Bridging DFT and DNNs: A neural dynamic process model of scene representation, guided visual search and scene grammar in natural scenes

Raul Grieben (raul.grieben@ini.rub.de)
Ruhr-Universität Bochum, Institut für Neuroinformatik
Universitätsstraße 150, 44801 Bochum, Germany

Gregor Schöner (gregor.schoener@ini.rub.de)
Ruhr-Universität Bochum, Institut für Neuroinformatik
Universitätsstraße 150, 44801 Bochum, Germany

Abstract

We extend our previous neural dynamic models of visual search and scene memory (Grieben et al. (2020); Grieben and Schöner (2021)) to move beyond classical "laboratory" stimuli. The new model can autonomously explore a natural scene and build a scene memory of recognized objects and their locations. It is also capable of guided visual search for object categories in the scene. This is achieved by learning *object templates* for object recognition, and feature *guidance templates* for visual search and associating them to categorical concepts. We address how preattentive shape can be extracted from the visual input and how *scene guidance*, specifically, *scene grammar* (Võ, 2021), emerge. For the first time, we embed feature extraction by a headless deep convolutional neural network (CNN) in a neural dynamic (DFT) architecture by learning a mapping from the distributed feature representation of the CNN to the localist representation of a dynamic neural field.

Keywords: neural dynamic process model; dynamic field theory; visual search; natural scenes; scene guidance; scene grammar; supervised one shot continual online learning; convolutional neural network; deep neural networks;

Introduction

A farmer looking for a pig will probably not find it if it is flying around. Because pigs don't fly (which is why "when pigs fly" expresses impossibility). The world we live in is highly structured, and that structure induces expectations that strongly influence how we search and ultimately interact with objects in our environment (Võ (2021); Võ and Wolfe (2015); Hollingworth (2012)). Yet the vast majority of studies in visual perception use artificial visual scenes and simplified stimuli. These greatly advanced our understanding of the basic principles underlying visual search (Wolfe & Horowitz, 2017), but provided limited understanding of how humans search for real-world objects (Vickery, King, and Jiang (2005); Cunningham and Wolfe (2014)) in natural scenes (Hollingworth, 2012). The classical version of the perhaps most influential theory of visual search, "guided search 2.0" (GS) (Wolfe, 1994) did not directly transfer to natural scenes (Wolfe, Võ, Evans, & Greene, 2011). To address that, Wolfe, Võ, et al. (2011) proposed the concept of *scene guidance* and incorporated it in his updated theory "GS 6.0" (Wolfe, 2021). Wolfe (2021) distinguished between two categories of *scene guidance*: *syntactic guidance* (e.g.,"pigs don't fly") and *semantic guidance* (e.g.,"whales do not live in a closet"). A special case of such *scene guidance* has been formalized through the notion of *scene grammar* (Võ, 2021) based on experimental findings that *anchor objects* and their reproducible spatial relation to other objects (Boettcher, Draschkow, Dienhart, & Võ, 2018) enable humans to strongly reduce the area scanned in visual search (Võ, 2021). Alongside the increased interest in theories and models of *scene guidance* in the psychological domain (Castelhano & Krzyś, 2020), attention has also become an important topic in deep learning (see Niu, Zhong, and Yu (2021) for a review), although there is a considerable gap between the understanding of *attention* in these two fields. There have been attempts to incorporate top-down modulated guidance into deep feed-forward neural networks for visual search (Zhang et al. (2018); Gupta, Zhang, Wu, Wolfe, and Kreiman (2021)). In the brain, visual search and voluntary attention are both part of a top-down recurrent feedback loop (Knudsen, 2007). Unfolding this loop into a feed-forward network needs critical examination. We would argue that a model for visual search and attention always needs a recurrent feedback loop and stable memory representations to be able to interact with the environment in a goal-oriented way. Zelinsky et al. (2021) presented an inverse-reinforcement imitation-learning model that inferred reward functions and policies from target-specific behavioral fixation data. Their model generated reward maps that showed patterns which suggested not only feature guidance, but also guidance by scene context. We presented a neural process model for guided visual search and scene memory that did not only account for classical findings like feature vs. conjunction search, but also proposed answers to long-standing questions in the field of visual search: The influence of scene memory in the preview paradigm (Grieben et al., 2020) and the relationship between attention and feature binding (Grieben & Schöner, 2021) in the context of the unexpectedly efficient triple conjunction search (Nordfang & Wolfe, 2014). This model was limited to classical laboratory stimuli, however. Here we present a neural process model that substantially extends our previous models (Grieben et al. (2020); Grieben and Schöner (2021)) to natural scenes in a neurally plausible way. At the same time we incorporate a new neural process account for *scene grammar*. Enabling the model to interact with natural scenes required a major innovation, interfacing for the first time, a neural architecture based on Dynamic Field Theory (Schöner, Spencer, & DFT Research Group, 2016) with a pretrained headless deep convolutional neural network (CNN;

