to check whether that target can be detected at a certain fixation [48].

3.5. Implementation Details and Code Availability

Both target and contextual modulation in TCT are implemented on individual layers of the ViT and therefore it is easy to investigate the contribution of each type of modulation in different combinations of layer configurations (see layer ablation results in Sec. 5.4). TCT achieves highest search efficiency when target modulation is implemented at $l \in [6, 12]$ and context modulation solely at $l = 3$. The target or context modulator ($\mathbf{M}_{T,l}$ or $\mathbf{M}_{C,l}$) is set to a matrix of all ones $\mathbf{1}^{N_S \times N_S}$ or $\mathbf{1}^{N_S \times D}$, respectively, if no such modulation is implemented at the corresponding layer l , where N_S is number of patches in I_S and D is the hidden state size. We perform the visual search task using search image resolution and IOR size consistent with previous works that introduced the image datasets [44, 48]: 512×320 pixels as search image and 48×48 pixels as IOR for the COCO-18 and SCEGRAM datasets, and 1280×1024 as search image and 200×200 as IOR for the NatClutter dataset, respectively (Sec. 4.1). All the source code is available at [here](#).

4. Experiments

4.1. Datasets

COCO-18 [10] is a visual search dataset consisting of about 300,000 eye movements from 10 human subjects in searching for one of 18 target object categories in 6,202 natural indoor scenes from MSCOCO dataset [29]. The target object is congruent with the context in each natural

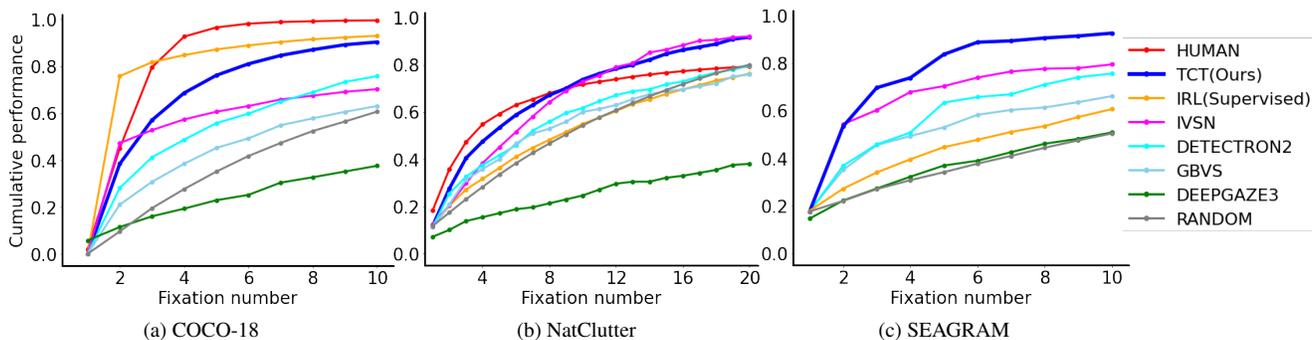


Figure 3. **Cumulative Performance of TCT and other baseline models on COCO-18 (a), NatClutter (b), SEGRAM (c).** Cumulative search performance is calculated as the cumulative probability $p(n)$ to find the target in $\leq n$ fixations for humans (red), TCT (blue) and other competitive methods (see Sec. 4.2). Human fixation data is not available in the SCEGRAM dataset.

| Attributes | # Subjects | # Trials | Target Ratio | Incongruency |
|-----------------|------------|----------|--------------|--------------|
| COCO-18 [44] | 10 | 612 | 0.04 | No |
| NatClutter [48] | 15 | 240 | 0.01 | No |
| SCEGRAM [35] | N/A | 187 | 0.01 | Yes |

Table 1. **Dataset summary.** Characteristics of the three image datasets used for performing visual search tasks: number of human subject with eye tracking data (“# Subjects”), number of search trials (“# Trials”), average ratio between the target and search image size (“Target ratio”), existence of search scenes with incongruent contexts (“Incongruency”).

image. As the target images are not provided in the original dataset, we pair each search image with a randomly selected target image from the same object category as the one present in the original search image.

