

# Hybrid primal sketch: Combining analogy, qualitative representations, and computer vision for scene understanding<sup>☆</sup>

Kenneth D. Forbus<sup>a,\*</sup>, Kezhen Chen<sup>b</sup>, Wangcheng Xu<sup>c</sup>, Madeline Usher<sup>c</sup>

<sup>a</sup> Computer Science, Northwestern University, 2233 Tech Drive, Evanston, IL 60208, USA

<sup>b</sup> Together.ai, San Francisco, CA, USA

<sup>c</sup> Computer Science, Northwestern University, 2233 Tech Drive, Evanston, IL 60208, USA

## ARTICLE INFO

### Keywords:

Computer vision  
Qualitative representation  
Analogy  
Sketch recognition  
Visual relation detection  
Video understanding

## ABSTRACT

One of the purposes of perception is to bridge between sensors and conceptual understanding. Marr's *Primal Sketch* combined initial edge-finding with multiple downstream processes to capture aspects of visual perception such as grouping and stereopsis. Given the progress made in multiple areas of AI since then, we have developed a new framework inspired by Marr's work, the *Hybrid Primal Sketch*, which combines computer vision components into an ensemble to produce sketch-like entities which are then further processed by CogSketch, a model of high-level human vision, to produce both more detailed shape representations and scene representations which can be used for data-efficient learning via analogical generalization. This paper describes our theoretical framework, summarizes several previous experiments, and outlines a new experiment in progress on diagram understanding.

## 1. Introduction

Perception provides organisms with information about the world around them, organized in ways that are useful given their capabilities. For people, and human-like AI systems, this includes providing a bridge between sensory data and conceptual understanding: What objects are like, how they are arranged into scenes, and how these change over time. One of the early revolutions in computer vision was the *Primal Sketch* (Marr, 1982; 2010), which proposed that edges actually formed a kind of hybrid representation, a combination of metric and symbolic information that served as a starting point for many downstream computations, such as stereopsis and lightness. For example, stereopsis was modeled by aligning edges from pairs of images, since edges provide more stable information about the world than pixels. The primal sketch was used to construct the 2 ½ D sketch, which carved images up into regions, so that properties of surfaces could be estimated (e.g. lightness and material composition).

Computer vision has made considerable progress since that time. Deep learning has produced components for building vision systems that have revolutionary capabilities. How should these advances be used to build end-to-end vision systems? One approach, which is receiving the

most attention, is to build end-to-end systems by training over massive datasets (e.g. Google's Gemini, OpenAI's GPT4/V). These monolithic end-to-end systems often treat language as an end goal, e.g. describing a scene or answering questions about it, which is an oversimplification of the roles of vision in cognition. Moreover, they require massive amounts of training and suffer from hallucinations. We think that ignoring the roles of explicit intermediate representations may be a mistake. For example, there is converging evidence from neuroscience (e.g. Amoranpanth et al., 2012) that schematic representations computed from images are used in relational processing. Moreover, there is evidence that analogical processing appears to be used in human visual reasoning (Forbus & Lovett, 2022), which suggests that relational representations provide an important intermediate form of representation. These are hybrid representations, grounded in symbolic visual elements tied to locations and properties of the image, just as in the original primal sketch.

This paper describes an alternative we have been exploring, the *hybrid primal sketch* (HPS), which builds on these hypotheses. The hybrid primal sketch combines advances in computer vision with advances in qualitative visual representation and reasoning. It uses an ensemble of low-level computations, including components constructed via deep

<sup>☆</sup> This article is part of a special issue entitled: 'RACS' published in Cognitive Systems Research.

\* Corresponding author at: Computer Science, Northwestern University, 2233 Tech Drive, Evanston, IL, 60208, USA.

E-mail addresses: [forbus@northwestern.edu](mailto:forbus@northwestern.edu) (K.D. Forbus), [kzchen0204@gmail.com](mailto:kzchen0204@gmail.com) (K. Chen), [wangcheng.xu@northwestern.edu](mailto:wangcheng.xu@northwestern.edu) (W. Xu), [usher@northwestern.edu](mailto:usher@northwestern.edu) (M. Usher).