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

VGG16: Simonyan and Zisserman (2014)) that provides feature extraction. The interface is based on neurally plausible learning and makes it possible to combine the strength of the two frameworks. DFT delivers autonomous process organization, sequence generation, and working memory. CNNs are undoubtedly able to extract the complex features needed for object recognition, similar to what the ventral stream of human vision does. Neural populations in *inferior temporal cortex* (IT) represent object identity over space in a manner sufficient for object recognition (DiCarlo, Zoccolan, & Rust, 2012). Lim et al. (2015) inferred that the rule of synaptic plasticity rule observed for IT neurons is akin to the Bienenstock-Cooper-Munro (BCM) (Bienenstock, Cooper, & Munro, 1982) learning rule. We propose, therefore, to use the biologically plausible BCM rule to map the distributed representation of the CNN feature maps to the localist representation of a 3D neural field defined over space and object identity as found for IT neural populations. This enables the model to learn new object concepts from single exposures in a *supervised one-shot continual online learning* fashion (Mai et al., 2022). To guide visual search in natural scenes we also had to solve the problem of learning the association between categorical (label) concepts and their corresponding preattentive shape guidance features (Wolfe, 2021). Our solution is to extract these guidance features from an intermediate layer of the CNN model, as proposed by Wolfe (2021).

Methods

The neural process model is based on Dynamic Field Theory (DFT; Schöner et al. (2016)) a mathematical framework that characterizes graded activation patterns of neural populations that evolve continuously in time. The model also embeds a headless CNN as a feature extraction network.

Neural Dynamic Fields

Neural populations tuned to a metric dimension $\boldsymbol{x}$, e.g., a feature or movement parameter, are modeled as neural activation fields. The continuous evolution, on the time scale τ, of field activation emerges from the neural dynamics

$$\begin{aligned}\tau\dot{u}(\boldsymbol{x},t) = &-u(\boldsymbol{x},t)+h+s(\boldsymbol{x},t)+\xi(\boldsymbol{x},t)\\ &+\int \omega(\boldsymbol{x}-\boldsymbol{x}')\,\sigma(u(\boldsymbol{x}',t))\,d\boldsymbol{x}'\end{aligned} \tag{1}$$

in which the negative resting level, h, and external input, $s(\boldsymbol{x},t)$, define the attractor state of the system, if the overall activation level remains below the threshold of the sigmoidal nonlinearity, $\sigma(u) = 1/(1+\exp[-\beta u])$. In the case of sufficiently strong localized input, the system transitions to a supra-threshold peak of activation, which is described as the *detection instability*. Supra-threshold activation engages in neural interaction defined by the kernel, $\omega(\boldsymbol{x}-\boldsymbol{x}')$, that is excitatory over small, and inhibitory over large distances, $\boldsymbol{x}-\boldsymbol{x}'$, within the field. Additive neural noise $\xi(\boldsymbol{x},t)$ allows for stochastic switches between stable states near instabilities. Stable supra-threshold activation peaks are the units of representation in DFT. Depending on the kernel parameters, fields may operate in different dynamic regimes. In the *self-stabilized* regime, peaks are stabilized against decay and changes in input. In the *selective* regime, only a single peak is stable at any point int time. In the regime of *sustained activation*, peaks may persist after the localized input that induced them is removed.

Networks of fields

Cognitive processes and motor behavior are realized through networks of fields in which sequences of processing emerge from the underlying dynamic instabilities. Networks are defined by directional coupling among different fields and eventually to sensory-motor systems. Directional coupling or projection means that supra-threshold activation of one field provides either excitatory or inhibitory input to another field. Projections from lower-dimensional to higher-dimensional fields perform *dimensionality expansion* by providing ridge or slice input. The reverse projections perform *dimensionality contraction* through marginalizing by integration.

The neural dynamic process model

The neural dynamic process model shown in Fig. 1 is able to autonomously explore the visual scene and build a scene representation, the scene memory, of recognized objects, their features and their position in space. In the presence of one or more *search cues* (*P*) it is able to perform neurally plausible guided visual search for real-world object categories in natural scenes. It can also use known semantic structure of the scene, the *scene grammar*, to reduce the search space by inducing a relational bias centered on a detected *anchor object* (Võ, 2021). Regardless of its figurative depiction the model is a system of coupled neural integro-differential equations. Neural activation evolves continuously in time and discrete events emerge from instabilities in the dynamics. Outside of the feature extraction no further algorithms are used in the model. All different cognitive modes of the model emerge naturally from the neural dynamics. Its real-time numerical simulation is achieved by implementing it in *cedar* (Lomp, Richter, Zibner, & Schöner, 2016). In the following the different sub-networks of the model are explained while following along Figure 1 with sub-networks being referenced via uppercase letters.