NatClutter [48] consists of 240 invariant target-search image pairs in complex scenes with congruent contexts. The sought target and the given target differ in geometric transformations. The target object category could be novel to all models of interest in the current study.

SCEGRAM [35] consists of 62 indoor target objects, each contextualized in 1 congruent search scene and 5 additional scenes with semantic and/or syntactic violations. The target rendered within the search images and the provided target image originate from the identical objects. To enable invariant visual search, we collect a novel set of 62 target objects varying in appearances as well as geometric transformations from the sought targets in the search images. Additionally, since context modulation is more important for smaller objects [7, 49], we sort the search images according to the ratio of target versus search image size in ascending order and only consider the bottom 50 percentile in the current study. In the modified SCEGRAM dataset, we have 187 images in total, with 32 congruent conditions and 155 incongruent conditions.

4.2. Baselines and metrics

IVSN [48] is a zero-shot visual search CNN with top-down target-driven modulation and outputs an attention map in each search trial. **IRL** [44] is an inverse reinforcement learning model trained on human fixation sequences on COCO-18 to predict visual search scanpaths. When using IRL to perform search on objects without explicit category labels such as in NatClutter and SCEGRAM, we iterate through the 18 training categories to compute a category-average attention map. **Detectron2** [43] has a collection of object detection algorithms. We used ResNet+FPN pre-trained on MSCOCO [29] as the backbone for visual search. The attention map is computed as the average of sigmoid objectness maps extracted from the RPN component. **GBVS** (Graph-based Visual Saliency) [17] is a bottom-up saliency model predicting attention maps based on low-level statistics of the search images. **DeepgazeIII** [28] is a bottom-up saliency model trained with human eye fixations in free-viewing tasks and outputs bottom-up attention maps for search images. We also consider a **Random** model that populates attention maps with values randomly drawn from a uniform distribution. This randomly sampling process is repeated 100 times.

For all the baseline models above, we generate a sequence of fixations by iterating through the maxima of the attention maps while applying IOR throughout the search process (see Sec. 3.4). We evaluate all computational models with the following evaluation metrics: **Cumulative performance across fixations** is calculated as the cumulative probability $p(n)$ that a human subjects or a computational model finds the target object in $\leq n$ fixations. We also compute **Average number of fixations** until the target object has been found.

