

# On the Binding Problem in Artificial Neural Networks

Klaus Greff\*

*Google Research, Brain Team*

*Tucholskystraße 2, 10116 Berlin, Germany*

KLAUSG@GOOGLE.COM

Sjoerd van Steenkiste

SJOERD@IDSIA.CH

Jürgen Schmidhuber

JUERGEN@IDSIA.CH

*Istituto Dalle Molle di studi sull'intelligenza artificiale (IDSIA)*

*Università della Svizzera Italiana (USI)*

*Scuola universitaria professionale della Svizzera italiana (SUPSI)*

*Via la Santa 1, 6962 Viganello, Switzerland*

## Abstract

Contemporary neural networks still fall short of human-level generalization, which extends far beyond our direct experiences. In this paper, we argue that the underlying cause for this shortcoming is their inability to dynamically and flexibly bind information that is distributed throughout the network. This *binding problem* affects their capacity to acquire a compositional understanding of the world in terms of symbol-like entities (like objects), which is crucial for generalizing in predictable and systematic ways. To address this issue, we propose a unifying framework that revolves around forming meaningful entities from unstructured sensory inputs (segregation), maintaining this separation of information at a representational level (representation), and using these entities to construct new inferences, predictions, and behaviors (composition). Our analysis draws inspiration from a wealth of research in neuroscience and cognitive psychology, and surveys relevant mechanisms from the machine learning literature, to help identify a combination of inductive biases that allow symbolic information processing to emerge naturally in neural networks. We believe that a compositional approach to AI, in terms of grounded symbol-like representations, is of fundamental importance for realizing human-level generalization, and we hope that this paper may contribute towards that goal as a reference and inspiration.

**Keywords:** binding problem, compositionality, systematicity, objects, artificial neural networks, representation learning, neuro-symbolic AI

---

\* This research was partially conducted while the author was affiliated with IDSIA, USI & SUPSI.

\*\* A preliminary version of this work was presented at an ICML Workshop ([van Steenkiste et al., 2019a](#)).

## 1. Introduction

Existing neural networks fall short of human-level generalization. They require large amounts of data, struggle with transfer to novel tasks, and are fragile under distributional shift. However, under the right conditions, they have shown a remarkable capacity for learning and modeling complex statistical structure in real-world data. One explanation for this discrepancy is that neural networks mostly learn about surface statistics in place of the underlying concepts, which prevents them from generalizing systematically. However, despite considerable effort to address this issue, human-level generalization remains a major open problem.

In this paper, we will view the inability of contemporary neural networks to effectively form, represent, and relate symbol-like entities, as the root cause of this problem. This emphasis on symbolic reasoning reflects a common sentiment within the community and others have advocated similar perspectives (Fodor and Pylyshyn, 1988; Marcus, 2003; Lake et al., 2017). Indeed, it is well established that human perception is structured around objects, which serve as compositional ‘building blocks’ for many aspects of higher-level cognition such as language, planning, and reasoning. This understanding of the world, in terms of parts that can be processed independently and recombined in near-infinite ways, allows humans to generalize far beyond their direct experiences.

Meanwhile, the persistent failure of neural networks to generalize systematically is evidence that neural networks do not acquire the ability to process information symbolically, simply as a byproduct of learning. Specialized inductive biases that mirror aspects of human information processing, such as attention or memory, have led to encouraging results in certain domains. However, the general issue remains unresolved, which has led some to believe that the way forward is to build hybrid systems that combine connectionist methods with inherently symbolic approaches. In contrast, we believe that these problems stem from a deeper underlying cause that is best addressed directly from *within* the framework of connectionism.

In this work, we argue that this underlying cause is the *binding problem*: The inability of existing neural networks to dynamically and flexibly *bind* information that is distributed throughout the network. The binding problem affects their ability to form meaningful entities from unstructured sensory inputs (segregation), to maintain this separation of information at a representational level (representation), and to use these entities to construct new inferences, predictions, and behaviors (composition). Each of these aspects relates to a wealth of research in neuroscience and cognitive psychology, where the binding problem has been extensively studied in the context of the human brain. Based on these connections, we work towards a solution to the binding problem in neural networks and identify several important challenges and requirements. We also survey relevant mechanisms from the machine learning literature that either directly or indirectly already address some of these challenges. Our analysis provides a starting point for identifying the right combination of inductive biases to enable neural networks to process information symbolically and generalize more systematically.

In our view, integrating symbolic processing into neural networks is of fundamental importance for realizing human-level AI, and will require a joint community effort to resolve. The goal of this survey is to support this effort, by organizing various related research into a unifying framework based on the binding problem. We hope that it may serve as an inspiration and reference for future work that bridges related fields and sparks fruitful discussions.

## 2. The Binding Problem

We start our discussion by reviewing the importance of symbols as units of computation and highlight several symptoms that point to the lack of emergent symbolic processing in existing neural networks. We argue that this is a major obstacle for achieving human-level generalization, and posit that the binding problem in connectionism is the underlying cause for this weakness. This section serves as an introduction to the binding problem and provides the necessary context for the subsequent in-depth discussion of its individual aspects in Sections 3 to 5.

### 2.1 Importance of Symbols

The human capacity to comprehend reaches far beyond direct experiences. We are able to reason causally about unfamiliar scenes, understand novel sentences with ease, and use models and analogies to make predictions about entities far outside the scope of everyday reality, like atoms, and galaxies. This seemingly infinite expressiveness and flexibility of human cognition has long fascinated philosophers, psychologists, and AI researchers alike. The best explanation for this remarkable cognitive capacity revolves around symbolic thought: the ability to form, manipulate, and relate mental entities that can be processed like symbols (Whitehead, 1985). By decomposing the world in terms of abstract and reusable ‘building blocks’, humans are able to understand novel contexts in terms of known concepts, and thereby leverage their existing knowledge in near-infinite ways. This compositionality underlies many high-level cognitive abilities such as language, causal reasoning, mathematics, planning, analogical thinking, etc.

Human understanding of the world in terms of objects develops at an early age (Spelke and Kinzler, 2007) and infants as young as five months appear to understand that objects continue to exist in the absence of visual stimuli (object permanence; Baillargeon et al., 1985). Arguably, this decoupling of mental representation from direct perception is a first step towards a compositional description of the world in terms of more abstract entities. By the age of eighteen months, young children have acquired the ability to use gestures symbolically to refer to objects or events (Acredolo and Goodwyn, 1988). This ability to relate sensory entities is then key to the subsequent grounding of language. As the child grows up, entities become increasingly more general and start to include categories, concepts, events, behaviors, and other abstractions, together with a growing number of universal relations such as “same”, “greater than”, “causes”, etc. This growing set of composable building blocks yields an increasingly more powerful toolkit for constructing structured mental models of the world (Johnson-Laird, 2010).

The underlying compositionality of such symbols is equally potent for AI, and numerous methods that model intelligence as a symbol manipulation process have been explored. Early examples included tree-search over abstract state spaces such as the General Problem Solver (Newell et al., 1959) for theorem proving, or chess (Campbell et al., 2002); Expert systems that made use of decision trees to perform narrow problem solving for hardware design (Sollow et al., 1987) and medical diagnosis (Shortliffe et al., 1975); Natural language parsers that used a dictionary and a fixed set of grammatical rules to interpret written English; And knowledge bases such as semantic networks (networks of concepts and relations) that could be used to answer basic questions (Weizenbaum, 1966), solve basic algebra word

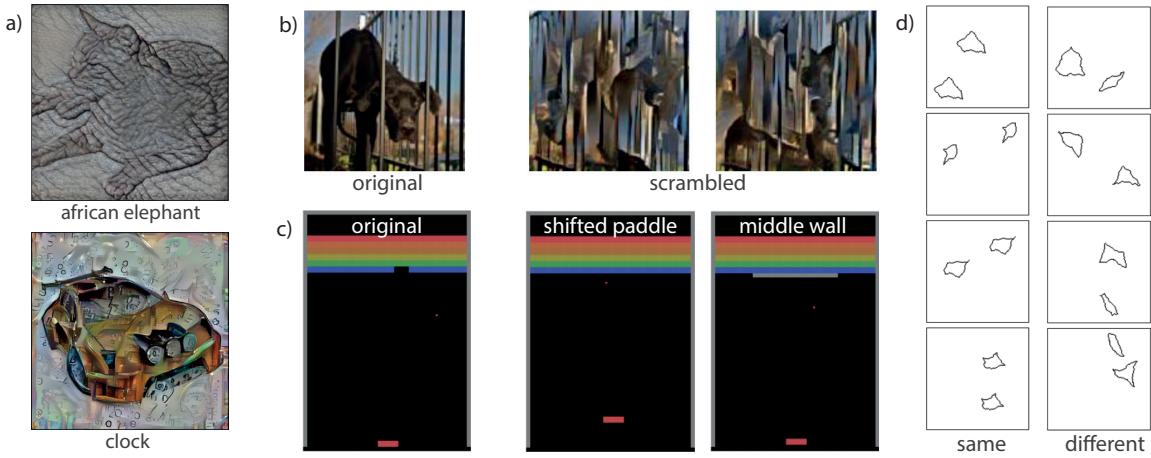


Figure 1: Various evidence for shortcomings of current neural networks. (a) CNN image classifiers are biased towards texture over shape (Geirhos et al., 2019) and (b) can be well approximated by bag-of-local-features models (Brendel and Bethge, 2019). Hence, scrambling the image in a way that preserves local (but not global) structures affects them less than humans. (c) Neural network based agents trained on Breakout, fail to generalize to slight variations of the game such as a shifted paddle or an added middle wall (Kansky et al., 2017). (d) Neural networks also struggle to learn visual relations such as whether two shapes are the same or different (Fleuret et al., 2011; Kim et al., 2018).

problems (Bobrow, 1964), or control simple virtual block worlds (Winograd, 1971). All of these examples of *symbolic AI* relied on manually designed symbols and rules of manipulation, which allowed them to generalize in *predictable* and *systematic* ways. Since then, many of these approaches have become part of the standard computer-science toolbox<sup>1</sup>.

## 2.2 Symbolic processing in Connectionist Methods

Connectionism takes a different, brain-inspired, approach to Artificial Intelligence that stands in contrast to symbolic AI and its focus on the conscious mind (Newell and Simon, 1981; Fodor, 1975). Rather than relying on hand-crafted symbols and rules, connectionist approaches such as neural networks focus on *learning* suitable distributed representations directly from low-level sensory data. In this way, neural networks have resolved many of the problems that haunted symbolic AI, including their brittleness when confronted with inconsistencies or noise, and the prohibitive amount of human engineering and interpretation that would be required to apply these techniques on low-level perceptual tasks. Importantly, the distributed representations learned by neural networks are directly grounded in their input data, unlike symbols whose connection to real-world concepts is entirely subject to human interpretation (see *symbol grounding problem*; Harnad, 1990). Modern neural networks have proven highly successful and superior to symbolic approaches in perceptual domains, such as in visual object recognition (Cireşan et al., 2011, 2012; Krizhevsky et al., 2012) or

1. They are hardly called AI anymore since it is now well understood how to solve the problems that they address. This redefinition of what constitutes AI is sometimes called the *AI effect*, summarized by Douglas Hofstadter as “AI is whatever hasn’t been done yet”.

speech recognition (Fernández et al., 2007; Hinton et al., 2012), and even in some inherently symbolic domains such as language modeling (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020), translation (Wu et al., 2016), board games (Silver et al., 2017), and symbolic integration (Lample and Charton, 2020).

On the other hand, it has become increasingly evident that neural networks fall short in many aspects of human-level generalization, including those that symbolic approaches exhibit by design. For example, it is difficult for neural language models to generalize syntactic rules such as verb tenses or embedded clauses in a systematic manner (Keysers et al., 2020; Lake and Baroni, 2018; Loula et al., 2018; Hupkes et al., 2020). Similarly, in vision, neural approaches often learn overly specialized features that do not easily transfer to different datasets or held-out combinations of attributes (Yosinski et al., 2014; Atzmon et al., 2016; Santoro et al., 2018b). In reinforcement learning, where the use of neural networks has led to superhuman performance in gameplay (Mnih et al., 2015; Silver et al., 2017; Berner et al., 2019), it is found that agents are fragile under distributional shift (Kansky et al., 2017; Zhang et al., 2018; Gamian and Goldberg, 2019) and require substantially more training data than humans (Tsividis et al., 2017). These failures at systematically reusing knowledge suggest that neural networks do not learn a compositional knowledge representation (although some mitigation is possible (Hill et al., 2019, 2020)). In some cases, such as in vision, it may appear that object-level abstractions can emerge naturally as a byproduct of learning (Zhou et al., 2015). However, it has repeatedly been shown that such features are best understood as “a texture detector highly correlated with an object” (Olah et al., 2020; Sundararajan et al., 2017; Ancona et al., 2017; Brendel and Bethge, 2019; Geirhos et al., 2019). In general, evidence indicates that neural networks learn mostly about surface statistics (e.g. between textures and classifications in images) in place of the underlying concepts (Jo and Bengio, 2017; Karpathy et al., 2015; Lake and Baroni, 2018).

A hybrid approach that combines the seemingly complementary strengths of neural networks and symbolic approaches may help address these issues, and several variations have been explored (Bader and Hitzler, 2005). A common variant uses a neural network as a perceptual interface (or pre-processor) tasked with learning symbols from raw data, which then serve as input to a symbolic reasoning system (e.g. Mao and Gan, 2019). Similarly, bottom-up neural networks have been used to make inference more tractable in probabilistic generative models that contain the desired symbolic structure (e.g. in the form of a symbolic graphics renderer Kulkarni et al., 2015). Neural networks have also been combined with search-based methods to improve their efficiency (Silver et al., 2016). Countless other variations that vary in terms of the division of labor between the symbolic and neural components and the choice of a mechanism used to couple them are possible (McGarry et al., 1999; Davidson and Lake, 2020).

In this work, we will adopt a more unified approach that addresses these problems from within the framework of connectionism. It is concerned with incorporating inductive biases in neural networks that enable them to efficiently learn about symbols and the processes for manipulating them (examples of such an approach are abound, even in early connectionist research, e.g. Smolensky (1990); Pollack (1990); McMillan et al. (1992); Das and Mozer (1993)). Compared to a hybrid approach, we believe that this is advantageous for a number

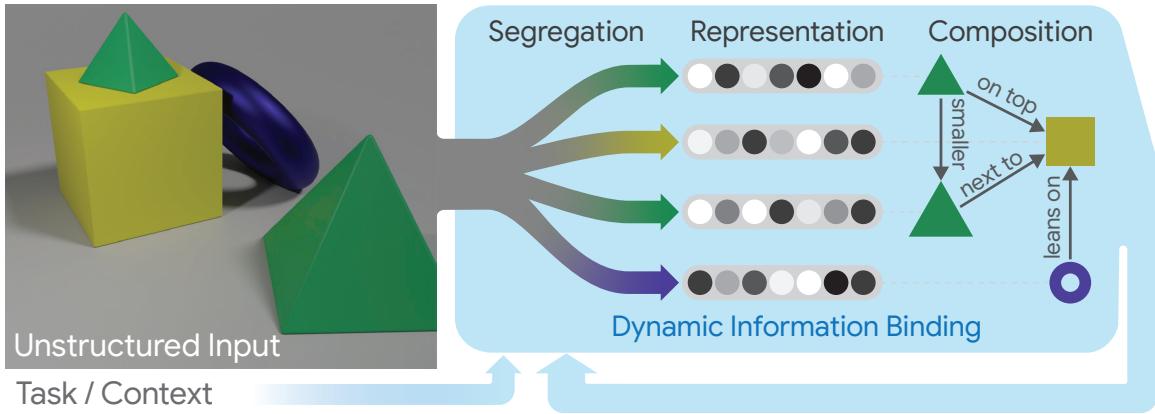


Figure 2: The binding problem in artificial neural networks can be understood from the perspectives of *segregation*, *representation*, and *composition*. Each of these subproblems focuses on a different functional aspect of dynamically binding neurally processed information with the aim of facilitating more symbolic information processing.

of reasons. Firstly, it reduces the required amount of task-specific engineering<sup>2</sup> and helps generalize to domains where expert knowledge is not available. Secondly, by tightly integrating multiple different layers of abstraction, they can continuously co-adapt, which avoids the need for rigid interfaces between connectionist and explicitly symbolic components. Finally, as is evident from the brain, it is sufficient to simply *behave* as an emergent symbol manipulator, and therefore explicit symbolic structure is not a requirement. The main challenge regarding this approach to AI is then to identify corresponding inductive biases that enable symbolic behavior to emerge.

### 2.3 The Binding Problem in Connectionist Methods

We claim that there exists an underlying cause for the lack of emergent symbolic processing in neural networks, which we refer to as the binding problem. The binding problem is about the inability to dynamically and flexibly combine (bind) information that is distributed throughout the network, which is required to effectively form, represent, and relate symbol-like entities. In regular neural networks, information routing is largely determined by the architecture and weights, both of which are fixed at training time. This limits their ability to dynamically route information based on a particular context and thereby accommodate different patterns of generalization.

The binding problem originates from neuroscience, where it is about the explanatory gap in our understanding of information processing in the brain. It includes perceptual binding problems such as visual binding (color, shape, texture), auditory binding (a voice from a crowd), binding across time (motion), cross-modal binding (sound and vision into joint event), motor-behavior (an action), and sensorimotor binding (hand-eye coordination) (Treisman, 1996; Roskies, 1999; Feldman, 2013). Another class—sometimes referred to as cognitive

2. This leaves the question of the innateness of aspects like causality or three-dimensional space open. Such priors might be helpful or eventually even necessary, however, an intelligent system must also be capable of independently discovering and using novel concepts and structures.

binding problems—includes binding semantic knowledge to a percept, memory reconstruction, and variable binding in language and reasoning<sup>3</sup>.

In the case of neural networks, the binding problem is not just a gap in understanding but rather characterizes a limitation of existing neural networks. Hence, it poses a concrete implementation challenge to address the need for binding neurally processed information, which we believe is common to all of the above subproblems. On the other hand, although we are convinced that this problem can be addressed by incorporating a general dynamic information binding mechanism, it is less clear how this can be implemented. Indeed, the search for an adequate mechanism for binding (in one form or another) is a long-standing problem, not just in neuroscience and cognitive psychology, but also in machine learning (Smolensky, 1987, 1988; Sun, 1992). Rather than focusing on a particular subproblem, here we propose to tackle the binding problem in its full generality, which touches upon all these related areas of research. In this way, we can connect ideas from otherwise disjoint areas, and thus draw upon a large body of research towards developing a general binding mechanism. Inspired by Treisman (1999), we organize our analysis along a functional division into three aspects pertaining to the role of binding for symbolic information processing in neural networks: 1) *representation*, 2) *segregation*, and 3) *composition*, each of which takes a different perspective on the binding problem.

THE REPRESENTATION PROBLEM is concerned with binding together information at a representational level that belongs to separate symbol-like entities. It revolves around so-called *object representations*, which act as basic building blocks for neural processing to behave symbolically. Like symbols, they are self-contained and separate from one another such that they can be related and assembled into structures without losing their integrity. But unlike symbols, they retain the expressive distributed feature-based internal structure of connectionist representations, which are known to facilitate generalization (Hinton, 1984; Bengio et al., 2013). Hence, object representations encode relevant information in a way that combines the richness of neural representations with the compositionality of symbols. We chose the term “object” representation because it is evocative of physical objects, which are processed as symbols in many important cognitive tasks. However, we emphasize that object representations are also meant to encode non-visual entities such as spoken words, imagined or remembered entities, and even more abstract entities such as categories, concepts, behaviors, and goals<sup>4</sup>.

Interestingly, even the seemingly basic task of incorporating object representations in neural networks faces several problems, such as the “superposition catastrophe” (von der Malsburg, 1986) portrayed in Figure 3. It suggests that fully-connected neural networks suffer from an “inherent tradeoff between distributed representations and systematic bindings among units of knowledge” (Hummel and Holyoak, 1993). A general treatment of object representation in neural networks involves addressing the superposition catastrophe, along with several other challenges, which we discuss in Section 3.

- 
- 3. The term binding problem has also been used in the context of consciousness, as the problem of how a single unitary experience arises from the distributed sensory impressions and processing in the brain (Singer, 2001)
  - 4. We have considered several other terms for “object” representations, including entity, gestalt, icon, and concept, which perhaps better reflect their abstract nature but are also less accessible at an intuitive level. The fact that objects are more established in the relevant literature gave them the final edge.

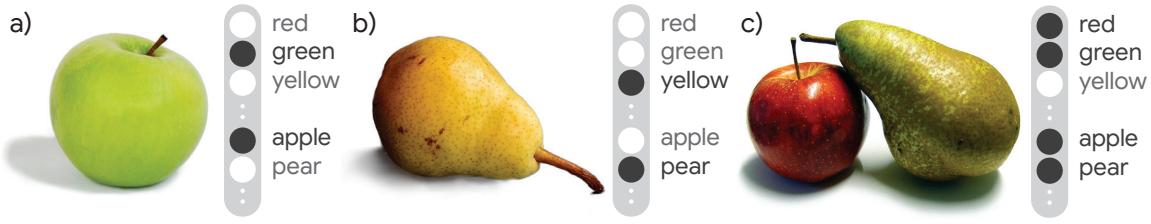


Figure 3: Illustration of the *superposition catastrophe*: A distributed representation in terms of disentangled features like color and shape (**a**, **b**) leads to ambiguity when confronted with multiple objects (**c**): The representation in (**c**) could equally stand for a red apple and a green pear, or a green apple and a red pear. It leads to an indiscriminate bag of features because there is no association of features to objects. A simple form of this problem in neural networks was first pointed out in Rosenblatt (1961), and has been debated in the context of neuroscience since (Milner, 1974; von der Malsburg, 1981).

THE SEGREGATION PROBLEM is about the process of structuring raw sensory information into meaningful entities. It is concerned with the information binding required for dynamically *creating* object representations, as well as the characteristics of objects as modular building blocks for guiding this process. This notion of an object is context and task-dependent, and difficult to formalize even for concrete objects like a tree, a hole, or a river, which are self-evident to humans. Hence, the segregation problem relates to the problem of instance segmentation in that it also produces a division of the input into meaningful parts, but it is complicated by the fact that it is concerned with objects in their most general form. The incredible variability among objects makes it intractable to resolve the segregation problem purely through supervision. Consequently, the segregation problem (Section 4) is about enabling neural networks to acquire an appropriate, context-dependent, notion of objects in a mostly unsupervised fashion.

THE COMPOSITION PROBLEM is about using object representations to dynamically construct compositional models for inference, prediction, and behavior. These structured models leverage the modularity of objects to support different patterns of generalization, and are the means by which more systematic ‘human-like’ generalization can be accomplished. However, this relies on the ability to learn abstract relations that can be arbitrarily and recursively applied to object representations, and requires a form of binding, not unlike the way variables can be bound to placeholder symbols in a mathematical expression. Moreover, the desired structure is often not known in advance and has to be inferred or adapted to a given context or task. To address the composition problem (Section 5), a neural network thus requires a mechanism that provides the flexibility to quickly restructure its information flow and ultimately enable it to generalize systematically.

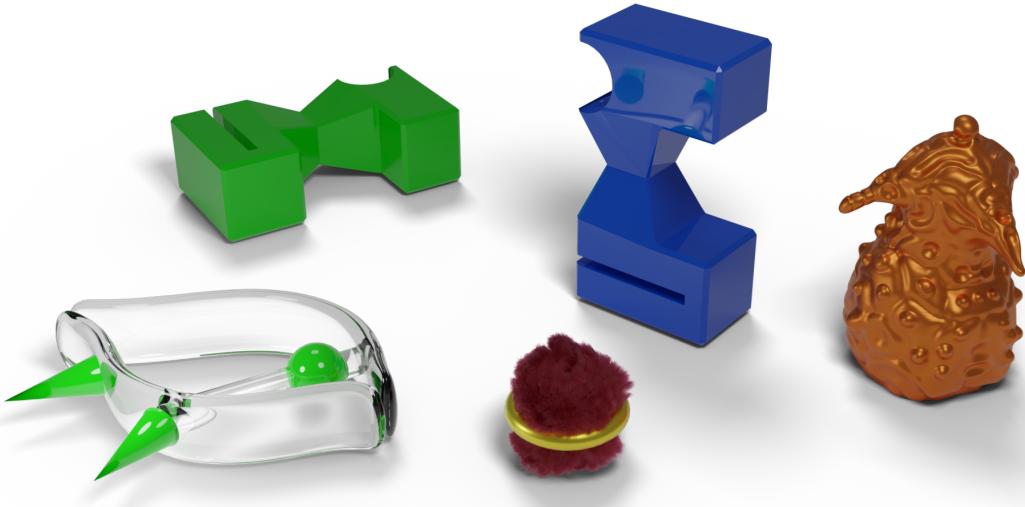


Figure 4: A visual scene composed of various unfamiliar objects.