Feed-forward feature and salience maps

The bottom-up pathway of the model is constituted by a parallel preattentive process that extracts low-resolution retinal (*C*) and high-resolution foveal (*J*) features in parallel from the input image(*D*). This is a simplified account for the two (ventral and dorsal) vision pathways in the human brain (Goodale & Milner, 1992).

Retinal Feature Extraction Preattentive *color* and preattentive *shape* are extracted from a scaled down version of the input image (*retinal image*) (*D*). Preattentive color is extracted from hue-space (*C*2). Preattentive shape is extracted

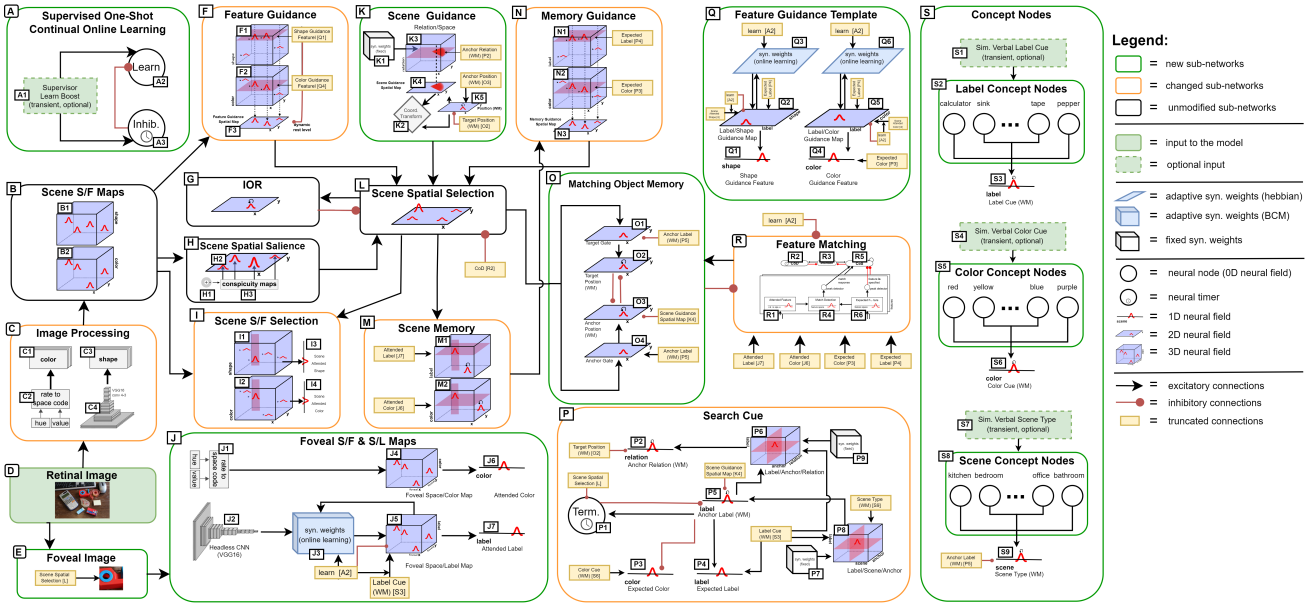

Figure 1: An overview of the neural dynamic process model.

from the intermediate *conv 4-3* convolutional layer of the VGG16 network (*C*4). Each feature filter generates the input to the corresponding *scene space/feature map* field (*B*). The feature *conspicuity maps* (*H*3) are the result of marginalization along the feature dimension of their corresponding *scene space/feature map* field (*B*), by using a center-surround filter (*H*1) as the projection kernel (see Grieben and Schöner (2021) for an explanation). These serve as the input to the *scene spatial salience* field (*H*2) that represents the non-linear spatial salience map (Itti & Koch, 2000) of the model.

Foveal Feature Extraction The *foveal image* (*E*) is extracted from the original input image at the currently attended location specified through the *scene spatial selection* field (*L*). From this attended *color* (*J*6) and *label* (*J*7) are extracted. Color is extracted from hue-space (*J*1) and serves as input to the *foveal space/color map* field (*J*4) that is defined over the two dimensions of foveal space and over one color feature dimension. Label is the result of a learned mapping from the last convolutional layer (*conv 5-3*) of the headless VGG16 network (*J*2) to the *foveal space/label map* field (*J*5). The marginalized activation over the color/label dimension of these foveal map fields (*J*4, *J*5) serves as the input to the *attended color/label* fields (*J*6, *J*7). These selective fields represent the dominant color/label at the attended location through a single peak along their feature dimension.