<https://doi.org/10.1016/j.cogsys.2025.101390>

Received 25 November 2024; Received in revised form 22 July 2025; Accepted 19 August 2025

Available online 24 August 2025

1389-0417/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

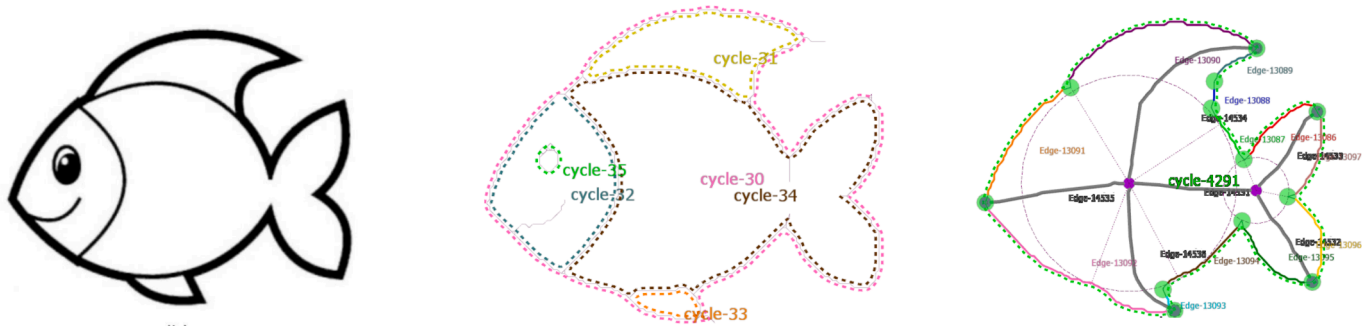


Fig. 1. Example of an image, an edge cycle decomposition of it, and its medial axis transform representation.

learning, to produce visual entities called *glyphs* for subsequent analysis. These glyphs are motivated by our work on sketch understanding, where a glyph consists of digital ink plus a conceptual label. The edges that make up regions or bounding boxes produced by visual components are treated as if they were digital ink. Any recognition information is treated as information about the conceptual label for the glyph. Thus the ensemble produces a set of glyphs, which are then analyzed at a higher level by CogSketch, a sketch understanding system that is also a model of aspects of high-level human vision (Forbus et al., 2011). This includes decomposing a glyph into fine-grained edge representations to capture its shape, and computing relationships between glyphs to construct relational representations of scenes. These descriptions can be used in analogical reasoning and learning. Analogical generalization provides incremental, data-efficient learning of inspectable models, producing performance on a par with, and sometimes better than, deep learning systems.

This paper describes the hybrid primal sketch (HPS) framework and illustrates it with several results, including one line of investigation in progress. We begin by providing brief key background and related work on computer vision, qualitative representations and analogical learning. Then we describe the HPS model, including how the ensemble works, glyphs as an intermediate representation, and highlights of shape and scene analysis. Completed experiments are summarized next, e.g. on sketch recognition, visual relation detection and Kinect video analysis. An experiment in progress on diagram understanding is described next. Finally, we discuss conclusions and future work.

## 2. Background and related work

Here we briefly summarize necessary background and closely related work.

### 2.1. Computer vision

A survey of even one area of computer vision is beyond the scope of this paper. Most relevant for this work are algorithms that extract and recognize visual structure. Broadly, these can be divided into edge finders, recognizers, and segmenters. Edge finders like Canny (1986) operate over natural images, finding zero crossings in intensity gradients that correspond to edges. The advantage of edge-finders is that they are (ideally) fully general-purpose, placing no limitations on the kinds of objects in the images. The disadvantage is that additional processing is required to use the edges found to propose segmentations of the image into regions and more complex structures. A special case is processing bitmaps that depict sketches (e.g. the MNIST dataset). For such images, a common technique is to treat them as binary images, by down-sampling and thresholding, then using a combination of edge thinners (e.g. Zhang & Suen 1984) and tracing algorithms (e.g. Potrace (Krenski & Selinger, 2003)) to produce edges. We use such techniques here to produce vector representations of sketches depicted in bitmaps.

By contrast, recognizers, like MASK-RCNN (He et al., 2017), directly

provide proposed regions and conceptual labels. However, they must be trained on the kinds of entities to be recognized, and hence are not as broadly applicable as edge finders. Nonetheless, recognizers are an extremely popular choice for building vision systems today. We have used both MASK-RCNN and Faster-RCNN (Ren et al., 2015) as object recognizers over natural images (Chen et al., 2020). Our techniques only assume that some form of boundary (bounding box or mask) is provided, plus a word or phrase as a conceptual label and should work with other recognizers. We note that an increasingly popular choice are open-vocabulary detectors, where one must provide the set of words corresponding to the kinds of entities sought in the image (Wu et al., 2024). We also have used OWLv2 (Minderer et al., 2023) as an input component, when top-down constraints suggest what kinds of entities should be sought (e.g. arrows).