### 3. Representation

In this section, we look at the binding problem from the perspective of representation. We have argued that, to take advantage of symbolic processing, neural networks require some form of object representations that combine the richness of neural representations with the compositionality of symbols. These object representations are intended as modular ‘building blocks’ from which to efficiently compose structured models of the world. This has direct consequences for the representational format and its underlying dynamics.

Consider for example Figure 4, where you are able to distinguish between five different objects. You can readily describe each object in terms of its shape, color, material, and other properties, despite most likely never having encountered them before. Notice also how these properties relate to individual objects as opposed to the entire scene, which is also evident from the fact that you can tell that the color green occurs multiple times for different objects. Finally, notice how you are readily able to perform comparisons, for example, to tell that the shape of the blue object is the same as that of the green one in the back, but that they differ in color.

In the following, we take a closer look at the *format* of object representations (Section 3.1). We work towards a format that separates information about objects and is general enough to accommodate unfamiliar objects in a meaningful way so that they can readily be compared. Additionally, we will also consider the representational *dynamics* that are required to support stable and coherent object representations over time (Section 3.2). Towards the end, we survey relevant approaches from the literature that may help incorporate these aspects of object representations into neural networks (Section 3.3).

#### 3.1 Representational Format

We seek a representational format that distinguishes objects, while retaining the advantages of learned distributed representations. These representations have proven highly successful

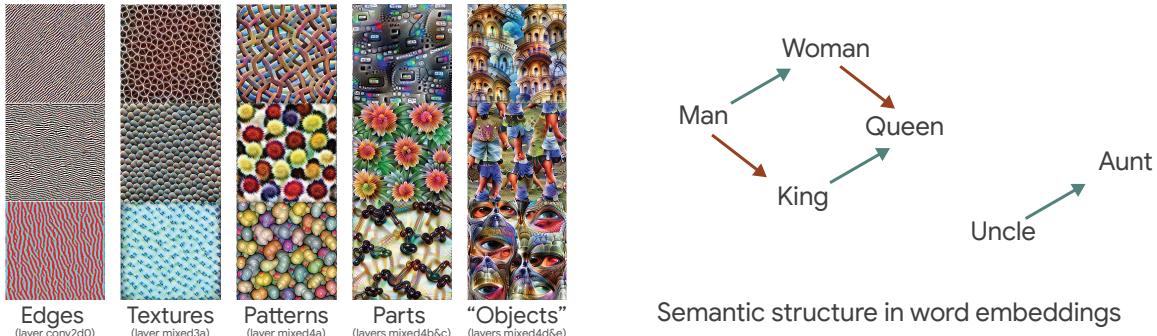


Figure 5: **Left:** Interpretable features learned on ImageNet as observed in Olah et al. (2017). **Right:** Learned word embeddings have been demonstrated to capture some of the semantic structure of text (Mikolov et al., 2013), although to a lesser extent than was initially reported (Nissim et al., 2019).

(e.g. Cireşan et al., 2011; Hinton et al., 2012; Krizhevsky et al., 2012) and are known to partially capture the semantic structure of a task (Figure 5), such as interpretable image features (Zeiler and Fergus, 2014; Olah et al., 2020), or the semantic structure of text (Mikolov et al., 2013; but compare Nissim et al., 2019). In this way learned object representations can also benefit from known inductive biases that focus on feature hierarchies, invariances, and spatio-temporal coherence (Becker and Hinton, 1992), sparsity (Olshausen and Field, 1996), or non-Euclidean feature spaces (Nickel and Kiela, 2017).

### 3.1.1 SEPARATION

To support the construction of structured models, object representations need to act as modular building blocks. This requires information about individual objects to remain separated at a representational level, such that their features do not interfere with one another, even when composed. Additionally, the features that belong to an object must be able to act as a unit, which implies strong dependencies between its features. For example, when an object representation appears or ceases to exist, all of its features are equally affected.

The separation of information has to be flexible enough to ensure that objects can be formed from novel (unseen) feature combinations. Hence, it is important that it is not purely determined by the representational content of the objects, but rather acts as an independent degree of freedom. Regarding capacity, it may suffice to represent only a few objects simultaneously, despite the fact that a typical scene potentially contains a large number of objects. Indeed, the capacity of the human working memory is generally believed to only be around 3–9 objects (Fukuda et al., 2010; Miller, 1956).

### 3.1.2 COMMON FORMAT

To be able to efficiently relate and compare a wide variety of object representations, they must be described in a common format. Recall how in Figure 4 you were able to freely compare a number of unfamiliar objects in terms of their properties, such as their size, shape, and location. On the one hand, this is possible because you have acquired a number of

general relationships, such as “bigger than”, “left of”, etc., which we will discuss in detail in Section 5. What is more important here is that such relations can only be applied if object representations provide a shared interface. More generally, a common format helps to ensure that *any* learned relation, transformation, or skill (like grasping) transfers between similar objects independent of context. Similarly, a common set of features helps carry over experiences between objects during learning.

### 3.1.3 DISENTANGLEMENT

Individual object representations need to be able to describe a large variety of (possibly unseen) objects in terms of attributes that are useful for down-stream problem-solving. This requires focusing on factors of variation in the data, that are sufficiently expressive, but also compact and reusable (i.e. they can be varied independently). Indeed, humans arguably manage to accomplish this by focusing on a relatively small, but consistent set of attributes such as color, shape, etc. (Devereux et al., 2014).

A *disentangled* representation aims to make these attributes explicit by establishing a local correspondence between (independent) factors of variation and features (Barlow et al., 1989; Schmidhuber, 1992c; Higgins et al., 2017a, 2018; Ridgeway and Mozer, 2018). In this case, information about a specific factor can be readily accessed and is robust to unrelated changes in the input, which improves sample efficiency and down-stream generalization (Higgins et al., 2017b; van Steenkiste et al., 2019b). In the context of object representations, disentanglement implies a factorized feature space that captures salient properties of objects. Together with a common format, it facilitates generalization to unseen feature combinations and enables useful comparisons between objects and other meaningful relations to be formed.

## 3.2 Representational Dynamics

When interacting with the real world, the stream of sensory information continuously evolves over time. It is therefore important to consider not only instantaneous representations, but also their *dynamics* over time.

### 3.2.1 TEMPORAL DYNAMICS

An object representation requires ongoing updates across time for a number of reasons: Firstly, with objects constantly moving and transforming in the real world, their corresponding representations need adjustments to remain accurate. Secondly, certain temporal attributes such as movement or behavior can only be estimated when considering the history of information. Finally, with the limited amount of information that can be observed about an object at any given time, accumulating information over multiple partial views can help produce more informative object representations.

An important aspect among all these cases is the need for an object representation to consider not only the input but also its own history (recurrence). This requires a stable identity to help ensure that information across time-steps is associated with the correct object representation. Note that the identity of an object cannot be tied exclusively to its visible properties, as illustrated by the extreme example of a fairytale prince that is transformed into a frog (Marcus, 2003; Bambini et al., 2012).

### 3.2.2 RELIABILITY

Structured mental models depend on object representations to provide a stable foundation for reasoning and other types of information processing (Johnson-Laird, 2010). The reliability of this foundation is especially important for more abstract computations to which object representations provide the only connection to the world. However, perfect reliability is unattainable since sensory information about the world is noisy and incomplete, and the capacity of any model is inherently limited.

Explicitly quantifying uncertainty can help mitigate this issue and prevent noise and errors from accumulating undetectably. In addition, certain small amounts of noise in an object representation may be continually corrected by leveraging dependencies among its features (i.e. through the features of an object acting as a unit). An important source of uncertainty accumulation is due to objects that are temporarily not perceived (e.g. as a result of occlusion). In this case, a ‘self-correcting’ representation may help maintain a stable object representation, even in the absence of sensory input (object permanence).

Uncertainty about object representations may also arise due to ambiguous inputs that allow for several distinct but coherent interpretations (for example see Figure 9 on page 16). The ability to (at least implicitly) encode multi-modal uncertainty is crucial to effectively treat such cases. Top-down feedback may then help disambiguate different interpretations (see also Sections 4.2.2 and 5.2.2).

## 3.3 Methods

In order to fulfill the desiderata outlined above, we require a number of specialized inductive biases. Indeed, it should now also be clear that a simple MLP falls short at adequately representing multiple objects simultaneously: If it attempts to avoid the superposition catastrophe by learning features that are specific to each object, then they lack a common format and become difficult to compare<sup>5</sup>. Therefore, in the following we will review several approaches for representing multiple objects in neural networks. We will focus on common format, temporal dynamics, reliability, and in particular on separation, which thus far has received little attention in the main-stream neural networks literature.

### 3.3.1 SLOTS

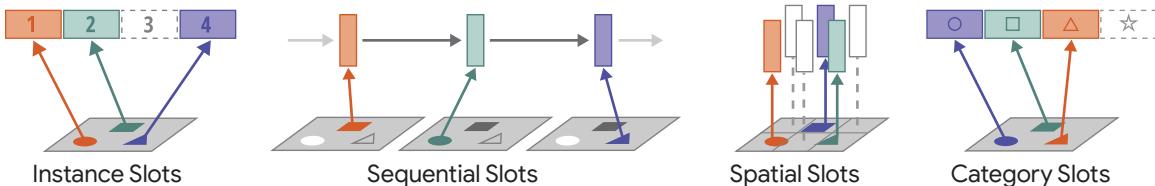


Figure 6: Illustration of the four different types of slot-based representations.

5. Others have suggested ways in which MLPs could *in principle* circumvent this problem (O'Reilly and Busby, 2002; Pollack, 1990). However, neither of these offer a solution that can convincingly fulfill all of the above desiderata simultaneously. In fact, even for plain RNNs it was found that when they are trained to remember multiple objects internally, they resort to a localist representation (Bowers et al., 2014).

The simplest approach to separation is to provide a separate representational slot for each object. This provides a (typically) fixed capacity working memory with independent object representations that can all be accessed simultaneously. Weight sharing can then be used to ensure a common format among the individual slots.

**INSTANCE SLOTS** In the most general form, which we call *instance slots*, all slots share a common format and their information can be kept separate, independent of their representational content. Instance slots are very flexible and general in that they have no preference for content or ordering. However, this generality introduces a *routing problem* when a common format is enforced via weight sharing: with all slots being identical, bottom-up information processing needs to break this symmetry to avoid assigning the same content to each one. Hence, the allocation of information to each slot must be determined by taking the other slots into account, which complicates the process of segregation (see also Section 4.2). Instance slots have been used in several approaches to learning object representations, including Masked Restricted Boltzman Machines (M-RBMs; Le Roux et al., 2011), Neural Expectation-Maximization (N-EM; Greff et al., 2017), and IODINE (Greff et al., 2019). They can also be found in the memory of memory-augmented neural networks (Joulin and Mikolov, 2015; Graves et al., 2016), in self-attention models (Vaswani et al., 2017; Dehghani et al., 2019; Locatello et al., 2020), in Recurrent Independent Mechanisms (RIMs; (Goyal et al., 2019)), albeit without having a common format, and in certain graph neural networks (Battaglia et al., 2018), where they are treated as internal representations that can be accessed simultaneously.

**SEQUENTIAL SLOTS** Sequential slots break slot symmetries by imposing an order on the representational slots, typically across time. They are commonly found in RNNs and, when paired with an attention mechanism that attends to a different object at each step, can serve as object representations. With weights typically being shared across (time)steps, sequential slots naturally share a common format and unlike other slot-based representations can dynamically adjust their representational capacity. Sequential slots in RNNs have been used as object representations, for example in Attend Infer Repeat (AIR; Eslami et al., 2016) and to a lesser degree in DRAW (Gregor et al., 2015). However, due to recurrence, these slots may not always be fully independent, which impedes their function as modular building blocks. Recent approaches, such as Multi-Object Networks (MONet; Burgess et al., 2019) and GENESIS (Engelke et al., 2019), alleviate this by using recurrence only for information routing, but not for the object representations themselves. In general, a potential limitation of sequential slots is that they are not simultaneously accessible at any given (time)step for down-stream processing. This can be addressed via a set function over sequential slots, such as the attention mechanism in certain neural machine translation methods (Bahdanau et al., 2014) or in pointer networks (Vinyals et al., 2015).

**SPATIAL SLOTS** In *spatial slots*, each slot is associated with a particular spatial coordinate (e.g. in an image), which helps to break slot symmetries and simplifies information routing. They can still accommodate a common format through weight-sharing, but lack generality because their content is tied to a specific spatial location. Because location and separation are entangled, changes to the location of an object potentially correspond to a change of slot, which complicates maintaining object identity across time. Spatial slots are commonly found in CNNs, where multiple convolutional layers share filter weights across the spatial dimensions to yield a spatial map of representational slots. Although they are not usually

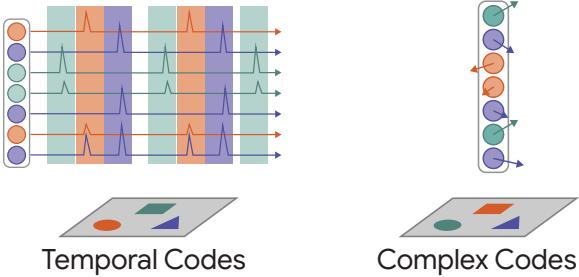


Figure 7: Illustration of the two main augmentation based approaches to object representations. **Left:** Neural activity over time for a temporal code, where synchronization is emphasized using color. **Right:** Complex valued activations are represented by arrows and colored according to their direction.

advertised as object representations in this way, several recent approaches, such as Relation Networks (Santoro et al., 2017), the Multi-Entity VAE (Nash et al., 2017), or the works by Zambaldi et al. (2019); Stanić et al. (2020) explicitly treat each spatial position in the filter-map of a CNN as a candidate object representation. Even more recent approaches, such as SPAIR (Crawford and Pineau, 2019), SPACE (Lin et al., 2020), and SCALOR (Jiang et al., 2020), expand on this by incorporating explicit features for the presence of an object and its bounding box into each spatial slot. Nonetheless, a current limitation of these approaches is that their spatial slots are typically tailored towards objects that are reasonably well separated, and whose size is compatible with the corresponding receptive field (or the bounding box) in the image.

**CATEGORY SLOTS** A related approach is to allocate slots according to some categorization of objects based on properties other than location. This too can serve to break slot symmetries for the purpose of information routing, and is further expected to mitigate the dependence of spatial slots on spatially separated inputs. In this case, however, because now category and separation are entangled, it is then no longer possible to represent multiple objects of the same category<sup>6</sup>. The main example of category slots are capsules (Hinton et al., 2011, 2018), although other approaches such as Recurrent Entity Networks (Henaff et al., 2017) can also be viewed from this perspective.

### 3.3.2 AUGMENTATION

Augmentation based approaches, unlike slot based ones, keep a single set of features shared among all object representations and instead augment each feature with additional grouping information. This grouping information is usually continuous, which may help to encode uncertainty about the separation. Object representations based on augmentation will trivially be in a common format, although extracting information about individual objects now requires first processing the grouping information. An important limitation of augmentation is that it requires substantial deviations from standard connectionist systems and is thus more difficult to integrate with state of the art systems. Due to features being shared, augmentation may also suffer from capacity and ambiguity problems when a feature is active in multiple object representations at the same time (e.g. two red objects), similar to when representing multiple objects of the same category using category slots Section 3.3.1.

6. There is some evidence that humans struggle with feature overlap too and show reduced working memory capacity in these cases (Mozer, 1989).

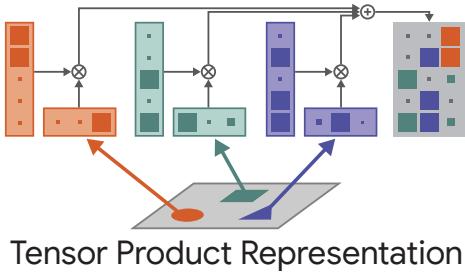


Figure 8: Illustration of a Tensor Product Representation (matrix on the right) that is formed through combining a role vector (horizontal) and a filler vector (vertical) for each object.

**TEMPORAL CODES** An early approach to object representation using augmentation in neural networks made use of the temporal structure of *spiking neurons* for separation (temporal codes). Here, the activation of a feature encoded by the firing rate is augmented with grouping information encoded by the temporal correlation between firing patterns (Singer, 2009). In other words, the features that form an object are represented by neurons that fire in synchrony (Milner, 1974; von der Malsburg, 1981; Singer, 1999; see also Section 6.3). Rather than using unrestricted spiking networks, most work on object representation using temporal codes focuses on *oscillatory networks*, where the firing pattern takes the form of a regular frequency rhythm (for an overview see Wang (2005)). Because temporal codes rely on spiking neurons, they are non-differentiable and also require simulating the dynamics of each neuron even for static inputs. This makes them incompatible with gradient-based training, and necessitates a completely different training framework (e.g. Doumas et al., 2008, 2019) typically based on Hebb’s rule (Kempter et al., 1999), or Spike-Timing-Dependent Plasticity (STDP; Caporale and Dan, 2008).

**COMPLEX-VALUED CODES** An alternative approach to augmentation uses *complex-valued* neurons (features) in place of oscillatory neurons. Hence, instead of explicitly simulating the temporal behavior of an oscillator, its activation and grouping information can now be described as the absolute value and angle of a complex-valued neuron. Similar to before, the grouping is implicit and smooth with neurons that “fire at similar angles” being grouped together. Complex-valued neurons are differentiable and more compatible with existing gradient-based learning techniques. On the other hand, they require specialized activation functions that consider both real and imaginary parts<sup>7</sup>, which tend to be difficult to integrate with existing methods. Successful integrations include complex-valued Boltzmann Machines (Reichert and Serre, 2014; Zemel et al., 1995) and complex-valued RNNs that could be trained either with backpropagation (Mozer et al., 1992) or via Hebbian learning (Rao et al., 2008).

### 3.3.3 TENSOR PRODUCT REPRESENTATIONS

A Tensor Product Representation (TPR) consists of a real-valued matrix (tensor) that is the result of combining distributed representations of *fillers* with distributed representations of *roles*. TPRs can be used for representing multiple objects by associating fillers with object representations and using roles to encode grouping information. A TPR is formed by combining each filler with a corresponding role via an outer product (“binding operation”),

7. In some sense, complex codes can be seen as an instance of a more general – yet unexplored – class of vector-valued activations that use the additional degrees of freedom for grouping.

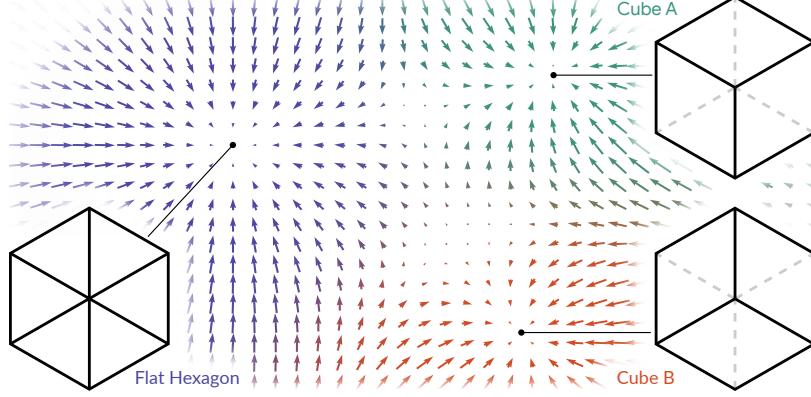


Figure 9: Correspondence of attractor states to visual interpretations for a tri-stable variant of the Necker cube. The vector field illustrates the (input-dependent) inference dynamics in feature space, with one attractor for each stable interpretation.

which are then composed to accommodate multiple object representations (“conjunction operation”). When the role representations are linearly independent, then the object representations can be retrieved from the TPR via matrix multiplication (“unbinding operation”). Notice that, when the role-vectors are one-hot encodings, the TPR reduces to instance slots. However, the additional freedom afforded by a general *distributed* role vector can be used to encode structural information or uncertainty about the separation of objects. TPRs always assume that the object representations are described in a common format. But note that, similar to augmentation, extracting information about individual objects first requires processing the grouping information (in this case via the unbinding operation). TPRs were first introduced in Smolensky (1990) and several modifications have since been proposed that consider different binding, unbinding, and conjunction operations (Plate, 1995; Kanerva, 1996; Gayler, 1998; see Kelly et al., 2013 for an overview). In the recent literature, TPR-like mechanisms have been incorporated into neural networks using fast-weights (Schlag and Schmidhuber, 2018) or self-attention (Schlag et al., 2019) to perform reasoning in language.

### 3.3.4 ATTRACTOR DYNAMICS

Up until this point, we have focused on methods that address the representational format of object representations. Now we consider *attractor dynamics* as an approach for addressing their representational dynamics (Section 3.2). Robust object representations are well described by a stable attractor state in a larger dynamical system that models the representational dynamics based on a given input. In this case, inferring a coherent object representation corresponds to running the dynamical system forward until it converges to an attractor state. A stable attractor is naturally self-correcting, and multiple competing interpretations (from ambiguous inputs) can easily be described by separate attractor states. Top-down feedback can then be used to switch interpretations by pushing the state of the system enough to cause it to cross over to a different basin of attraction. By adapting the system dynamics to changing inputs, they allow for moving attractors (changes of the object) or bifurcations (creation or vanishing of interpretations).

Attractor Networks incorporate attractor dynamics in neural networks and have a long history in connectionist research. Early work includes Hopfield networks (Hopfield, 1982), Boltzmann machines (Ackley et al., 1985), and associative memory (Kohonen, 1989). Attractor states were also found to occur naturally in RNNs, especially when using symmetric

recurrent weights (Almeida, 1987; Pineda, 1987). In recent years, however, they have received little attention (but see Mozer et al. (2018); Iuzzolino et al. (2019)), which might be in part because they can be difficult to train. In particular, the fact that each weight participates in the specification of many attractors can lead to spurious (unintended) attractors and ill-conditioned attraction basins (Neto and Fontanari, 1999). Localist attractor networks (Zemel and Mozer, 2001) and flexible kernel memory (Nowicki and Siegelmann, 2010) are two approaches that address this issue by introducing a separate representation for each attractor. However, note that spurious attractors that correspond to novel feature combinations may also be advantageous for generalization.

### 3.4 Learning and Evaluation

Object representations are the product of segregation and the foundation upon which compositional reasoning is built. To effectively connect high-level abstract reasoning with low-level sensory data they must be learned jointly, together with composition and segregation. Learning object representations requires incorporating architectural inductive biases to ensure a common format and to provide enough flexibility for dynamically separating information. Regarding separation, slot-based approaches offer a simple and minimal approach, while augmentation and TPRs are more difficult to incorporate, yet support more sophisticated use cases. The problem of learning representations that are disentangled can be approached by optimizing for some notion of (statistical) independence between features (e.g. Schmidhuber, 1992c; Chen et al., 2016; Higgins et al., 2017a), sparse feature updates across time (Whitney, 2016), or independent controllability of features (Thomas et al., 2017). In terms of temporal dynamics and robustness, the situation is less clear, although the use of attractor networks may serve as a good starting point.

Evaluation plays a critical role in guiding research to make measurable progress towards good object representations. A useful approach is to measure how well the system copes with particular generalization regimes such as to held-out-combinations of features for disentanglement (Esmaeili et al., 2019) and separation (Santoro et al., 2018b), prediction roll-outs for temporal dynamics (van Steenkiste et al., 2018), and robustness to injected noise for reliability (Mozer et al., 2018). However, in case of poor performance it may be difficult to diagnose the source of the problem in terms of properties of the representational format and dynamics. When ground truth information is available, an alternative is to directly measure selected properties of the object representations, such as local correspondence between ground-truth factors of variation and features for disentanglement (Eastwood and Williams, 2018). Finally, qualitative measures such as latent traversals or projections of the embedding space (van der Maaten and Hinton, 2008) can provide an intuition about the learned representations but due to their subjectivity, quantitative measures should be preferred.



Figure 10: Photo of two leaf-tailed geckos — “young and old” © 2015 by Paul Bertner.

#### 4. Segregation

In this section, we look at the binding problem from the perspective of segregation: the process of forming object representations. Unlike in [Section 3](#), where we focused on the need for binding at a representational level to maintain a separation of information for *given* entities, here we focus on the process of *creating* object representations through binding previously unstructured (raw) sensory information. Humans effortlessly perceive the world in terms of objects, yet this process of perceptual organization is surprisingly intricate ([Wagemans, 2015](#)). Even for everyday objects like a mirror, a river, or a house, it is difficult to formulate precise boundaries or a definition that generalizes across multiple different contexts. Nonetheless, we argue that an important aspect common to all objects is that they may act as stable and self-contained abstractions of the raw input. This then has important implications for the process of segregation.