Attentional selection

Biased attentional selection is the basis for flexible, goal-driven, human behavior. Therefore, all visual cognitive processes in the model emerge from attentional selection. This is achieved by the *scene spatial selection* field (*L*) that selects a salient region in the visual scene through biased competition (Desimone & Duncan, 1995). It receives an excitatory bottom-up bias from the *scene spatial salience* field (*H*2) and three excitatory top-down guidance biases from the *feature guidance*[1] (*F*3), the coordinate transformed *scene guidance* (*K*2), and *memory guidance* (*N*3) spatial map fields. The *inhibition of return* (IOR) memory trace (*G*) steers the selection away from previously attended locations, enabling sequences of attentional selection.

Exploration and scene memory

One important key to understand human visual cognition is the incidental scene memory of attended objects, the scene representation, that humans continuously build (Draschkow, Wolfe, and Võ (2014); Hollingworth (2006)). For this reason, autonomous visual exploration of the scene is the default behavior of the model. Since working memory (WM) does not contain any cues, attentional selection is only biased by the bottom-up salience of the visual input. The model sequentially attends salient regions in the scene and commits the label, if a known object is recognized, and color at the currently attended location to the *scene memory* fields (*M*). These fields operate in the regime of *sustained activation* (see Grieben et al. (2020) for an explanation of capacity limits in these memory fields). The neural timer *CoD* (*R*2) inhibits the *scene spatial selection* (*M*) field, so that a new cycle of attentional selection arises.

Visual search

Most real-world interactions with objects entail a preceding visual search for it. This requires the neural activation of a *guidance template* (Wolfe, 2021) for the object in work-

[1] We corrected the misnomer *scene guidance* in the Grieben et al. (2020) model to *feature guidance* to prevent a name clash with the new correct *scene guidance* sub-network.

ing memory, which biases selection through top-down recurrent feedback loops. Therefore, in the model visual search emerges naturally from the dynamics by top-down biasing the selection decision in the *scene spatial selection* field (*L*) through peaks in the working memory fields. Visual search terminates when there is a peak in the *target position (WM)* field (*O*2), or if there are no more salient locations left for selection.

Neural activation of working memory The model's visual search cue is evoked through an external task cue in the form of simulated language interaction, e.g., "Look for a red pepper in the kitchen". Cues are realized through the activation of concept nodes (*S*) that provide input to the *label cue (WM)* (*S*3), *color cue (WM)* (*S*6), *scene type (WM)* (*S*9) fields. Scene type information is part of the provided language cue, since recognition through scene gist is out of the scope of our model. We focus on how different active concepts interact with visual search and not how these get activated by language or gist. The selective *label/scene/anchor)* field (*P*8) is a simplified (hand-crafted) account for the long time memory (LTM) representation of known *label/scene/anchor* combinations obtained through experience. If an *anchor object* is known for the current scene and label it causes a peak to emerge in the *anchor label (WM)* field (*P*5), which then biases attentional selection towards the *anchor*. Inhibitory couplings between fields in the *search cue* sub-network (*P*) allow for the autonomous transition from the highly efficient scene guidance search to the efficient feature guidance search, when the anchor object was not found in the current scene. The selective *label/anchor/relation)* field (*P*6) is also a simplified (hand-crafted) account for the LTM representation of known *label/anchor/relation* combinations obtained through experience. It gives input to the *anchor relation (WM)* field (*P*2) and a peak in this field emerges, when there is a known relation for the current label and anchor.

Feature matching In DFT matches between *expected* (*P*3, *P*4) and *attended* (*J*6, *J*7) features are detected through the detection instability. Specifically, a *match detection* field (*R*4) receives localized input from these fields such that only if these inputs overlap sufficiently is goes through a detection instability. These matches are detected in parallel along each feature dimension. If all expected features match their attended counterpart the *condition of satisfaction* node (*R*5) (Sandamirskaya & Schöner, 2010) gets activated. It signals that a matching object was found, and the current attended location is committed to the *matching object memory* (*O*). Depending on the activation in the *anchor label* (WM) (*P*5) field the current position is stored in the *anchor position* (*O*3) or in the *target position* (*O*2) (WM) field. A supra-threshold peak in either of the *matching object memory* (*O*) position fields inhibits the *intention* node (*R*3) of the *feature matching* sub-network (*R*). This allows for further cognitive operations on the currently attended location, by effectively preventing the attentional selection of a new location. The content of the selective *expected color* (*P*3) and *expected label* (*P*4) fields are modulated by top-down task inference that is the result of their coupling to other fields in the *search cue* sub-network (*P*).

Memory guidance Visual search is additionally influenced by already present scene memory (Hollingworth (2006); Hollingworth (2012); Draschkow et al. (2014); Grieben et al. (2020)). In the model this bias is implemented through the *memory guidance spatial map* (*N*3), which gives previously attended locations with matching features a selective advantage (see Grieben et al. (2020) for experimental results and an in-depth analysis of this bias from memory).