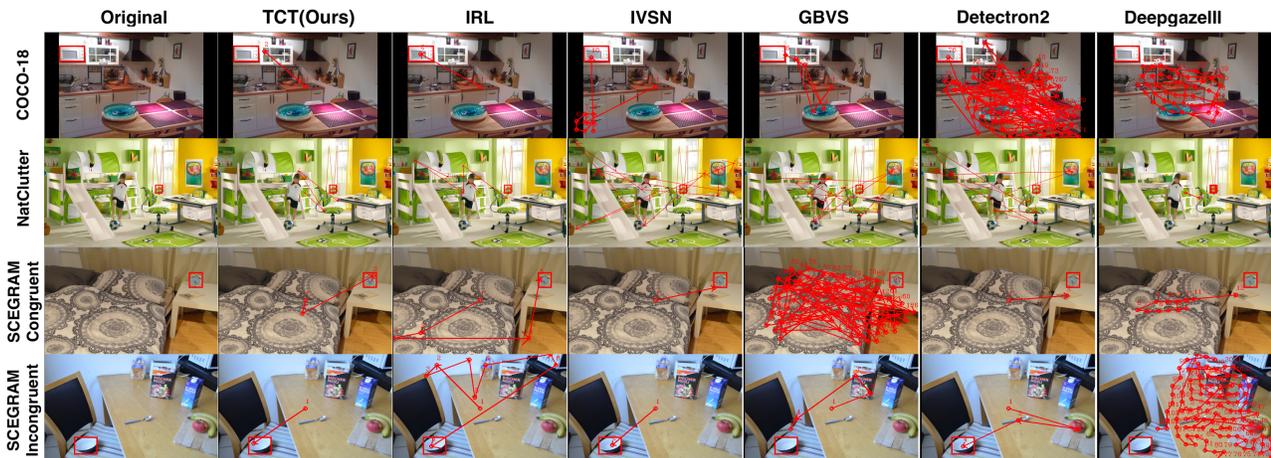


Figure 4. **TCT performs visual search tasks efficiently in multiple datasets and across different context conditions.** Visual search scanpaths of TCT (column 2) and other baseline models (column 3 to 7) on COCO-18 (row 1), NatClutter (row 2), SCEGRAM Congruent (row 3) and SCEGRAM Incongruent (row 4) datasets. The red bounding box shows the target object. Red dots denote predicted eye fixations. Red arrows indicate predicted eye movement directions.

5. Results

5.1. Visual Search on Natural Images

We first evaluate the visual search performance of TCT against the human benchmark as well as other baseline visual search models on natural images in the COCO-18 dataset (Fig. 3a). As shown in the example scanpaths in Fig. 4 row 1, TCT is able to locate the target object rapidly despite the object is far away from the initial fixation position (the center of the image). TCT manages to find the object using only 1 fixation for 38% of all trials and no more than 10 fixations for 92% of all trials.

While TCT behaves worse than human in the initial fixations, its cumulative performance increases rapidly across fixations and reaches close to human performance within 10 fixations. IRL serves as another upper bound performance for this task, since it was trained specifically on this dataset using human fixation sequence to mimic human search behavior.

Among all of the zero-shot visual search models, our TCT achieves the highest search efficiency (Fig 3a). In the example kitchen scene, TCT efficiently locates the microwave with 1 fixation (Fig 4), while other baseline models took more fixations to find the target. The outstanding performance can be most likely attributed to its human-inspired integration of both target and context information in guiding search attention. In contrast to TCT, IVSN is ignorant of any contextual cues and solely relies on target information to generate attention, leading to lower search efficiency. TCT is also notably better than Detectron2 since the latter utilizes object detection algorithms and is not aware of top-down modulation from the object of interest. GBVS and DeepgazeIII also

exhibit evidently lower performance, both of which are bottom-up saliency models that do not take the target object information into account.

5.2. Visual Search in Cluttered Environments

To evaluate visual search performance on more complex scenes where the target objects are smaller in size and possibly occluded by multiple distractors, we test TCT and other baseline models on the NatClutter dataset. Example scanpaths in row 2 of Fig 4 demonstrate a challenging search trial with a small teddy bear embedded in a highly cluttered scene with several distractors. Although the target object appears much less salient than in COCO-18, our TCT is still able to pinpoint the target position within a few fixations. TCT successfully locates the object using only 1 fixation for 28% of all trials and no more than 20 fixations for 92% of all trials.

While human subjects outperform TCT in the first 8 fixations, TCT surpasses human in subsequent fixations. One reason for this difference is that humans do not have infinite memory and potentially revisit the same fixation areas [47], whereas permanent IOR has been applied in all computational models, avoiding exhaustive search with repeated visits. Meanwhile, The cumulative search performance of IRL dropped dramatically in NatClutter compared to COCO-18, since IRL has been implemented to perform the search only within its limited training categories within COCO-18 and therefore fails to generalize to novel object categories or unlabelled objects. TCT manages to outperform IRL by a large margin due to its ability to generalize to novel target objects with minimal or even zero prior exposure.