Consider for example [Figure 10](#), which demonstrates several challenges for segregation that must be overcome. To recognize the two geckos sitting on a branch you have to segment out two unfamiliar objects (zero-shot) even though they belong to the same class (instance segmentation) and their use of camouflage (texture similarity). Both the large gecko and the branch are visually disconnected due to occlusion, and yet you perceive them as independent wholes (amodal completion). Beyond separating these objects, you have also formed separate representations for them that enable you to efficiently relate, describe, and reason about them.

In the following,, we take a closer look at this process of segregation<sup>8</sup>. We first work towards a general *notion* of an object built around modularity and hierarchy ([Section 4.1](#)). Next, we focus on the process of *forming* object representations based on this notion ([Section 4.2](#)). Unlike segmentation, which is typically only concerned with a static split at

---

8. We refer to this process as *segregation* rather than *binding*, to emphasize the fact that it typically requires a *separation* of the inputs and features into meaningful parts.

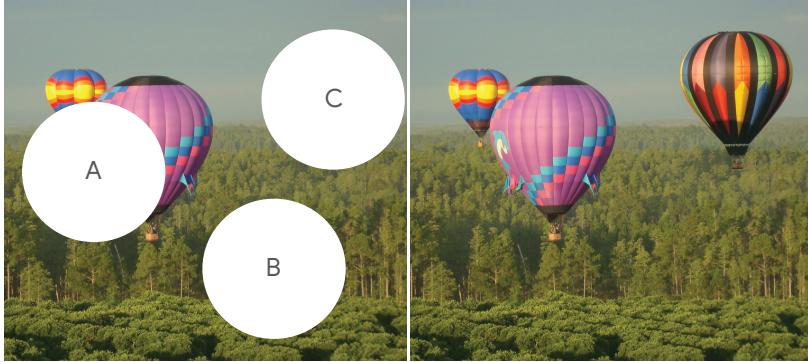


Figure 11: For partial objects (**A**) or only background (**B**), the occluded regions can be inpainted reasonably well, while in the case of full object occlusion (**C**) that is usually impossible.

the *input*-level, segregation is inherently task-dependent and aims to produce stable object representations that are grounded in the input and which maintain their identity over time. Towards the end, we survey relevant approaches from the literature that may help neural networks perform segregation (Section 4.3).

#### 4.1 Objects

The question of what constitutes a meaningful object (i.e. for building structured models of the world) is central to segregation. However, despite long-standing debates in many fields including philosophy, linguistics, and psychology, there exists no general agreed-upon definition of objects (Green, 2018; Cantwell-Smith, 1998). Here, we take a pragmatic stance that focuses on the *functional role* of objects as compositional building blocks. Hence, we are not interested in debating the “true” (i.e. metaphysical) nature of objects, but rather consider object representations as components of a useful representational “map” that refers to (but is not identical to) parts of the “territory” (world)<sup>9</sup>.

##### 4.1.1 MODULARITY

From a functional perspective, the defining quality of an object is that it is modular, i.e. it is self-contained and reusable independent of context. While this suggests choosing objects with minimal information content (to improve reusability), it is equally important that objects can be represented efficiently based on their internal predictive structure. We argue that this trade-off induces a Pareto front of valid decompositions into objects that have both strong internal structure, yet remain largely independent of their surroundings. By organizing information in this way, objects are expected to capture information that is due to independent causes, which matches our intuitive notion of objects in the real world (Green, 2018; Chater, 1996).

Consider the example of three balloons in front of a forest as depicted in Figure 11. When a balloon is partially occluded (as in A), you are still able to make a reasonable guess about the occluded part purely based on its internal predictive structure. On the other hand, when an entire balloon is occluded (as in B) it is impossible to infer its presence from the (unoccluded) context, and the most reasonable reconstruction is to fill in based on the

---

9. “*A map is not the territory it represents, but, if correct, it has a similar structure to the territory, which accounts for its usefulness.*” (Korzybski, 1958).

background (as in C). Notice that each balloon is modular in the sense that it is possible to reuse them in many different contexts (e.g. when placed in a different scene). In contrast, this would not be possible if an object were to be formed from the background *and* the balloon. Hence, by carving up perception at the “borders of predictability”, objects allow for an approximate divide and conquer (i.e. a compositional) approach to modeling the world.

#### 4.1.2 HIERARCHICAL

Objects are often hierarchical in the sense that they are composed of parts that can themselves be viewed as objects. Consider, for example, a house consisting of a roof and walls, which themselves may consist of several windows and a door, etc. Depending on the desired level of detail, a scene can therefore be decomposed in terms of coarser or finer scale objects, corresponding to different solutions on the Pareto front. In most cases, these decompositions relate to each other in the sense that they correspond to different levels in the *same* part-whole hierarchy. However, in rare cases, two decompositions may also consider incompatible parts, as, for example, in a page of text that can be decomposed either into lines or sentences<sup>10</sup>. Notice that there is a difference between this part-whole hierarchy and the feature hierarchy typically found in neural networks. Here, parts are themselves objects, which are the result of dynamically separating information into object representations (segregation). Hence, a part-whole hierarchy can be viewed in terms of a number of general “is-part-of” relations that can be reused between objects (see also [Section 5.1.1](#)).

#### 4.1.3 MULTI-DOMAIN

It is worth emphasizing that objects (as referred to in the context of this paper) are not restricted to vision, but also span sensory information from other domains such as audio or tactile<sup>11</sup> (and even be entirely abstract, although this is not the focus of segregation). For example, auditory objects may correspond to different sources of sound, such as speakers talking simultaneously in the same room (cocktail-party problem; [Cherry, 1953](#)). Objects in the tactile domain are perhaps less obvious, but consider the example of writing on a piece of paper with a pen, where you can clearly separate the sensations that arise from your fingers touching each other, touching the pen, and touching the paper (see also [Kappers and Tiest, 2015](#)). Notice how you are likely to associate the sensations of touching the pen and its visual perception with a common cause and therefore with the same object. This implies that objects can be simultaneously grounded in sensory information from multiple domains, which may help resolve ambiguities (e.g. McGurk Effect; [McGurk and Macdonald, 1976](#)).

## 4.2 Segregation Dynamics

Segregation needs not only infer a decomposition into objects, but also corresponding object representations. As is evident from our previous discussion, there is no universal choice of objects that is appropriate in all circumstances, which requires segregation to consider both

- 
- 10. A unique hierarchy is favored by modularity because in the case of incompatible decompositions (i.e. not corresponding to the same part-whole hierarchy) their objects *cross* “borders of predictability”, which implies a weaker internal structure.
  - 11. It is even discussed whether humans are capable of object perception in the olfactory domain ([Batty, 2014](#)).



Figure 12: Human perception is multistable, which is often demonstrated using visual illusions as in (a), yet it is also often encountered in the real world, e.g. for different groupings of tiles (b). To steer segregation towards a useful decomposition it is important to incorporate contextual information, for example to decide between a decomposition based on chairs or based on stacks in (c).

context- and task-dependent information. Together with the need for a stable outcome, this has several consequences for the segregation dynamics which we will consider next.

#### 4.2.1 MULTISTABILITY

Most scenes afford many different useful decompositions that either stem from choosing different levels of granularity (i.e. levels of hierarchy) or from ambiguous inputs that allow for multiple distinct but coherent interpretations (see multi-modal separation uncertainty Section 3.2.2). Together, these result in a massive number of potential object representations (e.g.  $\geq 3000$  letters per page of text). Simultaneously representing all of them is not only intractable, but also undesirable, as the majority of object representations will not be useful for any particular situation. A practical solution to this problem is a dynamical segregation process that has *multiple stable* equilibria that each correspond to a particular decomposition of a given scene. Indeed, humans resolve this problem via multistable perception, which allows us to seamlessly switch back and forth between different interpretations (Attneave, 1971). This effect is often demonstrated with visual illusions as in Figure 12a, but is in fact much more common than these constructed examples suggest. For example, a simple tile pattern (as in Figure 12b) can easily be perceived in several ways, including rows or columns of tiles. Multistability can also be observed in other sensory modalities such as audio, tactile, and even olfaction (Schwartz et al., 2012). Notice that it is possible to simultaneously perceive multiple objects from the same decomposition, but not from different decompositions (e.g. perceiving 13 and B simultaneously in Figure 12a). This inherent limitation of multistable segregation can also act as an advantage, since it ensures a single coherent decomposition of the input and avoids mixing objects from different incompatible decompositions. It implies that the process of segregation also has to be able to efficiently resolve conflicts from competing decompositions (explaining away).

#### 4.2.2 INCORPORATING TOP-DOWN FEEDBACK

Certain decompositions lead to a set of building blocks (objects) that are more useful than others for a given task or situation. For example, when moving a stack of chairs to another

room it is useful to group information about the individual chairs together as a single object (see Figure 12c). On the other hand, when the goal is to count each of the individual chairs, a more fine-grained decomposition is preferred (and perhaps when repairing a chair an even more fine-grained decomposition is needed). These building blocks underlie the structure of downstream models that can be used for inference, prediction, and behavior, and the choice of decomposition therefore affects the ability to generalize in predictable and systematic ways. Hence it is important that the outcome of the segregation process can be steered towards the most useful decomposition, based on contextual information. One of the main sources of contextual information is *top-down feedback*, for example in the form of task-specific information (e.g. to guide visual search) or based on a measure of success at performing the given task. Memory could act as another source of contextual information, for example by recalling a decomposition that has previously proven useful in the given situation.

#### 4.2.3 CONSISTENCY

It is important that the grounding of object representations, as provided by the segregation process, is both stable and consistent across time (i.e. it maintains object identity). This helps to correctly accumulate partial information about objects, to infer temporal attributes from prior observations (Section 3.2.1), and to ensure that the outcome of more abstract computations in terms of object representations remain valid in the environment (Section 3.2.2). It may also help to avoid “double-counting” of evidence (e.g. during learning)<sup>12</sup>. Object identity depends on a reliable mechanism for re-identification i.e. a mechanism for identifying an object as being the same despite changes in appearance, perspective, or temporary absence of sensory information. Consider, for example, a game of cups and balls, which involves tracking a ball hidden under one of three identical cups that are being moved around. In this case, a stable object identity requires maintaining separate identities for the cups despite their identical appearance, as well as re-identifying the ball as it reappears from under the cup. When an object is re-encountered after a prolonged period, re-identification may require interfacing with some form of long-term memory.

### 4.3 Methods

To succeed at segregation (in the sense outlined above) a neural network must acquire a comprehensive notion of objects and incorporate mechanisms to dynamically route their information. Due to the prohibitive amount of potentially useful objects, it is unlikely that an adequate notion can be engineered directly or taught purely through large-scale supervision. Therefore, in the following, we will review a wide range of approaches, including more classic non-neural approaches that have produced promising results despite incorporating domain-specific knowledge only to a lesser degree. By also discussing the latter, we aim to provide inspiration for the development of neural approaches that can learn about objects directly from raw data (e.g. by focusing on modularity).

---

12. Consider the example from Marcus (2003) about owning a three-legged dog. Despite the fact that you will likely see your dog much more often than other dogs, this series of observations does not affect your overall belief about the number of legs that dogs typically have, since these observations are all associated with *the same* dog.

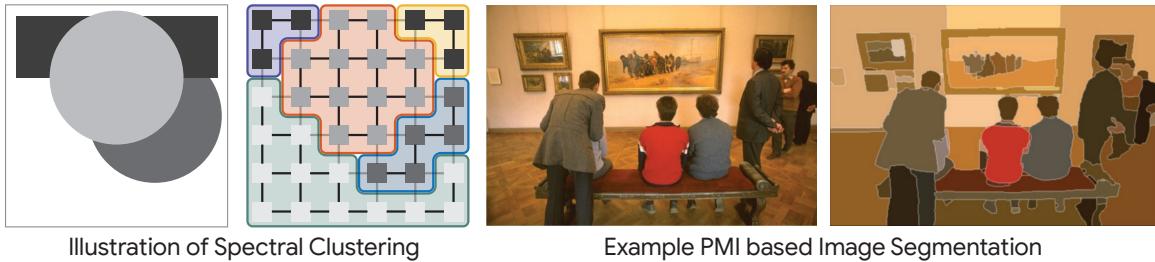


Figure 13: **Left:** An illustration of (spectral) clustering approaches, which treat image segmentation as a graph-partitioning problem. **Right:** Corresponding instance segments as obtained by Isola et al. (2014).

#### 4.3.1 CLUSTERING APPROACHES TO IMAGE SEGMENTATION

Image segmentation is concerned with segmenting the pixels (or edges Arbeláez et al., 2011) belonging to an image into groups (e.g. objects) and therefore provides a good starting point for segregation. A common approach to image segmentation is to cluster the pixels of an image based on some similarity function (Jain et al., 1988). One particularly successful approach is the spectral graph-theoretic framework of normalized cuts (Shi and Malik, 2000), which treats image segmentation as a graph-partitioning problem in which nodes are given by pixels and weighted edges reflect the similarity between pairs of (neighboring) pixels. Partitioning is performed by trading-off the total dissimilarity between different groups with the total similarity within the groups. To the extent that the similarity function is able to capture the predictive structure of the data, this is then analogous to the trade-off inherent to modularity. It is straightforward to achieve a hierarchical segmentation in this graph clustering framework, either via repeated top-down partitioning (Shi and Malik, 2000) or bottom-up agglomerative merging (Mobahi et al., 2011; Hoiem et al., 2011).

In the context of segregation, a central challenge is to define a good similarity function between pixels that leads to useful objects. As we have argued, a hardwired similarity function (e.g. as in Shi and Malik, 2000; Malik et al., 2001) has little chance at facilitating the required flexibility, although different initial seedings of the clustering may still account for multiple different groupings (i.e. multistability). Labeled examples can be used to address this challenge in a multitude of ways, e.g. to learn a similarity function between segments (Ren and Malik, 2003; Endres and Hoiem, 2010; Kong and Fowlkes, 2018) or discrete graphical patterns (Lun et al., 2017), to learn boundary detection (Martin et al., 2004; Hoiem et al., 2011), or as a means of top-down feedback (Mobahi et al., 2011). Unsupervised approaches (based on self-supervision) provide a more promising alternative. One approach is to learn a similarity function between pairs of pixels, e.g. based on their point-wise mutual information using kernel-density estimation (Isola et al., 2014) or based on self-supervised prediction using a neural network (Isola et al., 2015). Alternatively, one can attempt to steer the clustering process based on the unsupervised principle of compressibility (minimum description length; Mobahi et al., 2011).

Notice that, since clustering-based approaches to image segmentation focus on low-level similarity structures, their understanding of objects at a more high-level is limited (i.e. at the level of object representations, but see Bear et al., 2020).

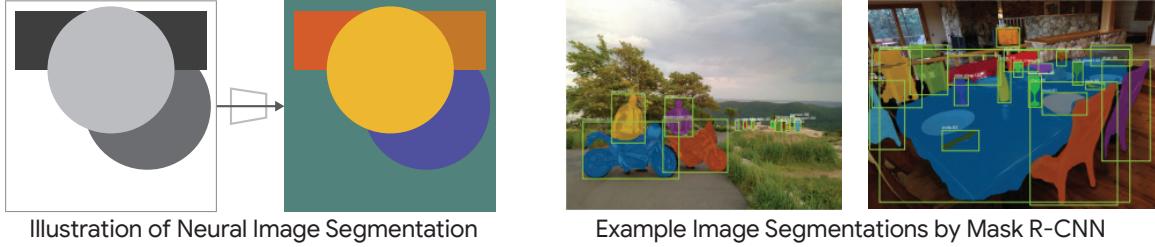


Figure 14: **Left:** An illustration of neural approaches that learn to directly output an image segmentation. **Right:** Corresponding bounding boxes and instance segments as obtained by He et al. (2017).

#### 4.3.2 NEURAL APPROACHES TO IMAGE SEGMENTATION

An alternative approach to image segmentation that leverages the success of end-to-end learning, is to directly output the segmentation with a deep neural network. Unlike clustering-based approaches, which focus on the similarity structure between pixels (or small segments), learning now takes place at the (global) image level, which allows objects to be modeled at multiple levels of abstraction. On the other hand, due to the one-to-one (feedforward) mapping from image to segmentation, it may now be more difficult to provide multiple different segmentations (multistability) or a hierarchical segmentation, for a given input.

Recent approaches based on supervised learning from ground-truth segmentation have produced high-quality instance segmentations of real-world images<sup>13</sup>. For example, approaches based on R-CNN (Girshick et al., 2014) decompose the instance segmentation problem into the discovery of bounding boxes using region-proposal networks (Ren et al., 2015) and mask prediction (Dai et al., 2016; He et al., 2017) to provide instance segmentations. The more recent DETR (Carion et al., 2020) was able to integrate these stages into a single Transformer-based network using a bipartite matching loss. Other approaches output an energy function from which the segmentation is easily derived, e.g. based on the Watershed transformation (Bai and Urtasun, 2017). Instance segmentation has also been phrased as an image-to-image translation problem using conditional generative adversarial networks (Mo et al., 2019). Approximate instance segments can also be obtained as a by-product of performing some other task, such as learning to interpolate between multiple images (Arandjelović et al., 2019) or minimizing mutual information between image segments (Yang et al., 2020).

Unsupervised approaches that directly infer the segmentation (and that do not require large-scale supervision) are more relevant in the context of segregation, but have received far less attention. (Ji et al., 2019) propose to train a neural network to directly output the segment that an input belongs to by maximizing the mutual information between paired inputs in representational space (although it operates at the level of patches as opposed to the global image). In the context of video, motion segmentation often produces segments

13. We would like to emphasize the distinction between *instance* segmentation and *semantic* segmentation. In the context of segregation we are more interested in the former, which is concerned with the more general notion of each segment being an object (instance). In contrast, semantic segmentation associates a particular semantic interpretation (in the form of a label) with each segment, and therefore can not segregate multiple objects belonging to the same class.

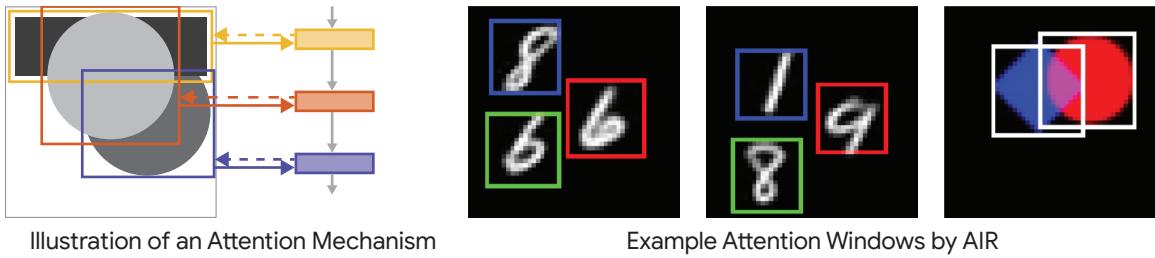


Figure 15: **Left:** An illustration of attention-based approaches, which sequentially attend to individual objects. **Right:** Corresponding attention windows as obtained by Eslami et al. (2016).

that correspond to instances (provided that they move, e.g. Cucchiara et al., 2003), which can for example be learned through unsupervised multi-task learning (Ranjan et al., 2019).

#### 4.3.3 SEQUENTIAL ATTENTION

In the context of segregation, attention mechanisms provide a means to selectively attend to different objects sequentially. Compared to image segmentation, this does not require exhaustively partitioning the image but instead allows one to focus only on the relevant locations in the image (e.g. as a result of top-down feedback). Here we focus mainly on *hard* attention mechanisms that attend to a strict (i.e. spatially delineated) subset of the available information in the form of an attention window, e.g. in the shape of a bounding-box (Stanley and Miikkulainen, 2004) or a fovea (Schmidhuber and Huber, 1991). Their strong spatial bias (due to the shape of the attention window) makes them particularly relevant for the domain of images, but more difficult to adapt to modalities in which meaningful objects are not characterized by spatial closeness. On the other hand, the rigid shape of the attention window may interfere with modularity due to potential difficulties in extracting information about objects with incompatible shapes or that are subject to occlusion.

The main challenge for incorporating attention mechanisms is in correctly placing the window. Early approaches bypass this problem by evaluating a fixed attention window exhaustively at each possible image location, or using several of many heuristics (Lampert et al., 2008; Alexe et al., 2010; Uijlings et al., 2013). A classifier can then be trained to determine which window contains an object (Rowley et al., 1998; Viola and Jones, 2001; Harzallah et al., 2009). Other approaches compute a two-dimensional topographical saliency map that reflects the presence of perceptually meaningful structures at a given location. This facilitates an efficient control strategy to direct an attention window in an image by visiting image locations in order of decreasing saliency (Itti et al., 1998). Salient regions can be learned based on bottom-up information, such as the self-information of local image patches (Bruce and Tsotsos, 2006). Alternatively, they can be derived by also incorporating top-down information, e.g. by highlighting locations that are (maximally) informative with respect to a discriminative task (Gao and Vasconcelos, 2005; Cao et al., 2015; Zhmoginov et al., 2019). Recently, there has been renewed interest in saliency-based approaches through the discovery of keypoints (Jakab et al., 2018; Kulkarni et al., 2019; Minderer et al., 2019; Gopalakrishnan et al., 2020).

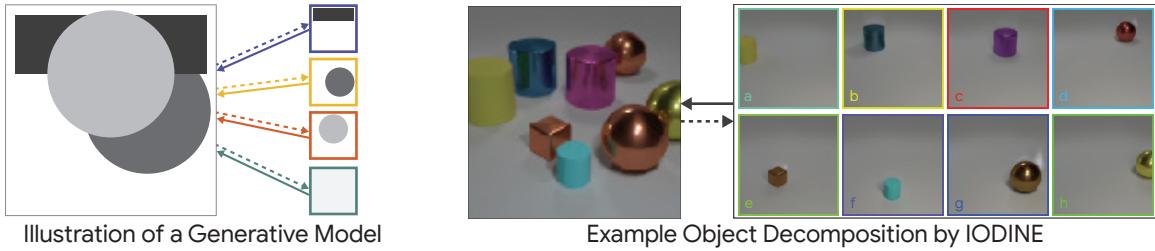


Figure 16: **Left:** An illustration of generative approaches to segregation that model an image as a mixture of components. **Right:** A corresponding decomposition in terms of individual objects as obtained by Greff et al. (2019).

It is also possible to directly learn the control strategy for placing the window of attention, which naturally accommodates top-down feedback. For example, learning the control strategy can be viewed as a reinforcement learning problem, in which the actions of an “agent” determine the location of the window. A policy for the agent (frequently implemented by a neural network) can then be evolved (Stanley and Miikkulainen, 2004), trained with Q-learning (Paletta et al., 2005), or via Policy Gradients (Butko and Movellan, 2009). Alternatively, it can be incorporated as a separate action in an agent trained to perform some task (e.g. classification) or to interact with an environment (Mnih et al., 2014; Ba et al., 2014). AIR (Eslami et al., 2016) and its sequential extension SQAIR (Kosiorek et al., 2018) deploy a similar strategy for an unsupervised learning task with the purpose of extracting object representations. They make use of an attention mechanism that is fully differentiable based on spatial transformer networks (Jaderberg et al., 2015), but see also DRAW (Gregor et al., 2015) for an alternative mechanism. Similarly, Tang et al. (2014) incorporates a window of attention in a deep belief network to extract object representations by performing (stochastic) inference over the window parameters alongside the belief states.

*Soft* attention mechanisms implement attention as a continuous weighing of the input (i.e. a mask) and can be seen as a generalization of hard attention. For example, in MONet (Burgess et al., 2019), GENESIS (Engelcke et al., 2019), and ECON (von Kügelgen et al., 2020) a recurrent neural network is trained to directly support the learning of object representations by outputting a mask that focuses on different objects at each step<sup>14</sup>. A similar soft-attention mechanism has also been used to facilitate supervised learning tasks such as caption generation (Xu et al., 2015), instance segmentation (Ren and Zemel, 2017), or (multi-)object tracking (Kosiorek et al., 2017; Fuchs et al., 2019). Soft attention mechanisms have also been applied *internally* (self-attention) to support segregation. For example, Mott et al. (2019) incorporates a form of dot-product attention (Vaswani et al., 2017) in an agent to attend to the internal feature maps of a bottom-up convolutional neural network that processes the input image. A similar self-attention mechanism was also used to support image classification (Zoran et al., 2020).