Feature guidance Feature guidance is the core working memory bias in visual search (Wolfe and Horowitz (2017)). In the model this bias comes from the *feature guidance spatial map* (*F*3), which gives locations with matching features a non-linear selective advantage. The non-linear bias that emerges naturally from the underling dynamics has been experimentally verified (for conjunction visual search and its interaction with scene memory see (Grieben et al., 2020) and against triple conjunction (Nordfang & Wolfe, 2014) visual search see (Grieben & Schöner, 2021)). In the current model we specifically addressed the question how guidance templates for object categories in natural scenes could be learned from the visual input. Therefore, the *feature guidance template* (*Q*) comes from an adaptive LTM representation that is updated through experience. The guiding features used in this model are restricted to *preattentive shape* and *color* since these are known guiding features for objects in natural scenes (Wolfe, 2021). The learned color guidance feature can be replaced by other top-down inference processes to enable flexible adaptation to specific search tasks.

Scene guidance/grammar In this model we now include a new neural process that utilizes anchor objects and their specific spatial relation regarding other objects in visual search. To this end the architecture utilizes a coordinate transformation of activation patterns that represent operators in relational spatial language Richter, Lins, and Schöner (2021) to provide an attentional bias relative to a found anchor object. The appropriate spatial pattern emerges in the *relation/space* field (*K*3) through an incoming overlap between the current *anchor relation (WM)* (*P*2) and the preshaped possible relation patterns. Due to the peak representation in the *anchor position (WM)* field (*O*3) it is possible to perform a coordinate transformation (*K*2; Schneegans and Schöner (2012)) that shifts the spatial pattern peak in the *scene guidance spatial map* field (*K*4) accordingly and provides it as bias input for attentional selection to the *scene spatial selection* field (*L*). A supra-threshold peak in the *scene guidance spatial map* field (*K*4) additionally inhibits the *anchor label (WM)* (*P*5) and the *anchor position (WM)* (*O*3) fields and thus stops the model from searching for the anchor.

Learning object and guidance templates

As long as the model is attending a location, a supervisor can provide an appropriate label (*S*3) (by activating a concept node), which is then associated with the features observed at the currently attended location. Learning is induced through the activation of a learn boost (*A*1) that gives rise to a transient activation pattern (*A*2,*A*3) (Kazerounian, Luciw, Richter, & Sandamirskaya, 2013). In the model two dynamic learn processes depend on the transient activation pattern. First, learning the association between the complex foveal shape feature at the attended location and the given object label (class) (*object template*). Second, learning the association between the given label and the retinal *preattentive* features (*guidance template*). A single transient activation is sufficient to learn a new object class. While learning takes place the *intention* node (*R*3) of the *feature matching sub-network* (*R*) is inhibited, effectively preventing the selection of a new location in the *scene spatial selection* field (*L*) during learning.

Learning of an object template A headless VGG16 network (*J*2) provides the complex feature maps m_f (f is the feature index), that are needed for object recognition, from the foveal image (*E*). The object template consists of connection weights, $w_{m_\mathrm{f},u_\mathrm{fsml}}$, that perform the transformation from the distributed representation in the feature maps m_f to the localist representation in the *foveal space/label map* field (u_fsml, *J*5). These connections weights are updated according to a dynamic version of the BCM (Law and Cooper (1994); Udeigwe, Munro, and Ermentrout (2017)) rule:

$$\begin{aligned} \tau_w \dot{w}_{m_\mathrm{f},u_\mathrm{fsml}}(\boldsymbol{x},t) &= \eta\ \sigma(u_\mathrm{learn})\ y\ (y-\Theta)\ \frac{m_\mathrm{f}(x_1,x_2,t)}{\Theta} \\ y &= \sigma(u_\mathrm{fsml}(\boldsymbol{x},t)) \\ \tau_\Theta \dot{\Theta} &= (y^2-\Theta), \end{aligned} \tag{2}$$

with η being the learning rate and u_learn the activation of the transient *learn* node (*A*2). Before learning the *label cue (WM)* field (*S*3) provides slice input to the *foveal space/label map* field (*J*5) and the *learn* node (*A*2) down-regulates the resting level of the field by supplying a homogeneous inhibitory input to it. After learning the input s_{u_fsml} to the *foveal space/label map* field is:

$$s_{u_\mathrm{fsml}}(\boldsymbol{x},t) = \sum_{f=0}^{F-1} m_\mathrm{f}(x_1,x_2,t)\ w_{m_\mathrm{f},u_\mathrm{fsml}}(\boldsymbol{x},t). \tag{3}$$