In general, TCT performed better than other baseline

| Avg Fixations | Overall | Congruent | Incongruent |
|---------------|------------|------------|-------------|
| TCT(ours) | 4.2 | 3.6 | 4.9 |
| IRL | 12.8 | 8.9 | 16.8 |
| IVSN | 6.6 | 5.2 | 8.1 |
| GBVS | 15.3 | 10.7 | 19.8 |
| Detectron2 | 11.7 | 5.6 | 17.8 |
| Deepgaze3 | 24.1 | 18.3 | 29.8 |

Table 2. Average number of fixations for TCT and baseline models under congruent and incongruent conditions in the modified SCEGRAM dataset. Best is in bold.

models. IVSN, particularly, achieves performance similar to TCT. Both TCT and IVSN use target modulation and generalize well to novel target objects. Nonetheless, the cumulative performance of IVSN was still lower than TCT possibly due to the lack of context modulation.

5.3. Visual Search in Incongruent Contexts

So far, we have focused on searching in natural scenes with contextual cues congruent with the target objects. Next, we assess whether target-context incongruency diminishes the search efficiency of visual search models using the modified SCEGRAM dataset (Sec. 4.1). Search trials in SCEGRAM are split into congruent and incongruent conditions and performance of each model is computed for the two different conditions separately (Table 2). As demonstrated in the example scanpaths (Fig. 4, rows 3-4), both TCT and IVSN exhibit small performance difference across congruent and incongruent conditions, whereas other baseline models suffer severely from incongruent contexts. These results agree with our expectation because both TCT and IVSN modulate attention with target information and therefore are somewhat protected against misleading contextual information, enabling efficient search even under incongruent conditions.

5.4. Ablation

To further dissect individual contribution of target and contextual modulation on search efficiency, we repeat the identical experiments on COCO-18 using different ablated versions of TCT and compute their average number of fixations (Fig. 5 left). When both target and context modulation are removed, TCT becomes the original ViT (“ViT”) and performs poorly as it simply functions as a bottom-up feature extractor and self-attention does not contain information relevant for the visual search task. To evaluate the importance of context modulation, we remove the global context modulation while keeping the local target modulation intact (“TargetAlone”). The TargetAlone model is still able to provide goal-directed information in a top-down fashion similar to its CNN counterpart IVSN [48], yet the performance drops in comparison to TCT due

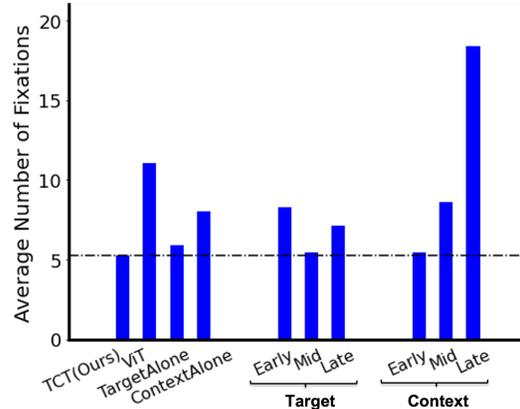


Figure 5. COCO-18 search performance for TCT with ablated components or with modulation applied on different layers. Performance is evaluated as the average number of fixations and the dash line indicates the best TCT performance out of all different ablations of TCT. Left 4 columns: our default TCT, TCT with target modulation alone (TargetAlone), TCT with context modulation alone (ContextAlone), the original ViT with no modulation (ViT). Mid 3 columns: TCT with target modulation applied starting from early, mid, and late TCAB blocks onwards, respectively. Right 3 columns: our TCT with context modulation applied at individual early, mid, and late TCAB blocks, respectively.

to lack of the contextual guidance. Similarly, we remove the local target modulation while leaving the global context modulation intact (“ContextAlone”). The ContextAlone model performs significantly worse than TCT, indicating that target-driven modulation is essential in guiding search attention. While context information guides attention on a global scale to emphasize regions of interest for contextual reasoning given a target object, by itself it does not provide information that is high resolution enough to pinpoint the object location.