#### 4.3.4 PROBABILISTIC GENERATIVE APPROACHES

A probabilistic approach to segregation is via inference in a generative model that models the observed data in terms of multiple components (objects)<sup>15</sup>. An advantage of explicitly modeling the constituent objects is that it is easy to incorporate assumptions about their structure, including modularity and hierarchy. This then enables inference (segregation) to go beyond low-level similarities or spatial proximity, and recover object representation based on their high-level structure as implied by the model. On the other hand, as we will see below, inference usually becomes more difficult as the complexity of the generative model increases, and especially when considering multi-modal distributions (i.e. for multistability).

The most basic assumption to incorporate in a generative model, for the purposes of segregation, is to assume that the input is *directly* composed of multiple parts (objects) that are each modeled individually. Inference in such models then allows one to recover a partitioning of the input in addition to a description of each part (object representation). Early approaches model images with a *mixture model* that treats the color values of individual pixels as independent data points that are identically distributed (Samadani, 1995; Friedman and Russell, 1997). Alternatively, the decomposition can be based on other features such as optical flow (Jepson and Black, 1993) or the coefficients of a wavelet transform (Guerrero-Colón et al., 2008). Mixture models can also be biased towards spatial coherence to explicitly account for the spatial structure of visual objects (Weiss and Adelson, 1996; Blekas et al., 2005). Independent Component Analysis (ICA) models the observed data as linear combinations (mixtures) of unobserved random variables (sources) that are statistically independent (Hyvärinen and Oja, 2000). This approach has been particularly successful at blind source separation (segregation) in the auditory domain (e.g. the cocktail party problem Cherry, 1953), although it has also seen application in the context of images (Lee and Lewicki, 2002).

To more accurately model complex data distributions, it is possible to incorporate domain-specific assumptions in the generative model (and thereby improve the result of inference). For example, a generative model that captures the geometry of 3D images of indoor scenes as well as the objects that are in it “[...] integrates a camera model, an enclosing room ‘box’, frames (windows, doors, pictures), and objects (beds, tables, couches, cabinets), each with their own prior on size, relative dimensions, and locations” (Del Pero et al., 2012). The results that can be obtained by incorporating domain-specific knowledge are impressive (Zhao and Zhu, 2011; Del Pero et al., 2012, 2013; Tu et al., 2005; Tu and Zhu, 2002). However, performing inference in highly complex generative models of this type is problematic and frequently relies on custom inference methods tailored to this particular task (e.g. Markov Chain Monte Carlo using jump moves to remove or add objects or specific initialization strategies). In recent years, *probabilistic programming languages* have emerged as a general-purpose framework to simplify the design of complex generative models and the corresponding inference process. For example, they have enabled the use of symbolic graphic renderers as forward models (Mansinghka et al., 2013) and incorporated deep neural networks to help make inference more tractable (Kulkarni et al., 2015; Romaszko et al., 2017).

- 
- 14. Notice, however, that these particular methods enforce an *exhaustive* partition of the image similar to image segmentation methods.
  - 15. Human perception is also said to be generative in the sense that we often perceive objects as coherent wholes even when they are only partially observed (amodal completion; Michotte et al., 1991).

Nonetheless, in the context of segregation, the amount of domain-specific engineering that is still required limits their generality and applicability to other domains (similar to overly relying on supervised labels from a particular domain).

An alternative approach to more accurately modeling complex data distributions is to incorporate fewer assumptions, and rather parameterize the generative model with a neural network that can *learn* a suitable generative process from many different observations. For example, van Steenkiste et al. (2020) demonstrates how a (spatial) mixture model that combines the output of multiple deep neural networks is able to learn to generate images as compositions of individual objects and a background (see also Nguyen-Phuoc et al., 2020; Ehrhardt et al., 2020; Niemeyer and Geiger, 2020). However, in order to perform segregation, we must also be able to perform inference in these models, which can be very challenging. This has been addressed by simultaneously learning an amortized iterative inference process based on de-noising (Greff et al., 2016), generalized expectation-maximization (Greff et al., 2017), iterative variational inference (Greff et al., 2019), slot attention (Locatello et al., 2020), or parallel spatial (bounding-box) attention (Lin et al., 2020; Jiang and Ahn, 2020). Further improvements can be made by assuming access to multiple different views to explicitly model 3D structure at a representational level (Chen et al., 2020; Nanbo et al., 2020). Even though these methods still struggle at modeling complex real-world images, they are capable of learning object representations that incorporate many of the previously mentioned desiderata (e.g. common format, disentangled, modular), in a completely unsupervised manner.

#### 4.4 Learning and Evaluation

The main challenge in segregation is in coping with the immense variability of useful objects that depend on both task and context. We have argued that this effectively precludes solutions that overly rely on supervision or domain-specific engineering. This raises the question of how a useful notion of an object can be discovered mainly via unsupervised learning (and later refined based on task-specific information). A key part of the answer is to focus on the modularity of objects, which only depends on the statistical structure of the observed data and interfaces directly with the functional role of objects as compositional building blocks. Indeed, evidence suggests that human object perception is based on similar principles (Orbán et al., 2008; Chater, 1996). In the machine learning literature, several approaches have also shown to be able to successfully leverage modularity to learn about objects, either in combination with spectral clustering (Isola et al., 2014), attention (Burgess et al., 2019), or by using neural mixture models (Greff et al., 2019), or an adversarial formulation (Yang et al., 2020). Additionally, also focusing on other properties of objects such as common fate (e.g. motion) may play an important role in further improving these results (e.g. Pathak et al., 2017; Ranjan et al., 2019).

Regarding segregation dynamics, we have seen that it is important to provide architectural inductive biases that help with dynamic information routing, e.g. in the form of attention or masking specific parts of the input. Consistency and top-down feedback are mostly affected by the interplay between segregation, representation, and composition, and it is difficult to evaluate these properties in isolation. However, in order to facilitate this interaction, it is critical that segregation is part of a fully-differentiable neural approach, which may be

most problematic for clustering-based approaches to image segmentation and probabilistic programs based on symbolic models.

Segregation is best evaluated in the context of a larger system, where the resulting object representations form the foundation of structured models for inference, behavior, and prediction. In this case, the ability to transfer learned object representations to other tasks, and improving sample-efficiency (semi-supervised) is of particular interest (Wei et al., 2020). Alternatively, when ground-truth information about objects is available, individual aspects of segregation can be evaluated more directly. For example, when a pixel-level segmentation is produced as part of segregation, then metrics such as AMI (Vinh et al., 2010) can be used to compare against the ground-truth. This also provides a means to probe multi-stability for inputs that are known to have multiple stable interpretations. Finally, consistency can also be evaluated in this way, namely by measuring how stable the inferred notion of an object is across a temporal sequence (e.g. object tracking).



Figure 17: Three different objects ( $\blacksquare$ ,  $\bullet$ ,  $\star$ ) appear in different pairings on a scale (a) and (b). By evaluating their relationships (d), it can be inferred how the scale will tip in (c).

## 5. Composition

In this section, we look at the binding problem from the perspective of composition: building structured models of the world that are *compositional*. Here we encounter the need for variable binding: the ability to combine object representations and relations without losing their integrity as constituents (as is needed for compositionality). As we have seen in Section 2, compositionality is a core aspect of human cognition and underlies our ability to understand novel situations in terms of existing knowledge. Similarly, in the context of AI, it supports the systematic reuse of familiar objects and relations to dynamically construct novel inferences, predictions, and behaviors, as well as the ability to efficiently acquire new concepts in relation to existing knowledge.

Consider the sequence of observations in Figure 17, which allows you to infer the relative weights of the three depicted objects ( $\blacksquare$ ,  $\bullet$  and  $\star$ ). Several interesting observations can be made. For example, from panel (a) you can tell that  $\bullet$  is heavier than  $\blacksquare$ , and likewise, that  $\star$  is heavier than  $\bullet$  from panel (b). This information does not describe a property of any of the individual objects, but rather a *relation* between them. On the other hand, it can still be used to update the properties of the participating objects in response to new information (e.g. the precise weight of  $\blacksquare$ ) or to respond to generic queries, such as answering which of the objects is the heaviest. The latter, in this case, also requires comparing the weights of  $\blacksquare$  and  $\star$  (panel (c)). Notice how this is only possible through transitivity of the ‘‘heavier than’’ relation, which allows you to combine the relations from panels (a) and (b) to infer that  $\star$  is heavier than  $\blacksquare$ .

In the following, we take a closer look at how to enable neural networks to dynamically implement *structured models* for a given task, with the ultimate goal of generalizing in a more systematic (human-like) fashion. First, we focus on incorporating a compositional structure that combines relations and object representations without undermining their modularity (Section 5.1). Next, we consider how a neural network can dynamically infer the appropriate structure and leverage it for the purpose of reasoning (Section 5.2). Towards the end, we survey relevant approaches from the literature that address these aspects of composition (Section 5.3).

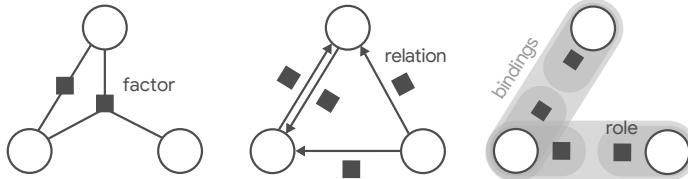


Figure 18: Three different ways in which structure can be defined in terms of relations between objects: As a factor graph, a directed graph, or as nested role-filler bindings.

## 5.1 Structure

To implement structured models, a neural network must organize its computations to reflect the desired *structure* in terms of objects and their relations. This structure is generally described by a graph where nodes correspond to objects and edges to relations<sup>16</sup>. By representing relations separately (independent of object representations) it is possible to freely compose relations and objects to form arbitrary structures (i.e. corresponding to different graphs). However, certain types of relations may also impose constraints on the structure to ensure internal consistency between relations (e.g. symmetry, transitivity).

### 5.1.1 RELATIONS

Relations encode the different computational interactions between the object representations in a structured model. Many different types of relations are possible, including causal relations (e.g. “collides with”), hierarchical relations (“is part of”), or comparative relations (e.g. “bigger than”). Moreover, these general relations can often be specialized to include the nature or strength of an interaction (e.g. “elastic collision”, “much bigger than”). To efficiently account for this variability and support learning, relations are best encoded using flexible (neural) representations. Similar to object representations, it may then also be desirable to use a common format that provides a measure of similarity between relations and ensures that they can be used interchangeably<sup>17</sup>. The way structure is defined in terms of relations may also have implications for their corresponding representations. When the structure is given by a regular (directed) graph or a factor graph (see Figure 18 a & b), then each relation is encoded by a single representation corresponding to either an edge or a factor. Alternatively, it is possible to encode a relation with multiple representations that correspond to the different *roles* that the participating objects play (see Figure 18 c). Finally, it is important that relations are represented separate from and independent of the object representations (see also *role-filler-independence*; Hummel et al., 2004). This enables relations and objects to be composed in arbitrary ways to form a wide variety of (potentially novel) structures.

16. In our discussion, we focus mainly on binary relations (e.g. A is bigger than B) that are well represented by individual edges. However, keep in mind that it is also possible to represent higher-order relations (e.g. A divides B from C), either by using a higher-order graph (e.g. a factor graph) or with the help of auxiliary nodes (e.g. by adding a ‘division node’ with binary relations to A, B, and C).

17. Doumas et al. (2008) even argues that objects and relations should in fact use a *shared* ‘feature pool’ with which both can be described.

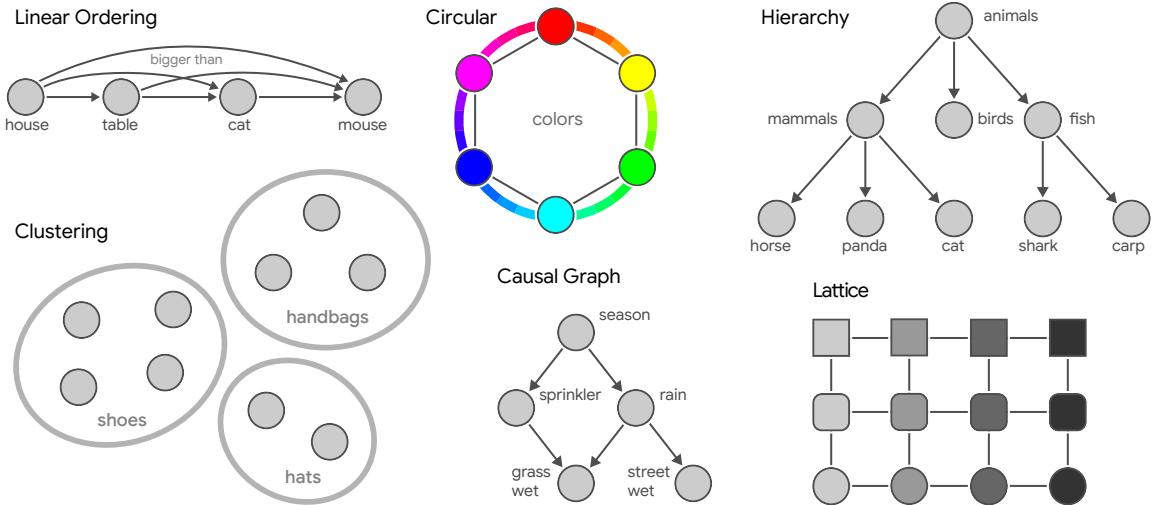


Figure 19: Examples of different structural forms (Kemp and Tenenbaum, 2008) that each can be used to define relations among objects and imply different patterns of generalization.

### 5.1.2 VARIABLE BINDING

To enable a single neural network to implement different structured models, it requires a suitable ‘variable binding’ mechanism<sup>18</sup> that can *dynamically* combine modular object representations and relations. Consider the classic example of Mary and John adapted from Fodor and Pylyshyn (1988): Depending on a given task or context it may be more important to consider that “Mary loves John”, that “John is taller than Mary”, or that “Mary hit John”. In general, the number of possible structures that can be considered is potentially very large, and it is, therefore, intractable to represent all of them simultaneously. Apart from being dynamic, a suitable variable binding mechanism should also preserve the modularity of individual object representations. This is critical to implement structured models that are *compositional*, which ensures that the neural network generalizes systematically and predictably with respect to the underlying objects.

In many cases, only a single level of variable binding that directly combines individual object representations and relations is needed. However, in certain other cases (e.g. “Bob knows that Mary loves John”) it may be required to first build composite structures that can themselves act as ‘objects’, and that can then be combined recursively. When using a role-based representation for relations, multiple levels of variable binding are also needed to avoid ambiguity when a low-level object representation plays the same role in multiple relations.

### 5.1.3 RELATIONAL FRAMES

Each *type* of relation focuses on a particular aspect of the broader interaction among objects, and thereby defines a particular *relational frame* that is internally consistent. Consider again the example in Figure 17, which was concerned with the “heavier than” relation. This

18. The term variable binding is adapted from mathematics, where it refers to binding the free variables in an expression to specific values. In our case, variables correspond to object representations that are bound to the structure determined by the relations.

corresponds to a relational frame of comparison that induces an ordering among the objects in terms of their weight. In this case, an internally consistent ordering requires the relation to be transitive (i.e.  $A > B \cap B > C \Rightarrow A > C$ ) and anti-symmetric (i.e.  $A > B \Rightarrow B \not> A$ ). More generally, a relational frame is characterized by a particular type of relation, and by the logical consequences (i.e. different entailments) that are implied by having (multiple) relations of this type within the structure. We adopted the term relational frame from Relational Frame Theory (RFT; see also [Section 6.4](#)), which distinguishes two types of entailment that humans primarily use to derive (unobserved) relations: *mutual entailment* and *combinatorial entailment*. Mutual entailment is used to derive additional relations between two objects based on a given relation between them, e.g. anti-symmetry for a frame of comparison, or symmetry for a frame of coordination (i.e. deriving  $B = A$  from  $A = B$ ). Analogously, combinatorial entailment is used to derive new relations between two objects, based on their relations with a shared third object, e.g. transitivity for a frame of coordination (i.e. deriving  $A = C$  from  $A = B$  and  $B = C$ ).

Many different types of relational frames can be distinguished, which can be organized into a number of general classes ([Hughes and Barnes-Holmes, 2016](#)), including ‘coordination’ (e.g. same as) , ‘comparison’ (e.g. larger than), ‘hierarchy’ (e.g. part of) , ‘temporal’ (e.g. after), or ‘conditional’ (e.g. if then). Their corresponding rules for entailment give rise to different *structural forms* ([Kemp and Tenenbaum, 2008](#)) among their relations, such as trees, chains, rings, and cliques (see [Figure 19](#)). In this way, each relational frame can also be seen as encoding a particular (systematic) *pattern of generalization* among the objects. Multiple different relational frames may co-occur within the same structure, which allows for rules of entailment to interact across different frames to facilitate more complex generalization patterns (e.g.  $A = B$  and  $B > C$  implies  $A > C$ ).

## 5.2 Reasoning

The appropriate structure for a model depends on the task and context, and should therefore be dynamically inferred by the neural network to focus only on relevant interactions between the objects. Likewise, it is important to consider the computational interactions between relations and object representations, in order to make use of the inferred structure for prediction and behavior.

### 5.2.1 RELATIONAL RESPONDING

To leverage a given structure in terms of relations between object representations, a neural network must be able to organize its computations accordingly. A common use case involves adjusting the (task-specific) response to an object based on its relation to other objects (relational responding). For example, if it is known that ■ is heavier than •, then learning that • is too heavy for a particular purpose (task) also changes your behavior concerning ■. More generally, relational responding of this kind may involve evaluating multiple (derived) relations between objects and combining information across different relational frames. Another use case is in implementing so-called *structure sensitive operations* ([Fodor and Pylyshyn, 1988](#)) that require responding *directly* to the structure given by the relations (independent of the object representations). This is especially important for solving abstract

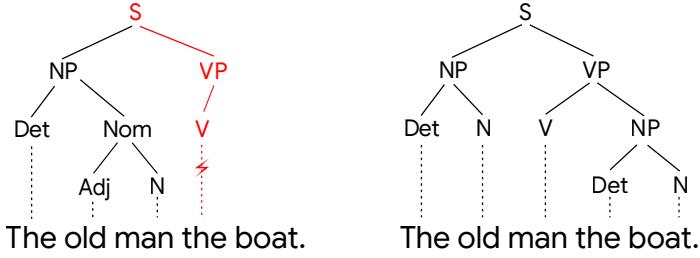


Figure 20: Two parse-trees of a garden-path sentence: The intuitive parsing (on the **left**) fails, even though the sentence is grammatically correct (see parse-tree on the **right**).

reasoning tasks, e.g. when applying the distributive law to a given mathematical expression (i.e. turning  $a \cdot (b + c)$  into  $a \cdot b + a \cdot c$ ).

A natural choice for facilitating relational responding in a neural network is to organize its internal information flow (i.e. computations) in a way that reflects the graph structure of relations and objects. This ensures that newly available information affects the object representations in accordance with the dependency structure implied by the relations (and therefore also with the generalization patterns due to the relational frames). Most information processing of this kind can then be implemented in terms of only local interactions between objects representations and relations, which maximally leverages their modularity. These local interactions, which can either be instantaneous (e.g. collides with) or persistent (e.g. is part of), can facilitate both directed (e.g. for causal relations) and bidirectional (e.g. for comparison) information flow. On the other hand, local interactions are ill-suited for implementing structure sensitive operations that require simultaneously considering multiple different parts of the larger structure.

### 5.2.2 INFERRING STRUCTURE

Inferring the most desirable structure is an inherently difficult task, which requires making many individual choices at the level of relations that all have to be coordinated to ensure that the structure as a whole is useful. One important guiding constraint is the internal consistency of the structure with respect to the rules of entailment as implied by the choice of relational frames. Inconsistencies between the observed information and predictions by the structured model are another indicator of a wrong or incomplete structure. The ‘garden-path’ sentence “The old man the boat.” (see Figure 20) provides a good example for a violation of expectations, which then triggers a revision of the structure. Upon first reading, “The old man” is likely parsed as the subject of the sentence, which implies a structure where the next word is expected to be a verb. However, since “the boat” is not a verb (and therefore does not match this expectation), the sentence cannot be parsed in this way. The problem is resolved by revising the structure so that it takes “The old” as the subject and “man” as the *verb* of the sentence. This example also illustrates the need for collaboration between composition and segregation: It was the initial grouping of “The old man” as a single object that gave rise to inconsistencies at the level of structure, which could only be resolved by also changing the outcome of the segregation process. Hence, it is vital that the process of inferring structure is able to provide (top-down) feedback to help guide the process of segregation.

Inferring structure at the level of individual relations between objects involves making choices about the type of relation, or which of the properties of an object to relate. These

decisions can be guided by *contextual cues* from the environment, such as the scales in Figure 18 that trigger a comparison of the objects in terms of their masses (as opposed to e.g. their relative position or shape). Inferring a relation between objects may also be triggered upon discovering their relation to other objects (e.g. due to combinatorial entailment). However, for the sake of efficiency it may not always be desirable to explicitly represent such relations, but rather model their effect implicitly due to appropriately organizing the computations of the network (i.e. relational responding). More generally, the process of inferring structure has to interface closely with the mechanism for variable binding (i.e. for dynamically combining modular object representations and relations in a way that preserves their modularity).

### 5.3 Methods

To succeed at composition, a neural network requires a mechanism for organizing its internal computations in a way that facilitates relational responding based on the desired structure. A natural approach is to incorporate the structure at an *architectural level* by focusing directly on the local interactions between objects representations and relations. Alternatively, one can also use a more generic (recurrent) neural network “processor” that (sequentially) operates on a *representation* of the desired structure. In the following we will review both of these different approaches, focusing in particular on relational responding and the difficulty of inferring structure<sup>19</sup>.

#### 5.3.1 GRAPH NEURAL NETWORKS

Graph Neural Networks (GNNs; Scarselli et al., 2009; Pollack, 1990) are a promising approach for composition that incorporates the desired structure for relational responding at an architectural level (see Wu et al. 2020 for an overview). At a high level, a GNN is a neural network that is structured according to a graph whose edges determine how information is exchanged among the nodes. In the context of composition, nodes correspond to object representations and edges to relations, which together form the structure, i.e. using (static) variable binding at the architectural level. A GNN fundamentally distinguishes two kinds of information processing, one that requires evaluating the relations between the object representations, and another that is concerned with combining (aggregating) the effect of the incoming relations to update the object representations (Battaglia et al., 2018). By implementing these in a general way that applies equally to different objects and relations, a GNN can accommodate many different structures. In general, the local information processing in a GNN ensures that information affects the object representations in a way that follows the dependency structure implied by the relations (relational responding).

**GRAPH CONVOLUTIONAL NETWORKS** Graph Convolutional Networks (GCNs) are a type of GNNs based on a generalization of convolutional neural networks (which operate on grids) to non-Euclidean geometries such as graphs (Bronstein et al., 2017). A GCN consists

---

19. We note that the problem of inferring structure has also received considerable attention in the causality literature, often specifically focusing on cause-effect discovery (e.g. see Hoyer et al. (2009); Lopez-Paz et al. (2015); Peters et al. (2016) or Peters et al. (2017) for an overview). Generally, we expect structural causal models to become highly relevant for composition, due to their robustness under intervention and utility for reasoning about hypothetical or unobserved scenarios (Pearl, 2019; Schölkopf, 2019).

of several layers that each produce an updated set of node representations by applying graph-convolutions to a local neighborhood in the graph. They have been successfully applied to a wide variety of graph-structured data including social networks (Hamilton et al., 2017), citation networks (Kipf and Welling, 2017), 3D surfaces (Litany et al., 2018), knowledge base completion tasks (Schlichtkrull et al., 2018), and bio-chemical modeling (Atwood and Towsley, 2016). However, while they excel at modeling large-scale graphs, one disadvantage of GCNs in the context of composition is that they assume a given graph in the form of an adjacency matrix and node representations as input. For the purpose of composition, scalability is less important since we are most interested in relatively small graphs (restricted by working memory) that are composed dynamically. On the other hand, some GCNs (e.g. Henaff et al., 2015; Lee et al., 2019) have used a mechanism for coarsening (down-sampling) the graph between layers, to reduce computational complexity, which could provide a mechanism for refining the structure (i.e. structure inference).

**MESSAGE PASSING NEURAL NETWORKS** Message Passing Neural Networks (MPNNs; Gilmer et al., 2017) iteratively update the node representations of a given graph by exchanging messages along its edges (until convergence)<sup>20</sup>. Compared to GCNs, both the graph structure and weights are shared across layers (iterations), and the messages (corresponding to the incoming relations) are typically implemented as a pairwise *non-linear* function of both adjacent node representations. Hence, edges play a more prominent role in information processing and by explicitly considering pair-wise interactions it is easier to model comparative relations between objects. MPNNs were initially conceived as a generalization of RNNs to graph-structured inputs (Sperduti and Starita, 1997; Gori et al., 2005) and have since been adapted to consider modern deep neural networks (Li et al., 2016). A more general framework that accommodates both MPNNs and GCNs was proposed in Battaglia et al. (2018), which additionally includes a global representation of the graph that interacts with all the nodes and edges (and may thereby more easily provide for structure-sensitive operations).