Learning of the guidance templates The synaptic weight pattern w_lcgm for the *label/color guidance map* field (u_lcgm, *Q*6) is updated according to a dynamic version of the Hebbian learning rule (Tekülve, Fois, Sandamirskaya, & Schöner, 2019):

$$\tau \dot{w}_\mathrm{lcgm}(\boldsymbol{x},t) = -\eta\ \sigma(u_\mathrm{learn})\ (\sigma(u_\mathrm{lcgm}(\boldsymbol{x},t)) - w_\mathrm{lcgm}(\boldsymbol{x},t)) \tag{4}$$

with η being the learning rate and u_learn the activation of the transient *learn* node (*A*2). The weight pattern w_lcgm for the *label/shape guidance map* field (u_lsgm, *Q*3) is updated using an analog rule. Activation of the transient *learn* node (*A*2) enables the formation of peaks in the *guidance map* (*Q*2, *Q*5) representing the currently attended preattentive (retinal) features (*I*3, *I*4), which are required for association. After learning the synaptic weight pattern w_lcgm and w_lsgm serve as preshape input to their corresponding *guidance map* fields.

Results

The neural architecture inherits all the properties of our previous models (Grieben et al. (2020); Grieben and Schöner (2021)) retaining the ability to fit and explain the relevant experimental findings. We used six class images to train our model and three scenes to test it shown in Fig. 2.

Figure 2: One-shot training images for six different classes (bottom) and the three test scenes (top). Bathroom image adapted from Võ (2021) (Fig. 5).

Exploration and scene memory

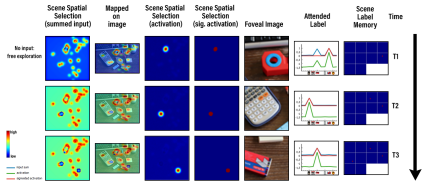

Figure 3: Demonstration of the exploration behavior of the model.

In Fig. 3 we see how the model autonomously selects salient locations in the visual scene and stores the labels (and colors) at the attended location in the scene memory fields.

Visual search

In Fig. 4 we see the two possible cases of guided visual search that emerge from the model: 1) for a label (top) and 2) for a label-color combination (bottom). In the first case, where only a label *pepper* is supplied, the guidance features of the learned representation emerge in working memory. Since the model was trained on a yellow *pepper*, the *color guidance feature* field contains a peak at yellow. But since this feature is only used for guiding, the model will stop at the first *pepper* it attentionally selects, since the *expected color* field contains no peak. In the second case, where the color cue *red* was supplied with the label *pepper* both the *expected color* and the *color guidance feature* fields contain a peak representing red. In the summed input to the *scene spatial selection* field one can see that this small change in working memory has a noticeable effect. Activation in the *shape guidance feature* field unfolds the same in both search cases.

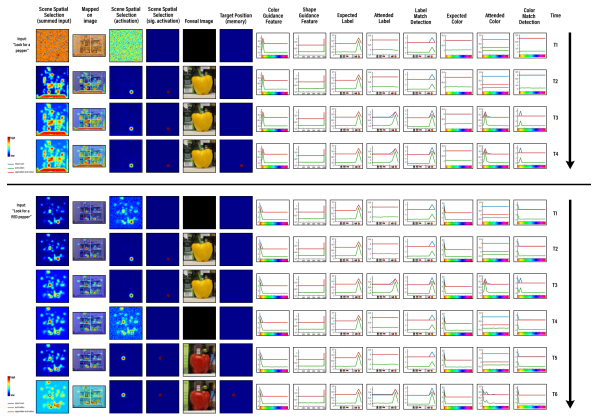

Figure 4: Demonstration of the model performing two visual search tasks, with (bottom) or without (top) an extra color cue.

Scene guidance/grammar

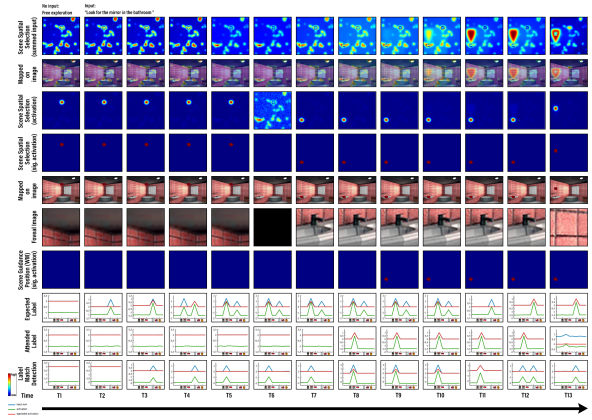

Figure 5: Demonstration of the model performing a scene guidance/grammar task (looking for the invisible mirror). Task image adapted from Võ (2021) (Fig. 5).