We further investigate the layer specificity of target and contextual modulation by applying the modulation on different subsets of TCABs for each modulation type independently. The target modulation is applied starting from a certain block and continues across subsequent blocks. We group the blocks where the modulation happens into early ($l \in [1, 4]$), middle ($l \in [5, 8]$), and late ($l \in [9, 12]$) groups. For each group, we average the search performance. The best performance is achieved when target modulation is applied from the middle blocks onward (Fig. 5 middle), potentially because the early blocks in transformer mainly extract low-level nuisance visual properties such as contrast or color that do not serve as essential cues in visual search. Similarly, we apply the contextual modulation onto different individual layers and group the modulation blocks in the same fashion. TCT with contextual modulation in earlier blocks significantly outperforms than that in later layers (Fig. 5

right), aligning with the intuition that context functions as global guidance that redirects attention at the initial stage of visual processing.

6. Conclusion

We propose a biologically plausible deep learning architecture for visual search, **TCT** (Target and Context-aware Transformer). TCT integrates target object and contextual information via a multi-head transformer encoder to guide attention during visual search tasks. TCT transfers contextual knowledge learnt from object recognition to visual search in a zero-shot manner. TCT manages to approximate human search efficiency in invariant zero-shot visual search tasks with various levels of difficulty. Remarkably, in contrast with standard object detection approaches, TCT generalizes across tasks with novel target object search *without retraining or fine-tuning*. Furthermore, TCT outperforms the SOTA visual search models, especially in highly complex datasets and across congruent and incongruent contexts in a newly introduced dataset.

Supplementary Materials

| | Top-1 Accuracy(%) |
|-----------------------------------|-------------------|
| Original Context, Original Target | 98.8 |
| Original Context, Random Target | 98.6 |
| Zero Context, Zero Target | 84.9 |

Table S1. **Top-1 object recognition accuracy of the context feature extractor on COCO-18.** Classification accuracy of the context feature extractor (see Sec. 3.2) under different context conditions in COCO-18 test dataset: (1) the target image I_T is cropped out directly from the context image I_S and the context image remains intact (“Original context, Original target”); (2) the original target image I_T is replaced with a randomly selected target image belonging to the same target category as the original I_T , while the context image I_S remains intact (“Original context, Random target”); (3) both the original target image I_T and the bounding box encompassing the target region on the original context image I_S is blacked out with pixel value 0 (“Zero context, Zero target”).

References

- [1] Hossein Adeli and Gregory Zelinsky. Deep-bcn: Deep networks meet biased competition to create a brain-inspired model of attention control. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1932–1942, 2018. 2
- [2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 3
- [3] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10231–10241, 2021. 3
- [4] Narcisse P Bichot, Matthew T Heard, Ellen M DeGennaro, and Robert Desimone. A source for feature-based attention in the prefrontal cortex. *Neuron*, 88(4):832–844, 2015. 2
- [5] James W Bisley. The neural basis of visual attention. *The Journal of physiology*, 589(1):49–57, 2011. 2
- [6] Sage EP Boettcher, Dejan Draschkow, Eric Dienhart, and Melissa L-H Võ. Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of vision*, 18(13):11–11, 2018. 3
- [7] Philipp Bomatter, Mengmi Zhang, Dimitar Karev, Spandan Madan, Claire Tseng, and Gabriel Kreiman. When pigs fly: Contextual reasoning in synthetic and natural scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 255–264, 2021. 2, 3, 4, 6
- [8] Neil DB Bruce and John K Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3):5–5, 2009. 2
- [9] Timothy J Buschman and Earl K Miller. Serial, covert shifts of attention during visual search are reflected by the frontal eye fields and correlated with population oscillations. *Neuron*, 63(3):386–396, 2009. 2
- [10] 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. 5
- [11] 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. 3
- [12] Robert Desimone, John Duncan, et al. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995. 2
- [13] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In *2009 IEEE Conference on computer vision and Pattern Recognition*, pages 1271–1278. IEEE, 2009. 3
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3
- [15] Miguel P Eckstein, Kathryn Koehler, Lauren E Welbourne, and Emre Akbas. Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology*, 27(18):2827–2832, 2017. 2
- [16] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006. 2
- [17] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19, 2006. 6
- [18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2
- [19] Xiaodi Hou, Jonathan Harel, and Christof Koch. Image signature: Highlighting sparse salient regions. *TPAMI*, 34(1):194–201, 2012. 2
- [20] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *IEEE ICCV*, pages 262–270, 2015. 2
- [21] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, (11):1254–1259, 1998. 2
- [22] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 3
- [23] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 3
- [24] Raymond M Klein. Inhibition of return. *Trends in cognitive sciences*, 4(4):138–147, 2000. 5
- [25] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987. 2
- [26] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4
- [27] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017. 2
- [28] Matthias Kümmerer, Matthias Bethge, and Thomas SA Wallis. Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5):7–7, 2022. 2, 6