MPNNs have been shown to generalize more systematically (compared to standard neural networks) on many different tasks that require relational responding in terms of objects, including common-sense physical reasoning (Chang et al., 2017; Battaglia et al., 2016; Janner et al., 2019), hierarchical physical reasoning (Mrowca et al., 2018; Li et al., 2020; Stanić et al., 2020), visual question answering (Santoro et al., 2017; Palm et al., 2018), abstract visual reasoning (Andreas, 2019), natural language processing (Tai et al., 2015), physical construction (Hamrick et al., 2018) or multi-agent interactions (Sun et al., 2019). Similar to GCNs, the desired structure may either be specified directly or inferred dynamically based on some heuristic, e.g. based on proximity (Chang et al., 2017; Mrowca et al., 2018) or a language parser (Tai et al., 2015). Alternatively, MPNNs have been used to implement a relational inductive bias based on a generic structure, e.g. by assuming it to be fixed and fully connected (as in Relation Networks; Santoro et al., 2017). In this case, information can still be exchanged among all the nodes, although the generalization implied by having the correct structural dependencies is lost (e.g. for entailment).

A more desirable approach is to (dynamically) infer the desired structure, although this is challenging due to the discreteness of graphs and difficulties in comparing them efficiently. One approach is to first learn a continuous embedding for all possible graph structures and

---

20. Recently, MPNNs were extended to allow for continuous updates (Deng et al., 2019; Liu et al., 2019).

then optimize for the right structure in the corresponding space, e.g. using VAEs (Kusner et al., 2017; Zhang et al., 2019), or GANs (Yang et al., 2019). The other approach is to directly infer the connectivity between nodes iteratively based on message passing, e.g. for a fixed number of nodes as in Neural Relational Inference (NRI; Kipf et al., 2018) or adaptively as in Graph Recurrent Attention Networks (GRANs; Liao et al., 2019).

**APPROACHES BASED ON SELF ATTENTION** Graph Neural Networks based on *self-attention* are closely related to MPNNs. The main difference to MPNNs is that they use self-attention to compute a *weighted* sum of the incoming messages (based on the relations) for updating the node representations. This provides a useful mechanism for dynamically adapting the information routing (here a kind of soft variable binding) and thereby infer the desired structure for a fixed set of nodes. However, note that this may be computationally inefficient because it still requires computing all possible messages and only affects which of them end up being used in the final summation. Wang et al. (2018) makes use of a kind of (learned) dot-product attention to infer relations between spatial slots. In this case, the attention coefficients are computed for pairs of nodes while the messages are based only on a single node, which may make it more difficult to implement multiple different relations. The use of *multiple attention heads* (i.e. as in Vaswani et al., 2017) may help mitigate this issue and has been successfully applied for relational reasoning about objects (Zambaldi et al., 2019; van Steenkiste et al., 2020; Goyal et al., 2019; Santoro et al., 2018a), citation networks (Veličković et al., 2018), question answering (Dehghani et al., 2019), and language modeling (Devlin et al., 2019; Brown et al., 2020). Indeed, Transformers themselves may already be viewed as a kind of graph network (Battaglia et al., 2018). Alternatively, multiple different relations could be learned by also conditioning the message on the receiving object representation when using attention e.g. as in R-NEM (van Steenkiste et al., 2018). The idea of using (self-)attention as a mechanism for inferring structure (and dynamic information routing) has also been applied outside the scope of graph neural networks, e.g. in pointer networks (Vinyals et al., 2015), energy-based models (Mordatch, 2019), and capsules (Sabour et al., 2017; Kosiorek et al., 2019).

### 5.3.2 NEURAL COMPUTERS

Neural computers offer an alternative approach to composition by learning to perform reasoning operations sequentially on some appropriate representation of the desired structure. In this case, the ‘processor’ is typically given by an RNN that interfaces with other components, such as a dedicated memory, via a prescribed set of differentiable operations. Compared to a GNN, the architecture of a neural processor is more generic and does not directly reflect the desired dependency structure in terms of relations between object representations. Instead, by considering structure at a representational level, it can more easily be adjusted depending on task or context. Similarly, by having a *central* processor that is responsible for relational responding (as opposed to a distributed GNN) it is easier to support operations that require *global* information (e.g. structure-sensitive operations). On the other hand, the ability of neural computers to learn more general algorithms comes at the cost of a weaker inductive bias for relational reasoning specifically. Hence, it is often necessary to incorporate more specialized mechanisms to efficiently learn algorithms for relational responding that generalize in agreement with the desired structure.

The most common type of neural computer consists of an RNN (the processor) that interfaces with an external differentiable memory component. A dedicated memory component provides an interface for routing information content (now stored separately) to the *variables* that take part in processing (i.e. the program executed by the RNN processor). Indeed, while an RNN can in principle perform any kind of computation using only its hidden state as memory (Siegelmann and Sontag, 1991), its dual purpose for representing structure and information processing makes it difficult to learn programs that generalize systematically (Lake and Baroni, 2018; Csordás et al., 2020). Early examples of memory-augmented RNNs (Das et al., 1992; Mozer and Das, 1993) use a continuous adaptation of stacks based on the differentiable push and pop operations introduced by Giles et al. (1990) (cf. Joulin and Mikolov, 2015 for an alternative implementation). Although a stack-based memory has proven useful for learning about the grammatical structure of language(e.g. Das et al., 1992)), its utility for more general reasoning tasks is limited by the fact that only the top of the stack is accessible at each step.

The addressable memory used in the Neural Turing Machine (NTM; Graves et al., 2014) offers a more powerful alternative, which can be accessed via generic read and write operations (but see memory networks for a read-only version; Weston et al., 2015; Sukhbaatar et al., 2015). In this case, all memory slots (and thereby all parts of the structure) are simultaneously accessible through an attention mechanism (responsible for variable binding) that supports both content- and location-based addressing. Together, these operations have shown to provide a useful inductive bias for learning simple algorithms (e.g. copying or sorting) that generalize to longer input sentences (i.e. more systematically). Additional memory addressing operations, e.g. based on the order in which memory locations are accessed (DNC; Graves et al., 2016), based on when they were last read (Munkhdalai and Yu, 2017), or based on a key-value addressing scheme (Csordás and Schmidhuber, 2019) may confer additional generalization capabilities that are especially relevant for relational reasoning. For example, the DNC has shown capable of learning traversal and shortest path algorithms for general graphs by writing an input sequence of triples ('from node', 'to node', 'edge') to memory, and iteratively traverse this structure using content-based addressing (Graves et al., 2016). Moreover, given a family tree consisting of ancestral relations between family members, the DNC can successfully derive relationships between distant members, which demonstrates a form of combinatorial entailment.

Other memory-based approaches take a step towards GNNs by updating each memory location in parallel (Henaff et al., 2017; Kaiser and Sutskever, 2016) or incorporate specialized structure for reasoning into the processor, e.g. for the purpose of visual question answering using a read-only memory (knowledge base; see Hudson and Manning, 2018). Alternatively, certain (Hebbian) forms of fast weights (Schmidhuber, 1992a) can be viewed as a type of *internal* associative memory based on previous hidden states (Ba et al., 2016). TPR-RNN (Schlag and Schmidhuber, 2018) extends this idea by equipping a fast-weight memory with specialized matrix operations inspired by Tensor Product Representations (TPR; Smolensky, 1990), which makes it easier to respond to relational queries. In contrast, Reed and de Freitas (2015) and Kurach et al. (2016) take a step towards modern computer architectures by, respectively, incorporating a call-stack with an explicit compositional structure or a mechanism for manipulating and dereferencing pointers to a differentiable memory tape.

## 5.4 Learning and Evaluation

The problem of composition is about implementing structured models with neural networks that take advantage of the underlying compositionality of object representations. We have argued that this requires incorporating mechanisms for dynamic variable binding (such as attention), and for dynamically organizing internal information processing for the purpose of relational responding. Regarding the latter, the choice of a suitable mechanism is less clear, although evidence indicates that a GNN-based approach is promising.

With the right mechanisms in place, it is reasonable to expect that relations, relational frames, and structure inference can all be learned (jointly with segregation and representation) via mostly unsupervised learning (Kemp and Tenenbaum, 2008). On the other hand, learning about relations in particular may be challenging, since they can never be observed directly, but always occur in conjunction with concrete objects. Indeed, young children initially reason primarily based on the perceptual similarity between objects and learn to pay attention to their relational similarity only at a later stage (i.e. after undergoing a “relational shift”; Gentner and Rattermann, 1991). A key enabler for children to acquire progressively more general relations is multi-exemplar training: repeated exposure to the same relation, but in combination with different fillers (Barnes-Holmes et al., 2004; Luciano et al., 2007). This idea has been successfully adapted for learning abstract relations using spiking neural networks (Doumas et al., 2008), and shares similarities to more recent contrastive learning objectives that require a neural network to infer relations from a dataset of positive and negative pairings (Kipf et al., 2020; Hadsell et al., 2006). The ability to interact with the environment may additionally enable an (embodied) agent to autonomously acquire multi-exemplar data for a *particular* relation (Schmidhuber, 2015; Haber et al., 2018). An alternative approach to learning composition is to view dynamic structure inference as a meta-learning problem and directly optimize for (systematic) generalization, e.g. by minimizing the generalization regret in face of deliberate non-stationarity (Bengio et al., 2019).

The ultimate goal of composition is to facilitate more systematic generalization and several methods have been proposed that measure different aspects of this ability. A prototypical approach is to evaluate a trained system on a set of held-out combinations of parts (objects) as an approximate measure of systematicity (Santoro et al., 2018b; Lake and Baroni, 2018; Hupkes et al., 2020). A similar strategy can also be used to assess the capacity for interpolation or extrapolation, i.e. by varying the number of parts or range of values. Additionally, Hupkes et al. (2020) propose to measure (systematic) “overgeneralization errors” that are indicative of a bias towards a particular pattern of generalization.

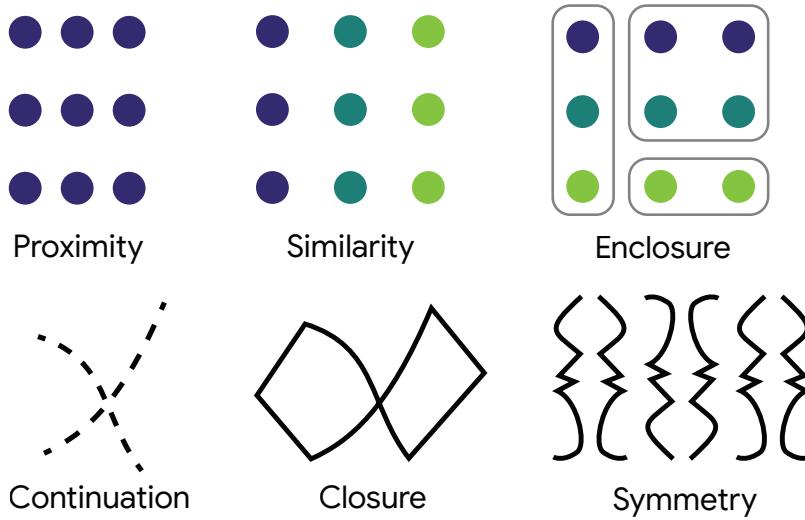


Figure 21: Illustration of several Gestalt Laws of visual perception. Note how the different cues influence which elements are perceived as belonging together.

## 6. Insights from Related Disciplines

Object perception and the symbolic nature of human cognition have been studied from various angles in Neuroscience, Psychology, Linguistics, and Philosophy. These complementary perspectives provide valuable inspiration for addressing the binding problem and we have frequently drawn upon their insights throughout this survey. While an exhaustive overview is outside the scope of this survey, we provide a brief discussion of the areas that were most influential to the development of the conceptual framework presented here. These fields have a lot more to offer and we encourage the reader to further explore this literature, for example by using the pointers and connections provided here as entry-points.

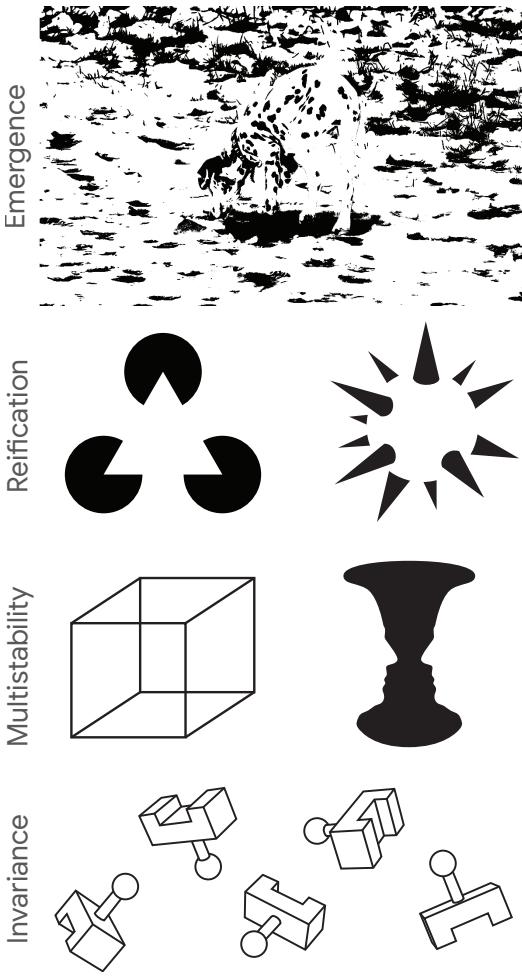
### 6.1 Gestalt

Gestalt Psychology describes many aspects of the subjective experience of perceptual organization (see [Wagemans et al., 2012a,b](#) for an overview). It is based on the observation that the perception of ‘wholes’ (or Gestalten<sup>21</sup>) can not be adequately described as a bottom-up agglomeration of more primitive percepts, but rather emerges in its entirety at once. Similarly, the perception of a Gestalt can fill in missing information, be invariant to transformations, and alternate discretely between multiple stable interpretations (see [Figure 22](#)). This holistic (as opposed to analytic; see [Section 6.2](#)) view of perception, was later summarized by Kurt Koffka as: “The whole *is other* than the sum of its parts” ([Koffka, 1935](#))<sup>22</sup>. The concept of a Gestalt closely resembles our notion of objects and Gestalt Psychology was arguably the first systematic investigation of human object perception (following the work by [Wertheimer, 1912](#)).

The best-known results of Gestalt research are their principles of perceptual grouping (also known as Gestalt Laws; see [Figure 21](#) for an overview). They describe which stimulus-cues influence the perceived grouping of a set of discrete elements ([Wertheimer, 1923](#); [Wagemans et al., 2012a](#)). They include among others: the law of proximity (closeby pieces tend to

21. “Gestalten” is plural of the German word “Gestalt” meaning “form” or “shape”.

22. Frequently misquoted as “The whole is *greater* than the sum of its parts”.



**Figure 22: Emergence:** At first encounter this image (a reproduction of the classic image from [Gregory 1970](#)) is perceived as an unstructured collection of black patches on white background. At some point perception shifts and suddenly reveals the image of a Dalmatian dog sniffing the ground. Perception of the whole arises at once, and not through hierarchically assembling of parts, such as legs, ears, etc.

**Reification:** Perception of a Gestalt carries information about its parts and leads to top-down “filling-in” of missing information. This is often demonstrated with the illusory contours of the *Kanizsa triangle* (left).

**Multistability:** Many scenes are ambiguous and afford multiple stable groupings. In such cases perception alternates periodically between different interpretations.

**Invariance:** Objects are recognized based on their overall shape invariant of: rotation, shift, scale, illumination, and many other factors.

be grouped), the law of similarity (similar pieces tend to be grouped), the law of closure (preference for closed contours), the law of symmetry (preference for symmetric objects) and the law of common fate (what moves together groups together). Several other Gestalt laws have been found over the years ([Alais et al., 1998](#); [Palmer, 1992](#); [Palmer and Rock, 1994](#)), including for other sensory modalities, such as audio ([Bregman, 1994](#)) and tactile ([Gallace and Spence, 2011](#)). Note that the laws of proximity and common fate can be seen as special cases of the law of similarity (with position and movement respectively being the compared attributes). In fact, it has been argued that the Gestalt Laws are all special cases of a single information-theoretic grouping principle ([Hatfield and Epstein, 1985](#)). Here the idea is that a ‘good’ Gestalt is one with a lot of internal redundancy ([Attneave, 1971](#)), and thus that the likelihood of a particular grouping is inversely proportional to the amount of information required to describe the Gestalt ([Hochberg and McAlister, 1953](#))<sup>23</sup>.

For our purposes, the existence of these general principles and their prevalence in multiple sensory domains is very interesting. It makes plausible the idea of a general segregation

23. There is disagreement about how to quantify information and the issue of simplicity versus likelihood has been debated extensively, though they might turn out to be identical ([Chater, 1996](#))

mechanism (e.g. based on modularity) that can generalize to novel objects and can help to steer the search for corresponding inductive biases. However, note that Gestalt Psychology has been criticized for its emphasis on subjective experience and the lack of successful physiological or mechanistic predictions (e.g. Ohlsson, 1984; Treisman and Gelade, 1980; but see Jäkel et al., 2016). Feature Integration Theory arose as a countermovement to provide an alternative, more mechanistic, account of the grouping process.

## 6.2 Feature Integration Theory

Feature Integration Theory (FIT) provides a model of human visual attention for perceiving objects (see Wolfe, 2020 for an overview). It is based on the idea that conscious object perception (i.e. as we experience it) is preceded by subconscious (mostly) bottom-up processing of visual information. FIT is motivated by a number of empirical findings, such as the different speeds at which humans are able to locate a visual target among a set of distractors (visual search). In this case, search is fast (subconscious and in parallel) if the target can be identified by a single characteristic feature (e.g. a particular orientation), which essentially causes it to *pop-out* (e.g. top panel in Figure 23a). In contrast, when the target is characterized by a *conjunction* of features, search becomes slow and requires serial attention (e.g. bottom panel in Figure 23a). Another important empirical finding occurs when attention is overloaded (or directed elsewhere), which sometimes causes humans to perceive *illusory conjunctions*: illusory objects that are the result of wrongly combining features from other objects (Treisman and Gelade, 1980; Figure 23c).

Feature Integration Theory distinguishes two stages of processing (see Figure 23b). First, a *pre-attentive stage* that registers features across the visual field (e.g. shape, color, size, etc.) automatically in parallel, and represents them in independent feature maps ('free-floating'). Then, at the *feature integration stage*, a 'spotlight of attention' is used to bind the features in these separate maps to form feature conjunctions in the form of objects (Kahneman et al., 1992). While initially objects are linked to specific locations as attention is focused on them, they may later be consolidated to form a more location invariant representation (Treisman and Zhang, 2006). Since its initial conception (Treisman, 1977; Treisman and Gelade, 1980), FIT has been refined and extended in various ways to account for new insights about human perception. There is now substantial evidence that the features of objects outside the focus of attention are more structured than initially assumed. For example, Humphrey and Goodale (1998) find that orientation and color are already represented jointly in the absence of attention (see also Vul and MacLeod (2006)). Similarly, Vul et al. (2019) find evidence for pre-attentive binding of color to parts based on the hierarchical (and geometric) structure of objects.

FIT has been a highly influential model of human visual attention and could serve as further inspiration for attention-based segregation. While FIT and Gestalt Psychology offer seemingly competing views of human perception, it has also been argued that these analytic and holistic views in fact complement each other (Prinzmetal, 1995). However, in either case, it is unclear how certain aspects of FIT should be implemented, such as top-down feedback to guide attention, especially in the context of non-visual domains (Spence and Frings, 2020).

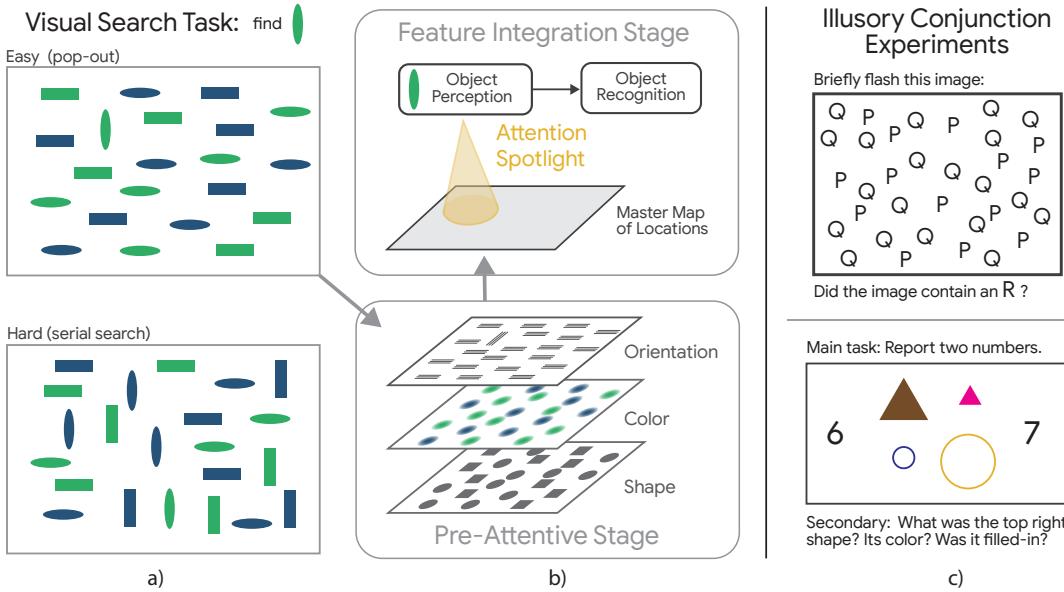


Figure 23: **Left:** Two examples of visual search tasks, an easy one where the target “pops-out” (top) and a hard one that requires serial search. **Middle:** Diagram of processing operations involved in the perception of objects according to FIT. **Right:** Two example tasks that have been used to demonstrate *illusory conjunctions*. Note, that the effect cannot be reproduced in print because it relies on showing the images very briefly.

### 6.3 The Binding Problem in Neuroscience

We have adapted the term binding problem from neuroscience, where it refers to a limitation of our understanding regarding information processing in the brain. In particular, its highly distributed nature raises the question of “[...] how the computations occurring simultaneously in spatially segregated processing areas are coordinated and bound together to give rise to coherent percepts and actions” — Singer (2007). For example, how is it that we typically do not wrongly mix the properties belonging to different objects, i.e. experience illusory conjunctions? The binding problem in neuroscience is thus concerned with understanding the mechanism(s) by which the brain addresses these challenges.

Several mechanisms have been proposed that range from static binding using conjunction cells (Ghose and Maunsell, 1999) to dynamic information routing through dedicated circuitry (Olshausen et al., 1993; Zylberberg et al., 2010) or attention using common location tags (Reynolds and Desimone, 1999; Robertson, 2005). A particularly promising hypothesis is the *temporal correlation hypothesis*, which holds that temporal synchrony of firing patterns is the mechanism responsible for binding (Milner, 1974; von der Malsburg, 1981). In this case, neurons whose activation encodes features of one object (e.g. color and shape) are expected to fire in synchrony (oscillating phase-locked), while neurons encoding features belonging to different objects would be out of phase with each other (see also Section 3.3.2). Other neurons are naturally capable of responding to this form of grouping since neuronal firing and synaptic learning (STDP; Caporale and Dan, 2008) are both sensitive to the relative timing of incoming activations (pre-synaptic spikes). Moreover, there is diverse experimental

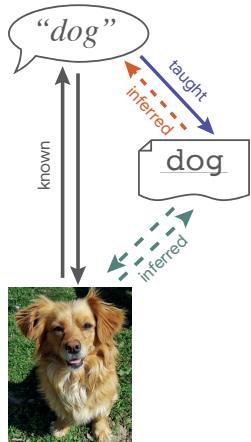


Figure 24: In an early experiment, Sidman (1971) examined a boy that could match spoken words to pictures and to name pictures (gray), but was unable to read. After being taught to match spoken words to written words (blue), he was then also able to read the written words aloud (red), and to match them to pictures (green). In this case, the dotted arrows represent relations that were never explicitly taught, and which were *derived* based on reflexivity (red) and transitivity (green) of the underlying equivalence relation (Sidman et al., 1989). Later, it was found that such derived relationships play an important role in systematically altering human behavior in response to feedback from the environment.

data in support of this interpretation relating synchronized oscillatory behavior of individual neurons to perceptual grouping (Usher and Donnelly, 1998; Tallon-Baudry and Bertrand, 1999), attention (Niebur et al., 2002), and sensory-motor integration (Pesce Ibarra, 2017; Engel and Fries, 2010; see also Uhlhaas et al., 2009 for an overview).