In Fig. 5 we see the model performing a task presented in Võ (2021) to test the ability of the scene guidance/grammar. It receives the label *mirror* and the scene type *bathroom*. We see that at first, in timesteps T2-T3, the expected label is *mirror*. Then in timesteps T4-T5 a strong input for the label *sink*, that originates from the *anchor label field* (not depicted), creates a peak in the same field and the peak for *mirror* decays due to the field's global inhibition. This small change in feature guidance is reflected in the summed input to the *scene spatial selection* field. The sink in the scene has now the highest salience (T5-T6) and is therefore selected in the next cycle of attentional selection (T7). The *label match detection* detects (T8), that it is attending the expected label, and the current position is stored through gating into the *position (WM)* of the *scene guidance* sub-network (T9). This destabilizes the working memory representation of the *anchor label (WM)*, and a peak at the *mirror* label forms again (T11-T12). Starting at timestep T10 we see the relational bias *above*, evolve over time, centered on the attended location of the *sink*. The strength of this relational bias alone is enough to force a selection decision for this specific region. This is consistent with the looking heatmap shown in Võ (2021) (Fig. 5).

Discussion

We have extended our neural process models of visual search and scene memory (Grieben et al. (2020); Grieben and Schöner (2021)) to enable autonomously building a scene representation and performing guided visual search on natural scenes. Along the way, we found solutions for three important open problems. First, little is known on how humans guide visual search for objects in natural scenes. The established guiding features (Wolfe & Horowitz, 2017) found through controlled experimental setups and oversimplified stimuli undoubtedly *can* guide visual search, but at the same time visual search for natural objects seems to be essentially unguided (Vickery et al., 2005), when presented outside of a meaningful scene. There is evidence, that *color* (Wolfe, Alvarez, Rosenholtz, Kuzmova, & Sherman, 2011) and *preattentive shape* (Wolfe, 2021) are some of the few features that also guide visual search for natural objects. But what this *preattentive shape* really is remains an unsolved question. In the paper we presented a solution on how the association between a object concept and *preattentive shape* can be learned from the intermediate layer of a CNN (as proposed by Wolfe (2021)). Second, humans can use the structure of the scene, *scene grammar*, to guide search for objects in natural scenes highly efficiently (Boettcher et al. (2018); Võ (2021)). Here we presented a new neural process on how *scene grammar* can emerge from the underlying dynamics of the model. To our knowledge this is the first model to account for it. Third, searching for objects implies learning *object templates* for object recognition. For this we embedded a headless CNN as a feature extracting network into our DFT model, for the first time. And presented a biologically inspired mapping from the distributed representation of the CNN feature maps to the localist representation of neural fields. The biggest strength of our model compared to an end-to-end learned CNN is besides the obvious neural plausibility, that the model operates in a closed behavioral loop. Stable memory representations allow for goal-oriented actions and the adaptive recurrent top-down feedback allows top-down inference processes to flexible switch between modes, without the need for specific algorithms. Given the nature of the DFT framework in which models are built from conceptually constrained building blocks, the present model may serve as the perceptual front-end for any DFT model that processes visual input. Our model could thus be combined with any previous DFT models that worked on simplified stimuli, extending these to natural images. Ultimately, this brings us a step closer to the conception of an autonomous agent that achieves higher cognition in natural environments. Future work must validate the model against human behavioral data.

References

Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, *2*(1), 32–48.

Boettcher, S. E., Draschkow, D., Dienhart, E., & Võ, M. L.-H. (2018). Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of vision*, *18*(13), 11–11.

Castelhano, M. S., & Krzyś, K. (2020). Rethinking space: A review of perception, attention, and memory in scene processing. *Annual Review of Vision Science*, *6*, 563–586.

Cunningham, C. A., & Wolfe, J. M. (2014). The role of object categories in hybrid visual and memory search. *Journal of Experimental Psychology: General*, *143*(4), 1585.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, *18*(1), 193–222.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Draschkow, D., Wolfe, J. M., & Võ, M. L.-H. (2014). Seek and you shall remember: Scene semantics interact with visual search to build better memories. *Journal of Vision*, *14*(8), 10–10.

Goodale, M. A., & Milner, A. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20-25. doi: https://doi.org/10.1016/0166-2236(92)90344-8

Grieben, R., & Schöner, G. (2021). A neural dynamic process model of combined bottom-up and top-down guidance in triple conjunction visual search. In T. Fitch, C. Lamm, H. Leder, & K. Teßmar-Raible (Eds.), *Proceedings of the 43nd annual conference of the cognitive science society*. Cognitive Science Society.