- [29] TY Lin, M Maire, S Belongie, J Hays, P Perona, D Ramanan, P Dollár, CL Zitnick, et al. Microsoft coco: Common objects in context. *neuroeuropean conference on computer vision 2014 sep 6* (pp. 740-755). 4, 5, 6
- [30] Yuetan Lin, Shu Kong, Donghui Wang, and Yueting Zhuang. Saliency detection within a deep convolutional architecture. 2
- [31] Julio Martinez-Trujillo. Searching for the neural mechanisms of feature-based attention in the primate brain. *Neuron*, 70(6):1025–1028, 2011. 2
- [32] Thomas Miconi, Laura Groomes, and Gabriel Kreiman. There’s waldo! a normalization model of visual search predicts single-trial human fixations in an object search task. *Cerebral cortex*, 26(7):3064–3082, 2015. 2
- [33] Earl K Miller, Jonathan D Cohen, et al. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001. 2
- [34] David A Nicholson and Astrid A Prinz. Could simplified stimuli change how the brain performs visual search tasks? a deep neural network study. *Journal of Vision*, 22(7):3–3, 2022. 2
- [35] Sabine Öhlschläger and Melissa Le-Hoa Võ. Scegram: An image database for semantic and syntactic inconsistencies in scenes. *Behavior research methods*, 49(5):1780–1791, 2017. 6
- [36] Rajesh PN Rao, Gregory J Zelinsky, Mary M Hayhoe, and Dana H Ballard. Eye movements in iconic visual search. *Vision research*, 42(11):1447–1463, 2002. 2
- [37] Antonio J Rodriguez-Sanchez, Evgueni Simine, and John K Tsotsos. Attention and visual search. *International Journal of Neural Systems*, 17(04):275–288, 2007. 2
- [38] David L Sheinberg and Nikos K Logothetis. Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. *Journal of Neuroscience*, 21(4):1340–1350, 2001. 2
- [39] Jin Sun and David W Jacobs. Seeing what is not there: Learning context to determine where objects are missing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5716–5724, 2017. 3
- [40] Melissa Le-Hoa Võ, Sage EP Boettcher, and Dejan Draschkow. Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current opinion in psychology*, 29:205–210, 2019. 3
- [41] Zijun Wei, Hossein Adeli, Minh Hoai Nguyen, Greg Zelinsky, and Dimitris Samaras. Learned region sparsity and diversity also predicts visual attention. *Advances in Neural Information Processing Systems*, 29, 2016. 2
- [42] Jeremy M Wolfe and Todd S Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):1–8, 2017. 2
- [43] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 6
- [44] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 193–202, 2020. 2, 3, 5, 6
- [45] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 153–160, 2013. 2
- [46] 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):32–32, 2008. 2
- [47] Mengmi Zhang, Marcelo Armendariz, Will Xiao, Olivia Rose, Katarina Bendtz, Margaret Livingstone, Carlos Ponce, and Gabriel Kreiman. Look twice: A generalist model predicts return fixations across tasks and species. *PLoS Computational Biology*, page In Press, 2022. 7
- [48] Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any waldo with zero-shot invariant and efficient visual search. *Nature communications*, 9(1):1–15, 2018. 2, 5, 6, 8
- [49] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12985–12994, 2020. 3, 6