In general, the role of synchrony in neuronal binding is still controversial. For example, it has been debated whether synchrony is necessary (Merker, 2013; Riesenhuber and Poggio, 1999), fast enough (Ray and Maunsell, 2015; Palmigiano et al., 2017), and is capable of providing sufficient (temporal) resolution<sup>24</sup> for separating multiple different objects. Likely, the brain does not rely on a single mechanism for addressing the binding problem but on a combination of several. In either case, it is clear that temporal synchronization plays an important role in neural information processing, and perhaps one that is still unaddressed in current artificial neural networks.

#### 6.4 Relational Frame Theory

Relational Frame Theory (RFT; Hayes et al., 2001; Hughes and Barnes-Holmes, 2016) is a theory of behavioral psychology about *relating* (i.e. responding to one event in terms of another) and offers interesting insights about composing and systematic generalization in humans. RFT was originally conceived to explain "stimulus equivalence" (Sidman, 1971): The emergent behavior to respond to events and objects through a derived "sameness" relation that has not been explicitly taught or reinforced. For example, when taught a correspondence between spoken words and pictures, and between spoken words and written words, children were able to match written words and pictures (see Figure 24). In a similar experiment, Dymond and Barnes (1995) showed that subjects were able to use such derived equivalence relations to "correctly" respond to stimuli for which no explicit feedback was provided.

RFT is based in behaviorism, focusing on observable behavioral responses that can be altered through reinforcement or punishment (learned operants). It argues that relational

24. In this case, the temporal accuracy of synchronization directly relates to the capacity of working memory (Wilhelm et al., 2013). The more objects need to be represented simultaneously, the more difficult it is to prevent cross-talk from corrupting and destabilizing individual representation, and such gradual decay has indeed been observed in Alvarez and Franconeri (2007).

responding<sup>25</sup> is a learned operant behavior, which can be acquired through repeated exposure to tasks that require responding to a particular relation (to receive positive feedback) but that varies across stimuli and contexts. Relational responding can be subdivided into different *relational frames* (see also Section 5.1.3), which each focus on a particular kind of relationship and differ in terms of three key properties: *mutual entailment*, *combinatorial entailment*, and *transformation of stimulus functions*. For example, the stimulus equivalence that was observed in Figure 24 corresponds to a particular relational frame with symmetry as mutual entailment and transitivity as combinatorial entailment. In this case, ‘transformation of stimulus function’ implies that when the reward associated with an object or event changes, this also alters the expected reward of other related events or objects in the same manner. Other examples include the relational frames of ‘opposition (e.g. opposite to)’, ‘comparison’ (e.g. larger than), ‘hierarchy’ (e.g. part of), ‘temporal order’ (e.g. after), or ‘condition’ (e.g. if then). It is easy to see how a vast number of possible relational structures can be constructed in this way, of which only very few are relevant in any given situation. RFT argues that people use (bottom-up) *contextual cues* from the environment to infer which relations when to apply.

Given the immediate relevance of RFT to systematic generalization and composition, it is surprisingly absent from the machine learning literature. This is likely in part due to the relative unpopularity of behaviorism compared to cognitive psychology. However, another reason may be due to the controversy that surrounds certain aspects of RFT, such as the clarity of the involved concepts and its novelty with regards to previous accounts of stimulus equivalence (Gross and Fox, 2009). Nonetheless, we find that RFT offers a useful conceptual framework for the problem of composition, and indeed it has helped shape our understanding of relational reasoning. Going forward, we would like to emphasize the value of RFT as a source of experimental designs to isolate and evaluate relational reasoning capabilities in neural networks.

## 6.5 Compositionality in Linguistics

Like many others in the field, we have used the term compositionality without giving a proper definition. Related terms such as systematicity, systematic generalization, and combinatorial generalization, unfortunately, do not provide a good alternative either. A good starting point for a definition may therefore be the so-called *principle of compositionality* from the field of linguistics:

“The meaning of a complex expression is determined by its structure and the meanings of its constituents.” — Szabó (2017)

Apart from its intuitive appeal, the main reason for its widespread adoption is the lack of a convincing alternative. However, there remains considerable disagreement about the exact phrasing and many interpretations of the principle exist (Szabó, 2017).

---

25. RFT distinguishes between two types of relational responding: Non-Arbitrarily Applicable Relational Responding (NAARR), which is only concerned with relations among physical attributes (e.g. choosing the *larger* among multiple objects), and the more general Arbitrarily Applicable Relational Responding (AARR) that allows for arbitrary relations between stimuli (or events). While NAARR is also encountered in animals, AARR has thus far only been observed in humans.

There are three main arguments for this notion of compositionality, namely *productivity*, *systematicity*, and *efficiency* of language. Productivity refers to the capacity of language to “make infinite use of finite means” (von Humboldt, 1999), i.e. the ability to form and understand a theoretically unbounded number of entirely novel sentences given only limited vocabulary and training. Systematicity is the observation that “the ability to produce/understand some sentences is intrinsically connected to the ability to produce/understand certain others” (Fodor and Pylyshyn, 1988). For example, anyone who understands “brown dog” and “black cat” also understands “brown cat”. Finally, the fact that we are able to communicate in real-time, puts clear bounds on the computational complexity of interpreting spoken language (Szabó, 2017). The principle of compositionality is thus an inference to the best explanation because it is difficult to imagine language being productive, systematic, and computationally efficient without its semantics being somehow compositional in the above sense.

Critique of the principle of compositionality, interestingly, ranges from it being too broad to it being too narrow. On the one hand, Zadrozny (1994) demonstrates how a function can be constructed that maps arbitrary meaning to any expression without violating compositionality. This suggests that the principle is formally vacuous unless the class of admissible functions is somehow restricted to exclude such a construction. On the other extreme, many have found violations of the principle in everyday language. Indeed, counterexamples such as ambiguities (“We saw her duck.”), references (“this dog”), and irony (“objectively the best example”) require context and thus contradict the principle. Similarly, idioms (“break the ice”) provide examples of obvious exceptions where the meaning differs substantially from a naive composition of the parts. However, few consider these problems severe enough to abandon the principle of compositionality entirely, and indeed most linguists have come to accept it as a guiding principle for developing syntactic and semantic theories. Though it was originally conceived for language, many believe that the principle of compositionality applies equally (or even more so) to mental representations (Butler, 1995; Fodor, 1975). A similar belief also underlies the interest in compositionality for understanding and encouraging productivity, systematicity, and efficient inference in neural networks (Santoro et al., 2018b; Hupkes et al., 2020; Andreas, 2019).

## 7. Discussion

The ultimate motivation of this work is to address the shortcomings of neural networks at human-level generalization. To this end, we have developed a conceptual framework centered around compositionality and the binding problem. Our analysis identifies the binding problem as the primary cause for these shortcomings, and thereby paves the way for a single unified solution. It rests on several (implicit) assumptions regarding the nature and importance of objects and the learning capabilities of neural networks. In the following, we explicate several of these assumptions and use them to contrast with other conceptual frameworks aimed at addressing (certain aspects of) human-level generalization.

One of the main assumptions behind our work is that objects are key to compositionality and that the latter plays a fundamental role in generalizing more systematically. This perspective has a long history in connectionism that goes back to at least Fodor and Pylyshyn (1988); Marcus (2003) and has been repeatedly emphasized (e.g. Smolensky, 1990; Bader and Hitzler, 2005), especially in recent years (e.g. Lake et al., 2017; Battaglia et al., 2018; Hamrick, 2019; Garnelo and Shanahan, 2019). However, our perspective stands out in that we focus on integrating symbolic reasoning and sensory grounding, which requires adopting a very broad notion of objects that spans all levels of abstraction. Importantly, we assume that objects at any level of abstraction are essentially the result of decomposing a given problem into modular building blocks, and thus share the same underlying computational mechanisms. It is our view that this broad notion of objects is necessary to accommodate the generality of human reasoning from concrete and physical to abstract and metaphorical.

Throughout this paper, we have assumed that learning objects in an unsupervised way is both feasible, and can be integrated directly into neural networks. Further, we have argued that unsupervised learning is, in fact, indispensable, due to the required scope and flexibility of objects, which renders adequate supervision or engineering infeasible. However, as we have seen (and discuss further below), evidence indicates that object representations are unlikely to emerge naturally simply by scaling current neural networks in terms of model size or by providing additional data. Here we have proposed to address this problem by incorporating a small set of inductive biases to enable neural networks to process information more symbolically, while also preserving the crucial benefits of end-to-end learning (Sutton, 2019).

Closely related to the mental framework proposed here is that of Lake et al. (2017), which is similarly concerned with addressing human-level generalization. They too emphasize the importance of (physical) objects, compositionality, and dynamic model building, although in their view these are only three instances of so-called ‘core ingredients’ necessary for realizing human intelligence. Other ingredients include an intuitive understanding of psychology as a form of “start-up software”, learning to learn, causality, and ingredients focusing on the speed of human comprehension. Hence, Lake et al. (2017) advocate the use of specialized inductive biases inspired by cognitive psychology, and using neural networks as a *means* for implementing fast inference within the context of larger structured models. In contrast, we argue that it is more fruitful to enable neural networks to directly implement structured models. This enables us to tackle a single shared underlying problem (the problem of dynamic binding) and, as much as possible, let learning account for the remaining, domain-specific, aspects of human cognition (e.g. psychology, physics, causality). Note that we do not wish

to argue against incorporating specialized inductive biases (which may still be beneficial), but rather advocate that learning should take priority whenever possible. Compared to Lake et al. (2017) our focus on integrating high-level reasoning with low-level perception in neural networks puts a lot more emphasis on symbol grounding and the associated problem of segregation. This is also reflected by our emphasis on end-to-end learning, whereas Lake et al. (2017) appear to argue for separating neural and symbolic information content, somewhat akin to hybrid approaches (Bader and Hitzler, 2005).

Our framework also relates to several other areas of machine learning research that aim towards human-level generalization. However, they center around composition and have mostly neglected the problem of segregation (and representation). For example, the field of causality is concerned with inferring and reasoning about structural causal models, which offer a particular kind of compositionality that is assumed to be essential to human-level generalization (Pearl, 2009; Peters et al., 2017). Using our terminology, structural causal models can be viewed as a specific set of relational frames composed of ‘independent causal mechanisms’ that define a structure, which can be used to systematically reason about novel situations (e.g. for interventions or counterfactuals). As was recently noted by Schölkopf (2019), traditional work in causality assumes given knowledge about the associated causal variables (e.g. objects), and the problem of discovering them (i.e. segregation) has mostly been neglected. In a similar vein, recent work on graph neural networks seeks to achieve systematic generalization by focusing on relations between *given* entities (Battaglia et al., 2018). Alternatively, Bengio (2019) argues for the importance of a low-dimensional ‘conscious state’ (working memory) composed of largely independent units of abstraction that can be selected via attention (perhaps reminiscent of Schmidhuber, 1992b). He relates the unconscious elements from which the conscious state is constructed to a more symbolic knowledge representation, and emphasizes their importance for systematic generalization. However, here too, it remains unclear how such elements should be obtained and represented in neural networks.

Finally, we acknowledge the promising results that recent large-scale language models have produced in terms of generalization and their (acquired) ability for few-shot learning (Radford et al., 2019; Brown et al., 2020). They are evidence for the possibility that human-level generalization may be achieved by scaling existing approaches using orders of magnitude more data and network parameters. However, we remain pessimistic as to whether similar results can be obtained on less structured domains, such as when learning from raw perceptual data. As we have argued throughout this work, the fundamental lack of a suitable mechanism for dynamic information binding precludes the emergence of the modular building blocks needed for acquiring a compositional understanding of the world.

## 8. Conclusion

Humans understand the world in terms of abstract entities, like objects, whose underlying compositionality allows us to generalize far beyond our direct experiences. At present, neural networks are unable to generalize in the same way. In this paper, we have argued that this limitation is largely due to the binding problem, which impairs the ability of neural networks to effectively incorporate symbol-like *object representations*. To address this issue, we have proposed a functional division of the binding problem that focuses on three different aspects: The ability to separately represent multiple object representations in a common format, without interference between them (representation problem); The process of forming grounded object representations that are modular from raw unstructured inputs (segregation problem); And finally, the capacity to dynamically relate and compose these object representations to build structured models for inference, prediction, and behavior (composition problem). Based on this division, we have offered a conceptual framework for addressing the lack of symbolic reasoning capabilities in neural networks that is believed to be the root cause for their lack of systematic generalization. Indeed, the importance of symbolic reasoning has been emphasized before (Fodor and Pylyshyn, 1988) and served as a starting point for several related perspectives (Marcus, 2003; Lake et al., 2017). Here we have provided a more in-depth analysis of the challenges, requirements, and corresponding inductive biases required for symbol manipulation to emerge naturally in neural networks.

Based on our discussion, we wish to highlight several important open problems for future research in three different areas.

First is the process of segregation, which is of foundational importance and requires a proper treatment of the dynamic and hierarchical nature of objects. In particular, we believe that the ability to segregate must therefore largely be learned in an unsupervised fashion, which is a major open problem that is often overlooked in the current literature. For a new situation, the most useful decomposition in terms of objects (and the associated level of abstraction) depends not only on the task, but also on the abstractions, relations, and general problem-solving capabilities available to the entire system. Therefore, another open problem is to integrate segregation, representation, and composition into a single system in a way that resolves these dependencies (through top-down feedback). Existing attempts fail to accommodate these interactions, e.g. because they rely on pre-trained vision modules (Mao and Gan, 2019) or overly specialized domain-specific components (de Avila Belbute-Peres et al., 2018). Addressing these open problems may pave the way for an integrated system that can learn to dynamically construct structured models for prediction, inference, and behavior in a way that generalizes similarly to humans.

Secondly, to facilitate progress on the binding problem, we require corresponding benchmarks and metrics that allow for meaningful comparisons. Current benchmarks fall short in the sense that they do not bridge the gap between simplistic ‘toy’ datasets and the complexity of real-world sensory information, or lack the appropriate meta-data required to support evaluation (such as object-level annotations). The latter is particularly important since standard approaches to measuring properties such as systematic generalization or disentanglement are supervised and require information about ‘ground truth’ objects or factors. However, this reliance on ground truth data seems problematic in real-world settings more generally, i.e. due to the task- and context-dependent nature of objects and the amount of manual

labor involved. This should motivate research on alternative ‘unsupervised metrics’ for these purposes, e.g. analogous to the FID score for the perceptual quality of images (Heusel et al., 2017). The design of benchmarks and metrics is hindered by a lack of agreed-upon definitions for behaviors like systematic generalization, combinatorial generalization, or compositionality. Going forward, it is therefore critical to develop a shared vocabulary of well-defined and measurable generalization patterns that can be explicitly characterized in terms of the type and amount of available information. Recent attempts at quantifying systematic generalization that distinguish between interpolation and extrapolation (Santoro et al., 2018b) or the categorization developed by Hupkes et al. (2020) provide a promising step in this direction.

Finally, we wish to highlight several other interesting research directions that are also important for human-level generalization but go beyond the scope of this survey. Concerning the binding problem, we focused primarily on encoding information about objects in working memory, although similar problems arise in the context of long-term memory. We speculate that several of the same insights can be applied here, e.g. memory recall as a type of segregation, or the need for a separation between relations and objects. However, for other challenges, such as the problem of representing information in a scalable way (despite a constantly evolving representational format), the connection is less clear. Another interesting direction is concerned with the arising and grounding of more abstract concepts like “mammals”, “capitalism” or “a transaction”. Although abstract objects may be more difficult to obtain, since they are further removed from sensory reality, it is precisely because of this gap that they are capable of participating in a wider range of situations. Indeed, this research direction is highly relevant to the broader problem of grounding language, which is concerned with abstract concepts in their most general form. In this context, it is interesting to note that most (if not all) abstract concepts seem to be grounded in basic physical metaphors (Lakoff and Johnson, 2008). Finally, a comprehensive treatment of causal reasoning likely goes beyond composition and should include an explicit treatment of interventions and the ability to reason about hypothetical or unobserved scenarios (counterfactuals). This is especially relevant due to the connection between systematic generalization and the increased robustness when considering so-called independent causal mechanisms (Peters et al., 2017). If a suitable causal relational frame can be learned, then this may allow the problem of planning to be phrased as connecting a current state and an imagined goal state, by means of combinatorial entailment.

We hope that this survey may serve as an inspiration and a guide for future work towards achieving human-level generalization in neural networks and that it may spark fruitful discussions that bridge the gap between related fields.

## Acknowledgements

We wish to thank Pina Merkert and Mike Mozer in particular, for their constructive feedback and support. We also wish to thank Sungjin Anh, Boyan Beronov, Paul Bertner, Matt Botvinick, Alexey Dosovitskiy, Sylvain Gelly, Leslie Kaelbling, Thomas Kipf, Alexander Lerchner, Paulo Rauber, Aleksandar Stanić, and Harri Valpola. Finally, we wish to thank many other colleagues and friends for useful discussions about binding throughout the last years. This research was supported by Swiss National Science (SNF) grant 200021\_165675/1 (successor project: no: 200021\_192356) and EU project “INPUT” (H2020-ICT-2015 grant no. 687795).

## References

- David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive science*, 9:147–169, 1985. ISSN 0364-0213.
- Linda Acredolo and Susan Goodwyn. Symbolic gesturing in normal infants. *Child development*, pages 450–466, 1988.
- David Alais, Randolph Blake, and Sang-Hun Lee. Visual features that vary together over time group together over space. *Nature neuroscience*, 1(2):160–164, 1998.
- Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. What is an object? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–80, 2010.
- Luis B. Almeida. A learning rule for asynchronous perceptrons with feedback in a combinatorial environment. In *Proceedings of the 1st First International Conference on Neural Networks*, volume 2, pages 609–618. ci.nii.ac.jp, 1987.
- George A. Alvarez and Steven L. Franconeri. How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of vision*, 7(13):14–14, 2007.
- Marco Ancona, Enea Ceolini, Cengiz Öztïreli, and Markus Gross. A unified view of gradient-based attribution methods for deep neural networks. In *Neural Information Processing Systems (NeurIPS) Workshop on Interpreting, Explaining and Visualizing Deep Learning-Now What?*, 2017.
- Jacob Andreas. Measuring compositionality in representation learning. In *International Conference on Learning Representations*, 2019.
- Relja Arandjelović, Andrew Zisserman, Yawei Luo, Changhui Hu, Xiaobo Lu, and Xin Yu. Object discovery with a copy-pasting GAN. *arXiv preprint arXiv:1905.11369*, 2019.
- Pablo Arbeláez, Michael Maire, Charless C. Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2011. ISSN 0162-8828. doi: 10.1109/TPAMI.2010.161.
- Fred Attneave. Multistability in perception. *Scientific American*, 225(6):62–71, 1971. ISSN 0036-8733(Print). doi: 10.1038/scientificamerican1271-62.
- James Atwood and Don Towsley. Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1993–2001, 2016.
- Yuval Atzmon, Jonathan Berant, Vahid Kezami, Amir Globerson, and Gal Chechik. Learning to generalize to new compositions in image understanding. *arXiv preprint arXiv:1608.07639*, 2016.

- Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014.
- Jimmy Ba, Geoffrey E. Hinton, Volodymyr Mnih, Joel Z. Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. In *Advances in Neural Information Processing Systems 29*, pages 4331–4339, 2016.
- Sebastian Bader and Pascal Hitzler. Dimensions of neural-symbolic integration — A structured survey. *arXiv preprint arXiv:0511042*, 2005.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5221–5229, 2017.
- Renee Baillargeon, Elizabeth S. Spelke, and Stanley Wasserman. Object permanence in five-month-old infants. *Cognition*, 20(3):191–208, 1985.
- Valentina Bambini, Cristiano Chesi, and Andrea Moro. A conversation with Noam Chomsky: New insights on old foundations. *Phenomenology and Mind*, (3):166–178, 2012.
- Horace B. Barlow, Tej P. Kaushal, and Graeme J. Mitchison. Finding minimum entropy codes. *Neural Computation*, 1(3):412–423, 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.3.412.
- Yvonne Barnes-Holmes, Dermot Barnes-Holmes, Paul M. Smeets, Paul Strand, and Patrick Friman. Establishing relational responding in accordance with more-than and less-than as generalized operant behavior in young children. *International Journal of Psychology and Psychological Therapy*, 4:531–558, 2004.
- Peter W. Battaglia, Razvan Pascanu, Matthew Lai, and Danilo Jimenez Rezende. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems*, pages 4502–4510, 2016.
- Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, H. Francis Song, Andrew Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matthew Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.
- Clare Batty. Olfactory objects. *Perception and its modalities*, pages 222–224, 2014.
- Daniel Bear, Chaofei Fan, Damian Mrowca, Yunzhu Li, Seth Alter, Aran Nayebi, Jeremy Schwartz, Li F. Fei-Fei, Jiajun Wu, Josh Tenenbaum, and Daniel L. Yamins. Learning physical graph representations from visual scenes. *Advances in Neural Information Processing Systems*, 33, 2020.
- Suzanna Becker and Geoffrey E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.
- Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2019.
- Yoshua Bengio, Aaron Courville, and Pierre Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. ISSN 0162-8828.

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, and Chris Hesse. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Konstantinos Blekas, Aristidis Likas, Nikolaos P. Galatsanos, and Isaac E. Lagaris. A spatially constrained mixture model for image segmentation. *IEEE transactions on Neural Networks*, 16(2):494–498, 2005. ISSN 1045-9227. doi: 10.1109/TNN.2004.841773.

Daniel G. Bobrow. Natural language input for a computer problem solving system. 1964.

Jeffrey S. Bowers, Ivan I. Vankov, Markus F. Damian, and Colin J. Davis. Neural networks learn highly selective representations in order to overcome the superposition catastrophe. *Psychological Review*, 121(2):248–261, 2014. ISSN 1939-1471, 0033-295X. doi: 10.1037/a0035943.

Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT press, 1994.

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. In *International Conference on Learning Representations*, 2019.

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Neil D. B. Bruce and John K. Tsotsos. Saliency based on information maximization. In Yair Weiss, Bernhard Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 155–162, 2006.

Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matthew Botvinick, and Alexander Lerchner. MONet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

Nicholas J. Butko and Javier R. Movellan. Optimal scanning for faster object detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2751–2758, 2009. doi: 10.1109/CVPR.2009.5206540.

Keith Butler. Content, context, and compositionality. *Mind & Language*, 10(1-2):3–24, 1995. ISSN 1468-0017. doi: 10.1111/j.1468-0017.1995.tb00003.x.

Murray Campbell, A. Joseph Hoane Jr, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1-2):57–83, 2002.

Brian Cantwell-Smith. *On the Origin of Objects*. A Bradford Book. MIT Press, Cambridge, Mass., 1st paperback ed edition, 1998. ISBN 978-0-262-69209-0.

Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, Deva Ramanan, and Thomas S. Huang. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *IEEE International Conference on Computer Vision*, pages 2956–2964, 2015.

Natalia Caporale and Yang Dan. Spike timing-dependent plasticity: A Hebbian learning rule. *Annual Review of Neuroscience*, 31:25–46, 2008.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020.

Michael B. Chang, Tomer Ullman, Antonio Torralba, and Joshua B. Tenenbaum. A compositional object-based approach to learning physical dynamics. In *International Conference on Learning Representations*, 2017.

Nick Chater. Reconciling simplicity and likelihood principles in perceptual organization. *Psychological review*, 103(3):566–581, 1996. doi: 10.1037/0033-295X.103.3.566.

Chang Chen, Fei Deng, and Sungjin Ahn. Object-centric representation and rendering of 3D scenes. *arXiv preprint arXiv:2006.06130*, 2020.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2016.

E. Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

Dan C. Cireşan, Ueli Meier, Jonathan Masci, Luca M. Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *International Joint Conference on Artificial Intelligence*, pages 1237–1242, 2011.

Dan C. Cireşan, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural Networks*, 32:333–338, 2012. ISSN 0893-6080. doi: 10.1016/j.neunet.2012.02.023.

Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3412–3420, 2019.

Róbert Csordás and Jürgen Schmidhuber. Improved addressing in the differentiable neural computer. In *International Conference on Learning Representations*, 2019.

Róbert Csordás, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Are neural nets modular? Inspecting functional modularity through differentiable weight masks. *arxiv preprint arXiv:2010.02066*, 2020.

Rita Cucchiara, Costantino Grana, Massimo Piccardi, and Andrea Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE transactions on pattern analysis and machine intelligence*, 25(10):1337–1342, 2003.

- Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016.
- Sreerupa Das and Michael C Mozer. A unified gradient-descent/clustering architecture for finite state machine induction. *Advances in Neural Information Processing Systems*, 6:19–26, 1993.
- Sreerupa Das, C. Lee Giles, and Guo-Zheng Sun. Learning context-free grammars: Capabilities and limitations of a recurrent neural network with an external stack memory. In *Proceedings of The Fourteenth Annual Conference of Cognitive Science Society. Indiana University*, page 14, 1992.
- Guy Davidson and Brenden M. Lake. Investigating simple object representations in model-free deep reinforcement learning. In *Annual Meeting of the Cognitive Science Society*, 2020.
- Filipe de Avila Belbute-Peres, Kevin Smith, Kelsey Allen, Josh Tenenbaum, and J. Zico Kolter. End-to-end differentiable physics for learning and control. In *Advances in Neural Information Processing Systems*, pages 7178–7189, 2018.
- Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. In *International Conference on Learning Representations*, 2019.
- Luca Del Pero, Joshua Bowdish, Daniel Fried, Bonnie Kermgard, Emily Hartley, and Kobus Barnard. Bayesian geometric modeling of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2719–2726. ieeexplore.ieee.org, 2012. doi: 10.1109/CVPR.2012.6247994.
- Luca Del Pero, Joshua Bowdish, Bonnie Kermgard, Emily Hartley, and Kobus Barnard. Understanding Bayesian rooms using composite 3d object models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 153–160. cv-foundation.org, 2013.
- Zhiwei Deng, Megha Nawhal, Lili Meng, and Greg Mori. Continuous graph flow. *arXiv preprint arXiv:1908.02436*, 2019.
- Barry J. Devereux, Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. The centre for speech, language and the brain (CSLB) concept property norms. *Behavior Research Methods*, 46(4): 1119–1127, 2014. ISSN 1554-3528. doi: 10.3758/s13428-013-0420-4.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.
- Leonidas A. A. Doumas, John E. Hummel, and Catherine M. Sandhofer. A theory of the discovery and predication of relational concepts. *Psychological Review*, 115(1):1–43, 2008. ISSN 1939-1471, 0033-295X. doi: 10.1037/0033-295X.115.1.1.
- Leonidas A. A. Doumas, Guillermo Puebla, Andrea E. Martin, and John E. Hummel. Relation learning in a neurocomputational architecture supports cross-domain transfer. *arXiv preprint arXiv:1910.05065*, 2019.
- Simon Dymond and Dermot Barnes. A transformation of self-discrimination response functions in accordance with the arbitrarily applicable relations of sameness, more than, and less than. *Journal of the Experimental Analysis of Behavior*, 64(2):163–184, 1995.
- Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

- Sebastien Ehrhardt, Oliver Groth, Aron Monszpart, Martin Engelcke, Ingmar Posner, Niloy Mitra, and Andrea Vedaldi. RELATE: Physically plausible multi-object scene synthesis using structured latent spaces. *Advances in Neural Information Processing Systems*, 33, 2020.
- Ian Endres and Derek Hoiem. Category independent object proposals. In *European Conference on Computer Vision*, pages 575–588, 2010. doi: 10.1007/978-3-642-15555-0\_42.
- Andreas K. Engel and Pascal Fries. Beta-band oscillations—signalling the status quo? *Current opinion in neurobiology*, 20(2):156–165, 2010.
- Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations. In *International Conference on Learning Representations*, 2019.
- S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Koray Kavukcuoglu, and Geoffrey E. Hinton. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances In Neural Information Processing Systems*, pages 3225–3233, 2016.
- Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, Narayanaswamy Siddharth, Brooks Paige, Dana H. Brooks, Jennifer Dy, and Jan-Willem van de Meent. Structured disentangled representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019.
- Jerome Feldman. The neural binding problem(s). *Cognitive neurodynamics*, 7(1):1–11, 2013. ISSN 1871-4080. doi: 10.1007/s11571-012-9219-8.
- Santiago Fernández, Alex Graves, and Jürgen Schmidhuber. Sequence labelling in structured domains with hierarchical recurrent neural networks. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007.
- François Fleuret, Ting Li, Charles Dubout, Emma K. Wampler, Steven Yantis, and Donald Geman. Comparing machines and humans on a visual categorization test. *Proceedings of the National Academy of Sciences*, 108(43):17621–17625, 2011.
- Jerry A. Fodor. *The Language of Thought*, volume 5. Harvard university press, 1975.
- Jerry A. Fodor and Zeno W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988. ISSN 0010-0277.
- Nir Friedman and Stuart Russell. Image segmentation in video sequences: A probabilistic approach. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 175–181, San Francisco, CA, USA, 1997.
- Fabian B. Fuchs, Adam R. Kosiorek, Li Sun, Oiwi Parker Jones, and Ingmar Posner. End-to-end recurrent multi-object tracking and trajectory prediction with relational reasoning. *arXiv preprint arXiv:1907.12887*, 2019.
- Keisuke Fukuda, Edward Awh, and Edward K. Vogel. Discrete capacity limits in visual working memory. *Current opinion in neurobiology*, 20(2):177–182, 2010. ISSN 0959-4388. doi: 10.1016/j.conb.2010.03.005.
- Alberto Gallace and Charles Spence. To what extent do Gestalt grouping principles influence tactile perception? *Psychological bulletin*, 137(4):538, 2011.
- Shani Gamrian and Yoav Goldberg. Transfer learning for related reinforcement learning tasks via image-to-image translation. In *International Conference on Machine Learning*, pages 2063–2072, 2019.

- Dashan Gao and Nuno Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. In Lawrence K. Saul, Yair Weiss, and Leon Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17, pages 481–488. MIT Press, 2005.
- Marta Garnelo and Murray Shanahan. Reconciling deep learning with symbolic artificial intelligence: Representing objects and relations. *Current Opinion in Behavioral Sciences*, 29:17–23, 2019. ISSN 2352-1546. doi: 10.1016/j.cobeha.2018.12.010.
- Ross W. Gayler. Multiplicative binding, representation operators & analogy. *Cogprints*, 1998.
- Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.
- Dedre Gentner and Mary Jo Rattermann. Language and the career of similarity. *Perspectives on language and thought: Interrelations in development*, 225, 1991.
- Geoffrey M. Ghose and John Maunsell. Specialized representations in visual cortex: A role for binding? *Neuron*, 24(1):79–85, 1999. ISSN 0896-6273. doi: 10.1016/S0896-6273(00)80823-5.
- C. Lee Giles, Guo-Zheng Sun, Hsing-Hen Chen, Yee-Chun Lee, and Dong Chen. Higher order recurrent networks and grammatical inference. In *Advances in Neural Information Processing Systems*, pages 380–387, 1990.
- Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, volume 70, pages 1263–1272, International Convention Centre, Sydney, Australia, 2017.
- Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587. cv-foundation.org, 2014.
- Anand Gopalakrishnan, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Unsupervised object keypoint learning using local spatial predictability. *arXiv preprint arXiv:2011.12930*, 2020.
- Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *Proceedings of the IEEE International Joint Conference on Neural Networks.*, volume 2, pages 729–734 vol. 2, 2005. doi: 10.1109/IJCNN.2005.1555942.
- Anirudh Goyal, Alex Lamb, Jordan Hoffmann, Shagun Sodhani, Sergey Levine, Yoshua Bengio, and Bernhard Schölkopf. Recurrent independent mechanisms. *arXiv preprint arXiv:1909.10893*, 2019.
- Alex Graves, Greg Wayne, and Ivo Danihelka. Neural Turing machines. *arXiv preprint arXiv:1410.5401*, 2014.
- Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià Puigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, 2016. ISSN 0028-0836. doi: 10.1038/nature20101.
- Edwin James Green. A theory of perceptual objects. *Philosophy and Phenomenological Research*, 106:7345, 2018. ISSN 0031-8205. doi: 10.1111/phpr.12521.

- Klaus Greff, Antti Rasmus, Mathias Berglund, Tele Hotloo Hao, Harri Valpola, and Jürgen Schmidhuber. Tagger: Deep unsupervised perceptual grouping. In *Advances in Neural Information Processing Systems*, pages 4484–4492, 2016.
- Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. In *Advances in Neural Information Processing Systems*, pages 6694–6704, 2017.
- Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nicholas Watters, Christopher P. Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, 2019.
- Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. DRAW: A recurrent neural network for image generation. In *International Conference on Machine Learning*, pages 1462–1471, Lille, France, 2015.
- Richard L. Gregory. *The Intelligent Eye*. 1970.
- Amy C. Gross and Eric J. Fox. Relational frame theory: An overview of the controversy. *The Analysis of verbal behavior*, 25(1):87–98, 2009.
- Jose A. Guerrero-Colón, Eero P. Simoncelli, and Javier Portilla. Image denoising using mixtures of Gaussian scale mixtures. In *IEEE International Conference on Image Processing*, pages 565–568, 2008.
- Nick Haber, Damian Mrowca, Stephanie Wang, Li F. Fei-Fei, and Daniel L. Yamins. Learning to play with intrinsically-motivated, self-aware agents. In *Advances in Neural Information Processing Systems*, pages 8388–8399, 2018.
- Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742, 2006.
- Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1024–1034, 2017.
- Jessica B. Hamrick. Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, 29:8–16, 2019. ISSN 2352-1546. doi: 10.1016/j.cobeha.2018.12.011.
- Jessica B. Hamrick, Kelsey R. Allen, Victor Bapst, Tina Zhu, Kevin R. McKee, Joshua B. Tenenbaum, and Peter W. Battaglia. Relational inductive bias for physical construction in humans and machines. In *Annual Meeting of the Cognitive Science Society*, 2018.
- Stevan Harnad. The symbol grounding problem. *Physica D*, 42(1):335–346, 1990. ISSN 0167-2789. doi: 10.1016/0167-2789(90)90087-6.
- Hedi Harzallah, Frédéric Jurie, and Cordelia Schmid. Combining efficient object localization and image classification. In *IEEE International Conference on Computer Vision*, pages 237–244, 2009. doi: 10.1109/ICCV.2009.5459257.
- Gary Hatfield and William Epstein. The status of the minimum principle in the theoretical analysis of visual perception. *Psychological Bulletin*, 97(2):155–186, 1985. ISSN 0033-2909.
- Steven C. Hayes, Dermot Barnes-Holmes, and Bryan Roche, editors. *Relational Frame Theory: A Post-Skinnerian Account of Human Language and Cognition*. Springer US, 2001. ISBN 978-0-306-46600-7. doi: 10.1007/b108413.

- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.
- Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.
- Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. Tracking the world state with recurrent entity networks. In *International Conference on Learning Representations*, 2017.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.
- Irina Higgins, Loic Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017a.
- Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. DARLA: Improving zero-shot transfer in reinforcement learning. In *International Conference on Machine Learning*, 2017b.
- Irina Higgins, David Amos, David Pfau, Sébastien Racanière, Loic Matthey, Danilo Jimenez Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.
- Felix Hill, Adam Santoro, David Barrett, Ari Morcos, and Timothy Lillicrap. Learning to Make Analogies by Contrasting Abstract Relational Structure. In *International Conference on Learning Representations*, 2019.
- Felix Hill, Andrew Lampinen, Rosalia Schneider, Stephen Clark, Matthew Botvinick, James L. McClelland, and Adam Santoro. Environmental drivers of systematicity and generalization in a situated agent. In *International Conference on Learning Representations*, 2020.
- Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, and Tara N. Sainath. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.
- Geoffrey E. Hinton. Distributed representations. Technical report, 1984.
- Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-21735-7\_6.
- Geoffrey E. Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with EM routing. In *International Conference on Learning Representations*, 2018.
- Julian Hochberg and Edward McAlister. A quantitative approach to figural "goodness". *Journal of Experimental Psychology*, 46(5):361–364, 1953. ISSN 0022-1015(Print). doi: 10.1037/h0055809.
- Derek Hoiem, Alexei A. Efros, and Martial Hebert. Recovering occlusion boundaries from an image. *International Journal of Computer Vision*, 91(3):328–346, 2011. ISSN 0920-5691. doi: 10.1007/s11263-010-0400-4.

- John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982. ISSN 0027-8424.
- Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, pages 689–696, 2009.
- Drew A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*, 2018.
- Sean Hughes and Dermot Barnes-Holmes. Relational frame theory: The basic account. *The Wiley Handbook of Contextual Behavioral Science*, page 129, 2016.
- John E. Hummel and Keith J. Holyoak. Distributing structure over time. *Behavioral and Brain Sciences*, 16(3):464–464, 1993. ISSN 1469-1825, 0140-525X. doi: 10.1017/S0140525X00031083.
- John E. Hummel, Keith J. Holyoak, Collin Green, Leonidas A. A. Doumas, Derek Devnich, Aniket Kittur, and Donald J. Kalar. A solution to the binding problem for compositional connectionism. In *Compositional Connectionism in Cognitive Science: Papers from the AAAI Fall Symposium*, Ed. SD Levy & R. Gayler, pages 31–34, 2004.
- G. Keith Humphrey and Melvyn A. Goodale. Probing unconscious visual processing with the McCollough effect. *Consciousness and cognition*, 7(3):494–519, 1998.
- Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, (67), 2020.
- Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.
- Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H. Adelson. Crisp boundary detection using pointwise mutual information. In *European Conference on Computer Vision*, pages 799–814, 2014. doi: 10.1007/978-3-319-10578-9\_52.
- Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H. Adelson. Learning visual groups from co-occurrences in space and time. *arXiv preprint arXiv:1511.06811*, 2015.
- Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998. ISSN 0162-8828. doi: 10.1109/34.730558.
- Michael Iuzzolino, Yoram Singer, and Michael C. Mozer. Convolutional bipartite attractor networks. *arXiv preprint arXiv:1906.03504*, 2019.
- Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 2017–2025, 2015.
- Anil K. Jain, Richard C. Dubes, et al. *Algorithms for Clustering Data*, volume 6. Prentice hall Englewood Cliffs, NJ, 1988.
- Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems*, pages 4016–4027, 2018.

- Frank Jäkel, Manish Singh, Felix A. Wichmann, and Michael H. Herzog. An overview of quantitative approaches in Gestalt perception. *Vision research*, 126:3–8, 2016. ISSN 0042-6989. doi: 10.1016/j.visres.2016.06.004.
- Michael Janner, Sergey Levine, William T. Freeman, Joshua B. Tenenbaum, Chelsea Finn, and Jiajun Wu. Reasoning about physical interactions with object-oriented prediction and planning. In *International Conference on Learning Representations*, 2019.
- Allan D. Jepson and Michael J. Black. Mixture models for optical flow computation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 760–761, 1993. doi: 10.1109/CVPR.1993.341161.
- Xu Ji, João F. Henriques, and Andrea Vedaldi. Invariant Information Clustering for Unsupervised Image Classification and Segmentation. In *IEEE International Conference on Computer Vision*, 2019.
- Jindong Jiang and Sungjin Ahn. Generative Neurosymbolic Machines. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- Jindong Jiang, Sepehr Janghorbani, Gerard de Melo, and Sungjin Ahn. SCALOR: Generative world models with scalable object representations. In *International Conference on Learning Representations*, 2020.
- Jason Jo and Yoshua Bengio. Measuring the tendency of CNNs to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.
- Philip N. Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010.
- Armand Joulin and Tomas Mikolov. Inferring algorithmic patterns with stack-augmented recurrent nets. In *Advances in Neural Information Processing Systems*, pages 190–198, 2015.
- Daniel Kahneman, Anne Treisman, and Brian J. Gibbs. The reviewing of object files: Object-specific integration of information. *Cognitive psychology*, 24(2):175–219, 1992. ISSN 0010-0285.
- \Lukasz Kaiser and Ilya Sutskever. Neural gpus learn algorithms. In *International Conference on Learning Representations*, 2016.
- Pentti Kanerva. Binary spatter-coding of ordered K-tuples. In *International Conference on Artificial Neural Networks*, pages 869–873, 1996. doi: 10.1007/3-540-61510-5\_146.
- Ken Kansky, Tom Silver, David A. Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, D. Scott Phoenix, and Dileep George. Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. In *International Conference on Machine Learning*, pages 1809–1818, Sydney, NSW, Australia, 2017.
- Astrid M. L. Kappers and Wouter M. Bergmann Tiest. Tactile and haptic perceptual organization. *The Oxford handbook of perceptual organization*, pages 621–638, 2015.
- Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.
- Matthew A. Kelly, Dorothea Blostein, and Douglas J. K. Mewhort. Encoding structure in holographic reduced representations. *Canadian Journal of Experimental Psychology*, 67(2):79–93, 2013. ISSN 1196-1961. doi: 10.1037/a0030301.

- Charles Kemp and Joshua B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0802631105.
- Richard Kempter, Wulfram Gerstner, and J. Leo Van Hemmen. Hebbian learning and spiking neurons. *Physical Review E*, 59(4):4498, 1999.
- Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, page 38, 2020.
- Junkyung Kim, Matthew Ricci, and Thomas Serre. Not-So-CLEVR: Learning same–different relations strains feedforward neural networks. *Interface focus*, 8(4):20180011, 2018.
- Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, pages 2688–2697, 2018.
- Thomas Kipf, Elise van der Pol, and Max Welling. Contrastive learning of structured world models. In *International Conference on Learning Representations*, 2020.
- Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.
- Kurt Koffka. *Principles of Gestalt Psychology*, volume 44. Routledge, 1935.
- Teuvo Kohonen. *Self-Organization and Associative Memory*. Springer, third edition, 1989. ISBN 978-0-387-18314-5.
- Shu Kong and Charless C. Fowlkes. Recurrent pixel embedding for instance grouping. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9018–9028, 2018.
- Alfred Korzybski. *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*. Institute of GS, 1958.
- Adam Kosiorek, Alex Bewley, and Ingmar Posner. Hierarchical attentive recurrent tracking. In *Advances in Neural Information Processing Systems*, pages 3053–3061, 2017.
- Adam Kosiorek, Hyunjik Kim, Yee Whye Teh, and Ingmar Posner. Sequential attend, infer, repeat: Generative modelling of moving objects. In *Advances in Neural Information Processing Systems*, pages 8606–8616, 2018.
- Adam Kosiorek, Sara Sabour, Yee Whye Teh, and Geoffrey E. Hinton. Stacked capsule autoencoders. In *Advances in Neural Information Processing Systems*, pages 15486–15496, 2019.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012.
- Tejas D. Kulkarni, Pushmeet Kohli, Joshua B. Tenenbaum, and Vikash K. Mansinghka. Picture: A probabilistic programming language for scene perception. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4390–4399, 2015.
- Tejas D. Kulkarni, Ankush Gupta, Catalin Ionescu, Sebastian Borgeaud, Malcolm Reynolds, Andrew Zisserman, and Volodymyr Mnih. Unsupervised learning of object keypoints for perception and control. In *Advances in Neural Information Processing Systems*, pages 10723–10733, 2019.

- Karol Kurach, Marcin Andrychowicz, and Ilya Sutskever. Neural random-access machines. In *International Conference on Learning Representations*, 2016.
- Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International Conference on Machine Learning*, 2017.
- Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pages 2873–2882, 2018.
- Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017. ISSN 0140-525X. doi: 10.1017/S0140525X16001837.
- George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago press, 2008.
- Christoph H. Lampert, Matthew B. Blaschko, and Thomas Hofmann. Beyond sliding windows: Object localization by efficient subwindow search. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. doi: 10.1109/CVPR.2008.4587586.
- Guillaume Lample and François Charton. Deep learning for symbolic mathematics. In *International Conference on Learning Representations*, 2020.
- Nicolas Le Roux, Nicolas Heess, Jamie Shotton, and John Winn. Learning a generative model of images by factoring appearance and shape. *Neural Computation*, 23(3):593–650, 2011. ISSN 0899-7667.
- Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *International Conference on Machine Learning*, pages 3734–3743, 2019.
- Te-Won Lee and Michael S. Lewicki. Unsupervised image classification, segmentation, and enhancement using ICA mixture models. *IEEE Transactions on Image Processing*, 11(3):270–279, 2002. ISSN 1057-7149. doi: 10.1109/83.988960.
- Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. In *International Conference on Learning Representations*, 2016.
- Yunzhu Li, Toru Lin, Kexin Yi, Daniel Bear, Daniel Yamins, Jiajun Wu, Joshua Tenenbaum, and Antonio Torralba. Visual grounding of learned physical models. In *International Conference on Machine Learning*, pages 5927–5936, 2020.
- Renjie Liao, Yujia Li, Yang Song, Shenlong Wang, Will Hamilton, David K. Duvenaud, Raquel Urtasun, and Richard Zemel. Efficient graph generation with graph recurrent attention networks. In *Advances in Neural Information Processing Systems*, pages 4257–4267, 2019.
- Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. SPACE: Unsupervised object-oriented scene representation via spatial attention and decomposition. In *International Conference on Learning Representations*, 2020.
- Or Litany, Alex Bronstein, Michael Bronstein, and Ameesh Makadia. Deformable shape completion with graph convolutional autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1886–1895, Salt Lake City, UT, USA, June 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00202.

- Jenny Liu, Aviral Kumar, Jimmy Ba, Jamie Kiros, and Kevin Swersky. Graph normalizing flows. In *Advances in Neural Information Processing Systems*, volume 32, pages 13578–13588, 2019.
- Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Ilya Tolstikhin. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pages 1452–1461, 2015.
- Joao Loula, Marco Baroni, and Brenden M. Lake. Rearranging the familiar: Testing compositional generalization in recurrent networks. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.
- Carmen Luciano, Inmaculada Gómez Becerra, and Miguel Rodríguez Valverde. The role of multiple-exemplar training and naming in establishing derived equivalence in an infant. *Journal of the Experimental Analysis of Behavior*, 87(3):349–365, 2007.
- Zhaoliang Lun, Changqing Zou, Haibin Huang, Evangelos Kalogerakis, Ping Tan, Marie-Paule Cani, and Hao Zhang. Learning to group discrete graphical patterns. *ACM Transactions on Graphics*, 36(6):225:1–225:11, 2017. ISSN 0730-0301. doi: 10.1145/3130800.3130841.
- Jitendra Malik, Serge Belongie, Thomas Leung, and Jianbo Shi. Contour and texture analysis for image segmentation. *International journal of computer vision*, 43(1):7–27, 2001. ISSN 0920-5691. doi: 10.1023/A:1011174803800.
- Vikash K. Mansinghka, Tejas D. Kulkarni, Yura N. Perov, and Joshua B. Tenenbaum. Approximate Bayesian image interpretation using generative probabilistic graphics programs. In *Advances in Neural Information Processing Systems*, volume 26, pages 1520–1528. Curran Associates, Inc., 2013.
- Jiayuan Mao and Chuang Gan. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*, page 28, 2019.
- Gary F. Marcus. *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT press, 2003. ISBN 978-0-262-63268-3.
- David R. Martin, Charless C. Fowlkes, and Jitendra Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE transactions on pattern analysis and machine intelligence*, 26(5):530–549, 2004. ISSN 0162-8828. doi: 10.1109/TPAMI.2004.1273918.
- Kenneth McGarry, Stefan Wermter, and John MacIntyre. Hybrid neural systems: From simple coupling to fully integrated neural networks. *Neural Computing Surveys*, 2(1):62–93, 1999.
- Harry McGurk and John Macdonald. Hearing lips and seeing voices. *Nature*, 264(5588):746, 1976. ISSN 1476-4687. doi: 10.1038/264746a0.
- Clayton McMillan, Michael C Mozer, and Paul Smolensky. Rule induction through integrated symbolic and subsymbolic processing. In *Advances In Neural Information Processing Systems*, pages 969–976, 1992.
- Bjorn Merker. Cortical gamma oscillations: The functional key is activation, not cognition. *Neuroscience & Biobehavioral Reviews*, 37(3):401–417, 2013. ISSN 0149-7634. doi: 10.1016/j.neubiorev.2013.01.013.