Grieben, R., Tekülve, J., Zibner, S. K., Lins, J., Schneegans, S., & Schöner, G. (2020). Scene memory and spatial inhibition in visual search: A neural dynamic process model and new experimental evidence. *Attention, Perception, & Psychophysics*. doi: 10.3758/s13414-019-01898-y

Gupta, S. K., Zhang, M., Wu, C.-C., Wolfe, J. M., & Kreiman, G. (2021). Visual search asymmetry: Deep nets and humans share similar inherent biases. *arXiv preprint arXiv:2106.02953*.

Hollingworth, A. (2006). Visual memory for natural scenes: Evidence from change detection and visual search. *Visual cognition*, *14*(4-8), 781–807.

Hollingworth, A. (2012). Guidance of visual search by memory and knowledge. In *The influence of attention, learning, and motivation on visual search* (pp. 63–89). Springer.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, *40*(10-12), 1489–1506.

Kazerounian, S., Luciw, M., Richter, M., & Sandamirskaya, Y. (2013). Autonomous reinforcement of behavioral sequences in neural dynamics. In *International joint conference on neural networks (ijcnn)*.

Knudsen, E. I. (2007). Fundamental components of attention. *Annu. Rev. Neurosci.*, *30*, 57–78.

Law, C. C., & Cooper, L. N. (1994). Formation of receptive fields in realistic visual environments according to the bienenstock, cooper, and munro (bcm) theory. *Proceedings of the National Academy of Sciences*, *91*(16), 7797–7801.

Lim, S., McKee, J. L., Woloszyn, L., Amit, Y., Freedman, D. J., Sheinberg, D. L., & Brunel, N. (2015). Inferring learning rules from distributions of firing rates in cortical neurons. *Nature neuroscience*, *18*(12), 1804–1810.

Lomp, O., Richter, M., Zibner, S. K. U., & Schöner, G. (2016). Developing Dynamic Field Theory Architectures for Embodied Cognitive Systems with cedar. *Frontiers in Neurorobotics*, *10*, 14.

Mai, Z., Li, R., Jeong, J., Quispe, D., Kim, H., & Sanner, S. (2022). Online continual learning in image classification: An empirical survey. *Neurocomputing*, *469*, 28-51. doi: https://doi.org/10.1016/j.neucom.2021.10.021

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, *452*, 48-62. doi: https://doi.org/10.1016/j.neucom.2021.03.091

Nordfang, M., & Wolfe, J. M. (2014). Guided search for triple conjunctions. *Attention, Perception, & Psychophysics*, *76*(6), 1535–1559.

Richter, M., Lins, J., & Schöner, G. (2021). A neural dynamic model of the perceptual grounding of spatial and movement relations. *Cognitive Science*, *45*(10), e13045. doi: https://doi.org/10.1111/cogs.13045

Sandamirskaya, Y., & Schöner, G. (2010). An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, *23*(10), 1164–1179.

Schneegans, S., & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological cybernetics*, *106*(2), 89–109.

Schöner, G., Spencer, J. P., & DFT Research Group, T. (2016). *Dynamic thinking: A primer on dynamic field theory*. Oxford University Press.

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

Tekülve, J., Fois, A., Sandamirskaya, Y., & Schöner, G. (2019). Autonomous sequence generation for a neural dynamic robot: Scene perception, serial order, and object-oriented movement. *Frontiers in Neurorobotics*, *13*, 95. doi: 10.3389/fnbot.2019.00095

Udeigwe, L. C., Munro, P. W., & Ermentrout, G. B. (2017). Emergent dynamical properties of the bcm learning rule. *The Journal of Mathematical Neuroscience*, *7*(1), 1–32.

Vickery, T. J., King, L.-W., & Jiang, Y. (2005, 02). Setting up the target template in visual search. *Journal of Vision*, *5*(1), 8-8. Retrieved from `https://doi.org/10.1167/5.1.8` doi: 10.1167/5.1.8

Võ, M. L.-H. (2021). The meaning and structure of scenes.

Vision Research, *181*, 10–20.

Võ, M. L.-H., & Wolfe, J. M. (2015). The role of memory for visual search in scenes. *Annals of the New York Academy of Sciences*, *1339*(1), 72.

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, *1*(2), 202–238.

Wolfe, J. M. (2021). Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, 1–33.

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., & Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception, & Psychophysics*, *73*(6), 1650.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*(3), 0058.

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in cognitive sciences*, *15*(2), 77–84.

Zelinsky, G. J., Chen, Y., Ahn, S., Adeli, H., Yang, Z., Huang, L., … Hoai, M. (2021). Predicting goal-directed attention control using inverse-reinforcement learning. *Neurons, behavior, data analysis and theory*, *2021*.

Zhang, M., Feng, J., Ma, K. T., Lim, J. H., Zhao, Q., & Kreiman, G. (2018). Finding any waldo with zero-shot invariant and efficient visual search. *Nature communications*, *9*(1), 1–15.