- Albert Michotte, Georges Thinès, and Geneviève Crabbé. Amodal completion of perceptual structures. *Michotte's experimental phenomenology of perception*, pages 140–167, 1991.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.
- George A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.
- Peter M. Milner. A model for visual shape recognition. *Psychological review*, 81(6):521, 1974. ISSN 0033-295X.
- Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole, Kevin P. Murphy, and Honglak Lee. Unsupervised learning of object structure and dynamics from videos. In *Advances in Neural Information Processing Systems*, pages 92–102, 2019.
- Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, volume 27, pages 2204–2212, 2014.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, and Georg Ostrovski. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
- Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. In *International Conference on Learning Representations*, 2019.
- Hossein Mobahi, Shankar R. Rao, Allen Y. Yang, Shankar S. Sastry, and Yi Ma. Segmentation of natural images by texture and boundary compression. *International journal of computer vision*, 95(1):86–98, 2011. ISSN 0920-5691. doi: 10.1007/s11263-011-0444-0.
- Igor Mordatch. Concept learning with energy-based models. In Ashok K. Goel, Colleen M. Seifert, and Christian Freksa, editors, *Annual Meeting of the Cognitive Science Society*, pages 58–59. cognitivesciencesociety.org, 2019.
- Alexander Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo Jimenez Rezende. Towards interpretable reinforcement learning using attention augmented agents. In *Advances in Neural Information Processing Systems*, volume 32, pages 12350–12359. Curran Associates, Inc., 2019.
- Michael C. Mozer. Types and tokens in visual letter perception. *Journal of experimental psychology: Human perception and performance*, 15(2):287–303, 1989. doi: <https://psycnet.apa.org/doi/10.1037/0096-1523.15.2.287>.
- Michael C. Mozer and Sreerupa Das. A connectionist symbol manipulator that discovers the structure of context-free languages. In *Advances in Neural Information Processing Systems*, pages 863–870, 1993.
- Michael C. Mozer, Richard S. Zemel, and Marlene Behrmann. Learning to segment images using dynamic feature binding. In *Advances in Neural Information Processing Systems*, volume 4, pages 436–443. Morgan-Kaufmann, 1992.
- Michael C. Mozer, Denis Kazakov, and Robert V. Lindsey. State-denoised recurrent neural networks. *cs.colorado.edu*, 2018.

- Damian Mrowca, Chengxu Zhuang, Elias Wang, Nick Haber, Li Fei-Fei, Joshua B. Tenenbaum, and Daniel L. K. Yamins. Flexible neural representation for physics prediction. In *Advances in Neural Information Processing Systems*, pages 8799–8810, 2018.
- Tsendsuren Munkhdalai and Hong Yu. Neural semantic encoders. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 397–407, 2017.
- Li Nanbo, Cian Eastwood, and Robert Fisher. Learning object-centric representations of multi-object scenes from multiple views. *Advances in Neural Information Processing Systems*, 33, 2020.
- Charles Nash, S. M. Ali Eslami, Christopher P. Burgess, Irina Higgins, Daniel Zoran, Theophane Weber, and Peter W. Battaglia. The multi-entity variational autoencoder. *Neural Information Processing Systems (NeurIPS) Workshop on Learning Disentangled Representations: from Perception to Control*, 2017.
- Camilo Rodrigues Neto and Jose Fernando Fontanari. Multivalley structure of attractor neural networks. *Journal of Physics A: Mathematical and General*, 30(22):7945, 1999. ISSN 0305-4470. doi: 10.1088/0305-4470/30/22/028.
- Allen Newell and Herbert A. Simon. Computer science as empirical inquiry: Symbols and search. *Mind design*, page 41, 1981.
- Allen Newell, John C. Shaw, and Herbert A. Simon. Report on a general problem solving program. In *IFIP Congress*, volume 256, page 64. Pittsburgh, PA, 1959.
- Thu Nguyen-Phuoc, Christian Richardt, Long Mai, Yong-Liang Yang, and Niloy Mitra. BlockGAN: Learning 3D object-aware scene representations from unlabelled images. In *Advances in Neural Information Processing Systems*, volume 33, 2020.
- Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, pages 6338–6347, 2017.
- Ernst Niebur, Steven S. Hsiao, and Kenneth O. Johnson. Synchrony: A neuronal mechanism for attentional selection? *Current opinion in neurobiology*, 12(2):190–194, 2002.
- Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. *arXiv preprint arXiv:2011.12100*, 2020.
- Malvina Nissim, Rik van Noord, and Rob van der Goot. Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2):487–497, 2019.
- Dimitri Nowicki and Hava T. Siegelmann. Flexible kernel memory. *PLoS One*, 5(6):e10955, 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0010955.
- Stellan Ohlsson. Restructuring revisited: I. Summary and critique of the Gestalt theory of problem solving. *Scandinavian Journal of Psychology*, 25(1):65–78, 1984.
- Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. ISSN 2476-0757. doi: 10.23915/distill.00007.
- Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024.001, 2020. ISSN 2476-0757. doi: 10.23915/distill.00024.001.

- Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. ISSN 0028-0836. doi: 10.1038/381607a0.
- Bruno A. Olshausen, Charles H. Anderson, and David C. Van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13(11):4700–4719, 1993.
- Gergo Orbán, József Fiser, Richard N. Aslin, and Máté Lengyel. Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, 105(7):2745–2750, 2008. ISSN 0027-8424. doi: 10.1073/pnas.0708424105.
- Randall C. O'Reilly and Richard S. Busby. Generalizable relational binding from coarse-coded distributed representations. In *Advances in Neural Information Processing Systems*, volume 14, pages 75–82, 2002.
- Lucas Paletta, Gerald Fritz, and Christin Seifert. Q-learning of sequential attention for visual object recognition from informative local descriptors. In *International Conference on Machine Learning*, pages 649–656, New York, NY, USA, 2005. doi: 10.1145/1102351.1102433.
- Rasmus Berg Palm, Ulrich Paquet, and Ole Winther. Recurrent relational networks. In *Advances in Neural Information Processing Systems*, pages 3368–3378, 2018.
- Stephen Palmer and Irvin Rock. Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic bulletin & review*, 1(1):29–55, 1994.
- Stephen E. Palmer. Common region: A new principle of perceptual grouping. *Cognitive psychology*, 24(3):436–447, 1992.
- Agostina Palmigiano, Theo Geisel, Fred Wolf, and Demian Battaglia. Flexible information routing by transient synchrony. *Nature neuroscience*, 20(7):1014–1022, 2017. ISSN 1097-6256. doi: 10.1038/nn.4569.
- Deepak Pathak, Ross B. Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2701–2710, 2017.
- Judea Pearl. *Causality*. Cambridge university press, 2009.
- Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.
- Luigi S. Pesce Ibarra. Synchronization matters for motor coordination. *Journal of neurophysiology*, page jn.00182.2017, 2017. ISSN 0022-3077. doi: 10.1152/jn.00182.2017.
- Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT press, 2017.
- Fernando J. Pineda. Generalization of back-propagation to recurrent neural networks. *Physical review letters*, 59(19):2229–2232, 1987. ISSN 0031-9007. doi: 10.1103/PhysRevLett.59.2229.

- Tony A. Plate. Holographic reduced representations. *IEEE Transactions on Neural networks*, 6(3):623–641, 1995. ISSN 1045-9227. doi: 10.1109/72.377968.
- Jordan B. Pollack. Recursive distributed representations. *Artificial Intelligence*, 46(1-2):77–105, 1990.
- William Prinzmetal. Visual feature integration in a world of objects. *Current Directions in Psychological Science*, 4(3):90–94, 1995. ISSN 0963-7214. doi: 10.1111/1467-8721.ep10772335.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.
- Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J. Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12240–12249, 2019.
- A. Ravishankar Rao, Guillermo A. Cecchi, Charles C. Peck, and James R. Kozloski. Unsupervised segmentation with dynamical units. *IEEE Transactions on Neural Networks*, 19(1):168–182, 2008. ISSN 1045-9227.
- Supratim Ray and John H. R. Maunsell. Do gamma oscillations play a role in cerebral cortex? *Trends in cognitive sciences*, 19(2):78–85, 2015. ISSN 1364-6613. doi: 10.1016/j.tics.2014.12.002.
- Scott Reed and Nando de Freitas. Neural programmer-interpreters. In *International Conference on Learning Representations*, 2015.
- David P. Reichert and Thomas Serre. Neuronal synchrony in complex-valued deep networks. In *International Conference on Learning Representations*, 2014.
- Mengye Ren and Richard S. Zemel. End-to-end instance segmentation with recurrent attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 293–301, Honolulu, HI, July 2017. IEEE. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.39.
- Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 91–99, 2015.
- Xiaofeng Ren and Jitendra Malik. Learning a classification model for segmentation. In *IEEE International Conference on Computer Vision*, pages 10–17 vol.1, 2003. doi: 10.1109/ICCV.2003.1238308.
- John H. Reynolds and Robert Desimone. The role of neural mechanisms of attention in solving the binding problem. *Neuron*, 24(1):19–29, 1999. ISSN 0896-6273. doi: 10.1016/S0896-6273(00)80819-3.
- Karl Ridgeway and Michael C. Mozer. Learning deep disentangled embeddings with the F-statistic loss. In *Advances in Neural Information Processing Systems*, pages 185–194, 2018.
- Maximilian Riesenhuber and Tomaso Poggio. Are cortical models really bound by the "binding problem"? *Neuron*, 24(1):87–93, 1999. ISSN 0896-6273. doi: 10.1016/S0896-6273(00)80824-7.
- Lynn C. Robertson. Attention and binding. In Laurent Itti, Geraint Rees, and John K. Tsotsos, editors, *Neurobiology of Attention*, pages 135–139. Academic Press, Burlington, 2005. ISBN 978-0-12-375731-9. doi: 10.1016/B978-012375731-9/50028-8.

- Lukasz Romaszko, Christopher K. I. Williams, Pol Moreno, and Pushmeet Kohli. Vision-as-inverse-graphics: Obtaining a rich 3D explanation of a scene from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 851–859, 2017.
- Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, DTIC Document, 1961.
- Adina L. Roskies. The binding problem. *Neuron*, 24(1):7–9, 111–25, 1999. ISSN 0896-6273.
- Henry A. Rowley, Shumeet Baluja, and Takeo Kanade. Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence*, 20(1):23–38, 1998. ISSN 0162-8828. doi: 10.1109/34.655647.
- Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, volume 30, pages 3856–3866, 2017.
- Ramin Samadani. A finite mixtures algorithm for finding proportions in SAR images. *IEEE Transactions on Image Processing*, 4(8):1182–1186, 1995. ISSN 1057-7149. doi: 10.1109/83.403427.
- Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, volume 30, pages 4967–4976, 2017.
- Adam Santoro, Ryan Faulkner, David Raposo, Jack Rae, Mike Chrzanowski, Theophane Weber, Daan Wierstra, Oriol Vinyals, Razvan Pascanu, and Timothy Lillicrap. Relational recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 7299–7310, 2018a.
- Adam Santoro, Felix Hill, David G. T. Barrett, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning*, volume 80, pages 4477–4486, Stockholm Sweden, 2018b.
- Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009. ISSN 1045-9227. doi: 10.1109/TNN.2008.2005605.
- Imanol Schlag and Jürgen Schmidhuber. Learning to reason with third order tensor products. In *Advances in Neural Information Processing Systems*, pages 9981–9993, 2018.
- Imanol Schlag, Paul Smolensky, Roland Fernandez, Nebojsa Jojic, Jürgen Schmidhuber, and Jianfeng Gao. Enhancing the transformer with explicit relational encoding for math problem solving. *arXiv preprint arXiv:1910.06611*, 2019.
- Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In Aldo Gangemi, Roberto Navigli, Maria-Ester Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web*, Lecture Notes in Computer Science, pages 593–607, Cham, 2018. ISBN 978-3-319-93417-4. doi: 10.1007/978-3-319-93417-4\_38.
- Jürgen Schmidhuber. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139, 1992a. ISSN 0899-7667.
- Jürgen Schmidhuber. Learning complex, extended sequences using the principle of history compression. *Neural Computation*, 4(2):234–242, 1992b. ISSN 0899-7667.

- Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992c. ISSN 0899-7667. doi: 10.1162/neco.1992.4.6.863.
- Jürgen Schmidhuber. On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models. *arXiv preprint arXiv:1511.09249*, 2015.
- Jürgen Schmidhuber and Rudolf Huber. Learning to generate artificial fovea trajectories for target detection. *International Journal of Neural Systems*, 2(01n02):125–134, 1991. ISSN 0129-0657.
- Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.
- Jean-Luc Schwartz, Nicolas Grimault, Jean-Michel Hupé, Brian C. J. Moore, and Daniel Pressnitzer. Multistability in perception: Binding sensory modalities, an overview. *Philosophical Transactions of the Royal Society B*, 367(1591):896–905, 2012. ISSN 0962-8436. doi: 10.1098/rstb.2011.0254.
- Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. ISSN 0162-8828. doi: 10.1109/34.868688.
- Edward H. Shortliffe, Randall Davis, Stanton G. Axline, Bruce G. Buchanan, C. Cordell Green, and Stanley N. Cohen. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Computers and biomedical research*, 8(4):303–320, 1975.
- Murray Sidman. Reading and auditory-visual equivalences. *Journal of speech and Hearing Research*, 14(1):5–13, 1971.
- Murray Sidman, Constance K. Wynne, Russell W. Maguire, and Thomas Barnes. Functional classes and equivalence relations. *Journal of the Experimental analysis of Behavior*, 52(3):261–274, 1989.
- Hava T. Siegelmann and David Sontag. Turing computability with neural nets. *Applied Mathematics Letters*, 4(6):77–80, 1991. ISSN 0893-9659.
- David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, and Marc Lanctot. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, and Thore Graepel. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.
- Wolf Singer. Neuronal synchrony: A versatile code for the definition of relations? *Neuron*, 24(1):49–65, 111–25, 1999. ISSN 0896-6273.
- Wolf Singer. Consciousness and the binding problem. *Annals of the New York Academy of Sciences*, 929(1):123–146, 2001.
- Wolf Singer. Binding by synchrony. *Scholarpedia*, 2(12):1657, 2007. ISSN 1941-6016. doi: 10.4249/scholarpedia.1657.
- Wolf Singer. Distributed processing and temporal codes in neuronal networks. *Cognitive neurodynamics*, 3(3):189–196, 2009. ISSN 1871-4080. doi: 10.1007/s11571-009-9087-z.

- Paul Smolensky. Analysis of distributed representation of constituent structure in connectionist systems. In *Neural Information Processing Systems*, pages 730–739, 1987.
- Paul Smolensky. On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1):1–23, 1988. ISSN 0140-525X. doi: 10.1017/S0140525X00052432.
- Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1):159–216, 1990. ISSN 0004-3702. doi: 10.1016/0004-3702(90)90007-M.
- Elliot Sollow, Judy Bachant, and Keith Jensen. Assessing the maintainability of XCQN-in-RIME: Coping with the problems of a VERY large rule-base. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, volume 2, pages 824–829, 1987.
- Elizabeth S. Spelke and Katherine D. Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007.
- Charles Spence and Christian Frings. Multisensory feature integration in (and out) of the focus of spatial attention. *Attention, Perception, & Psychophysics*, 82(1):363–376, 2020.
- Alessandro Sperduti and Antonina Starita. Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, 8(3):714–735, 1997.
- Aleksandar Stanić, Sjoerd van Steenkiste, and Jürgen Schmidhuber. Hierarchical relational inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
- Kenneth O. Stanley and Risto Miikkulainen. Evolving a roving eye for go. In *Genetic and Evolutionary Computation Conference*, pages 1226–1238, 2004.
- Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 2440–2448. Curran Associates, Inc., 2015.
- Chen Sun, Per Karlsson, Jiajun Wu, Joshua B. Tenenbaum, and Kevin Murphy. Stochastic prediction of multi-agent interactions from partial observations. In *International Conference on Learning Representations*, 2019.
- Ron Sun. On Variable Binding in Connectionist Networks. *Connection Science*, 4(2):93–124, 1992. ISSN 0954-0091. doi: 10.1080/09540099208946607.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.
- Richard Sutton. The bitter lesson. <http://www.incompleteideas.net/IncIdeas/BitterLesson.html>, 2019.
- Zoltán Gendler Szabó. Compositionality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition, 2017.
- Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1166, 2015.
- Catherine Tallon-Baudry and Olivier Bertrand. Oscillatory gamma activity in humans and its role in object representation. *Trends in cognitive sciences*, 3(4):151–162, 1999.

- Yichuan Tang, Nitish Srivastava, and Ruslan R. Salakhutdinov. Learning generative models with visual attention. In *Advances in Neural Information Processing Systems*, volume 27, pages 1808–1816, 2014.
- Valentin Thomas, Jules Pondard, Emmanuel Bengio, Marc Sarfati, Philippe Beaudoin, Marie-Jean Meurs, Joelle Pineau, Doina Precup, and Yoshua Bengio. Independently controllable factors. *arXiv preprint arXiv:1708.01289*, 2017.
- Anne Treisman. Focused attention in the perception and retrieval of multidimensional stimuli. *Perception & Psychophysics*, 22(1):1–11, 1977. ISSN 0031-5117. doi: 10.3758/BF03206074.
- Anne Treisman. The binding problem. *Current opinion in neurobiology*, 6(2):171–178, 1996. ISSN 0959-4388. doi: 10.1016/S0959-4388(96)80070-5.
- Anne Treisman. Solutions to the binding problem: Progress through controversy and convergence. *Neuron*, 24(1):105–10, 111–25, 1999. ISSN 0896-6273.
- Anne Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. ISSN 0010-0285.
- Anne Treisman and Weiwei Zhang. Location and binding in visual working memory. *Memory & cognition*, 34(8):1704–1719, 2006.
- Pedro A. Tsividis, Thomas Pouncy, Jacqueline L. Xu, Joshua B. Tenenbaum, and Samuel J. Gershman. Human learning in Atari. In *2017 AAAI Spring Symposium Series*, 2017.
- Zhuowen Tu and Song-Chun Zhu. Image segmentation by data-driven Markov chain Monte Carlo. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):657–673, 2002. ISSN 0162-8828. doi: 10.1109/34.1000239.
- Zhuowen Tu, Xiangrong Chen, Alan L. Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of computer vision*, 63(2):113–140, 2005. ISSN 0920-5691. doi: 10.1007/s11263-005-6642-x.
- Peter Uhlhaas, Gordon Pipa, Bruss Lima, Lucia Melloni, Sergio Neuenschwander, Danko Nikolić, and Wolf Singer. Neural synchrony in cortical networks: History, concept and current status. *Frontiers in integrative neuroscience*, 3:17, 2009.
- Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. ISSN 0920-5691. doi: 10.1007/s11263-013-0620-5.
- Marius Usher and Nick Donnelly. Visual synchrony affects binding and segmentation in perception. *Nature*, 394(6689):179–182, 1998.
- Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- Sjoerd van Steenkiste, Michael Chang, Klaus Greff, and Jürgen Schmidhuber. Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. In *International Conference on Learning Representations*, 2018.
- Sjoerd van Steenkiste, Klaus Greff, and Jürgen Schmidhuber. A perspective on objects and systematic generalization in model-based RL. *International Conference on Machine Learning (ICML) Workshop on Generative Modeling and Model-based Reasoning for Robotics and AI*, 2019a.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, pages 14245–14258, 2019b.

Sjoerd van Steenkiste, Karol Kurach, Jürgen Schmidhuber, and Sylvain Gelly. Investigating object compositionality in generative adversarial networks. *Neural Networks*, 130:309–325, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. In *International Conference on Learning Representations*, 2018.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010. ISSN 1532-4435.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 2692–2700, 2015.

Paul Viola and Michael J. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–511–I–518 vol.1, 2001. doi: 10.1109/CVPR.2001.990517.

Christoph von der Malsburg. The correlation theory of brain function. Technical report, 1981.

Christoph von der Malsburg. Am I thinking assemblies? In *Brain Theory*, pages 161–176. Springer, 1986.

Wilhelm von Humboldt. *Humboldt: 'On Language': On the Diversity of Human Language Construction and Its Influence on the Mental Development of the Human Species*. Cambridge University Press, 1999.

Julius von Kügelgen, Ivan Ustyuzhaninov, Peter Gehler, Matthias Bethge, and Bernhard Schölkopf. Towards causal generative scene models via competition of experts. *International Conference on Learning Representations (ICLR) Workshop on "Causal learning for decision making"*, 2020.

Edward Vul and Donald I. A. MacLeod. Contingent aftereffects distinguish conscious and preconscious color processing. *Nature neuroscience*, 9(7):873–874, 2006.

Edward Vul, Cory A. Rieth, Timothy F. Lew, and Anina N. Rich. The structure of illusory conjunctions reveals hierarchical binding of multipart objects. *Attention, Perception, & Psychophysics*, pages 1–14, 2019.

Johan Wagemans. *The Oxford Handbook of Perceptual Organization*. Oxford University Press, 2015. ISBN 978-0-19-968685-8.

Johan Wagemans, James H. Elder, Michael Kubovy, Stephen E. Palmer, Mary A. Peterson, Manish Singh, and Rüdiger von der Heydt. A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *psycnet.apa.org*, 2012a.

- Johan Wagemans, Jacob Feldman, Sergei Gepshtein, Ruth Kimchi, James R. Pomerantz, Peter A. van der Helm, and Cees van Leeuwen. A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological bulletin*, 138(6):1218–1252, 2012b. ISSN 0033-2909. doi: 10.1037/a0029334.
- Deliang Wang. The time dimension for scene analysis. *IEEE Transactions on Neural Networks*, 16(6):1401–1426, 2005. ISSN 1045-9227. doi: 10.1109/TNN.2005.852235.
- Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- Tianhan Wei, Xiang Li, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2869–2878, 2020.
- Yair Weiss and Edward H. Adelson. A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 321–326, 1996. doi: 10.1109/CVPR.1996.517092.
- Joseph Weizenbaum. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.
- Max Wertheimer. Experimentelle Studium über das Sehen von Bewegung. *Zeitschrift für psychologie*, 61(3):161–265, 1912. ISSN 2190-8370.
- Max Wertheimer. Untersuchungen zur Lehre von der Gestalt II. *Psychologische forschung*, 4(1):301–350, 1923. ISSN 0033-3026.
- Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *International Conference on Learning Representations*, 2015.
- Alfred North Whitehead. *Symbolism: Its Meaning and Effect*. Fordham University Press, New York, revised ed. edition edition, 1985. ISBN 978-0-8232-1138-8.
- William Whitney. *Disentangled Representations in Neural Models*. Masters Thesis, MIT, 2016.
- Oliver Wilhelm, Andrea Hildebrandt Hildebrandt, and Klaus Oberauer. What is working memory capacity, and how can we measure it? *Frontiers in Psychology*, 4, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00433.
- Terry Winograd. Procedures as a representation for data in a computer program for understanding natural language. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC, 1971.
- Jeremy M. Wolfe. Forty years after feature integration theory: An introduction to the special issue in honor of the contributions of Anne Treisman. *Attention, Perception, & Psychophysics*, 82(1):1–6, 2020. ISSN 1943-393X. doi: 10.3758/s13414-019-01966-3.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, and Klaus Macherey. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.
- Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

- Kelvin Xu, Jimmy Ba, Jamie Ryan Kiros, Aaron Courville, Ruslan R. Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.
- Carl Yang, Peiye Zhuang, Wenhan Shi, Alan Luu, and Pan Li. Conditional structure generation through graph variational generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 1338–1349, 2019.
- Yanchao Yang, Yutong Chen, and Stefano Soatto. Learning to manipulate individual objects in an image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6558–6567, 2020.
- Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.
- Wlodek Zadrozny. From compositional to systematic semantics. *Linguistics and Philosophy*, 17(4):329–342, 1994. ISSN 1573-0549. doi: 10.1007/BF00985572.
- Vinicio Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, Murray Shanahan, Victoria Langston, Razvan Pascanu, Matthew Botvinick, Oriol Vinyals, and Peter Battaglia. Deep reinforcement learning with relational inductive biases. In *International Conference on Learning Representations*, 2019.
- Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.
- Richard S. Zemel and Michael C. Mozer. Localist attractor networks. *Neural Computation*, 13(5):1045–1064, 2001. ISSN 0899-7667.
- Richard S. Zemel, Christopher K. I. Williams, and Michael C. Mozer. Lending direction to neural networks. *Neural Networks*, 8(4):503–512, 1995. ISSN 0893-6080. doi: 10.1016/0893-6080(94)00094-3.
- Chiyuan Zhang, Oriol Vinyals, Remi Munos, and Samy Bengio. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018.
- Muhan Zhang, Shali Jiang, Zhicheng Cui, Roman Garnett, and Yixin Chen. D-VAE: A variational autoencoder for directed acyclic graphs. In *Advances in Neural Information Processing Systems*, pages 1588–1600, 2019.
- Yibiao Zhao and Song-Chun Zhu. Image parsing with stochastic scene grammar. In *Advances in Neural Information Processing Systems*, volume 24, pages 73–81, 2011.
- Andrey Zhmoginov, Ian Fischer, and Mark Sandler. Information-bottleneck approach to salient region discovery. *International Conference on Machine Learning (ICML) Workshop on Self-Supervised Learning*, 2019.
- Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene CNNs. In *International Conference on Learning Representations*, 2015.
- Daniel Zoran, Mike Chrzanowski, Po-Sen Huang, Sven Gowal, Alex Mott, and Pushmeet Kohli. Towards robust image classification using sequential attention models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- Ariel Zylberberg, Diego Fernández Slezak, Pieter R. Roelfsema, Stanislas Dehaene, and Mariano Sigman. The brain’s router: A cortical network model of serial processing in the primate brain. *PLoS Computational Biology*, 6(4):e1000765, 2010.