Segmenters carve up an image into regions, roughly corresponding to entities (or pieces of them). For example, Meta's Segment Anything Model (SAM; Kirillov et al., 2023) was trained on over a billion masks across 11 million images. Downstream processes must still pull together interpretations of regions into objects, but given the smaller number of regions, this is a much easier problem than not having the segmentation.

## 3. Qualitative visual representations

Qualitative representations provide symbolic representations of continuous phenomena. In visual processing, qualitative representations have been introduced for both shapes (i.e. within-object relationships) and for scenes (i.e. between-object relationships). We discuss CogSketch's representations for each in turn.

For shapes, we focus here on the special case of sketches of objects, consisting of a set of visual edges. This factors out, for example, the color of regions (but see Chen et al., 2020). Edges can be qualitatively described in terms of their shape (i.e. straight or curved, and when part of a larger contour, concave or convex), their orientation (e.g. quadrants), and relative size (e.g. within a distribution provided by a diagram) – see (Forbus et al., 2017a) for details. Edges can be combined into *edge cycles*, which provide an intermediate representation that captures aspects of surfaces (McLure et al., 2011). There is evidence that people are sensitive to the medial axis transform in shape perception (Lowet et al., 2018). Fig. 1 illustrates edge cycles and the medial axis transform. These are used in the work discussed here, but there are other qualitative shape representations, e.g. (Museros & Escrig, 2004) which has been used in matching for robot assembly tasks.

For scenes, a variety of qualitative representations have been used. Many are qualitative spatial calculi, which provide a set of mutually exclusive and collectively exhaustive set of relationships (e.g. Moratz 2017). The most widely used is the eight-relation region connection calculus (RCC8, Cohn et al., 1997), which captures topological relationships like disjoint, edge-connected, overlapping, and inside. In purely qualitative spatial reasoning, transitivity tables are used to squeeze out implications of combinations of statements relating a set of objects. In vision settings like ours, such relationships can simply be

computed from metric shape representations and positions of the objects. Coordinate information is also used to compute common positional relationships, e.g. above/below, left/right.

### 3.1. Sketch understanding versus scene understanding

Work in artificial intelligence on sketch understanding tends to equate understanding with recognition (Forbus, 2019). In general, recognition is insufficient for sketch understanding because the mapping between sketches and the objects or systems they can depict is many-to-many. When people sketch, they usually talk, providing another channel of information that enables others in the interaction to understand them. In other words, even people do not rely solely on recognition when sketching with each other. Recognition is a catalyst, not a requirement. Sketching also often provides timing information, i.e. watching the ink being laid down. This extra source of information can help improve recognition, e.g. it is well known that in handwriting recognition, timing data is crucial for accuracy. In CogSketch (Forbus et al., 2011), people sketching lay down digital ink, manually segment it into entities, and provide conceptual labels from a large knowledge base. Thus, errors in both recognition and segmentation are entirely factored out, at the cost of putting a little more work on the sketcher. As discussed further in Section 4, these visual entities, called *glyphs*, are proposed here as an intermediate representation between low-level vision and high-level vision. The more tentative nature of segmentation and recognition when building atop computer vision outputs will sometimes require probabilistic abductive interpretation, as discussed below.

A recent approach to scene understanding is to train a multimodal foundation model, e.g. LLaVA (Liu et al., 2023) & Ferret v2 (Zhang et al., 2024). However, while such systems can have high performance, they also have massive hallucinations (Lu et al., 2024; Huang et al., 2024), require batch training, and produce uninspectable models. By contrast, the approach presented here incrementally produces inspectable models that are reliable.

### 3.2. Analogical learning

We use an *analogy stack* based on Structure-Mapping Theory (Gentner, 1983) that is built into CogSketch. Matching is performed via the structure-mapping engine (SME, Forbus et al., 2017b), which takes as input two structured representations and produces one or more mappings between them. Each mapping indicates what statements and entities in one description correspond with the other, what inferences follow from these correspondences, and a numerical estimate of their similarity. A mapping's inferences are what provide transfer between situations. Retrieval is performed by MAC/FAC (Forbus et al., 1995), a two-stage map reduce process. The first stage uses sparse vector dot-products over a massive library of cases to find a small number of items for the second stage, which uses SME in parallel with the structured version of the probe and the MAC output cases to select the few most similar as reminders. Generalization is performed by the Sequential Analogical Generalization Engine (SAGE; Kandaswamy & Forbus, 2012). SAGE models concepts via *generalization pools*, which are incrementally constructed by adding examples. These pools contain both generalizations and outliers. Generalizations are formed when a new example retrieves an outlier, i.e. a prior example that wasn't similar enough to anything else so far. Each new example added to a generalization causes the probability of statements in that generalization to be updated, e.g. statements that don't show up frequently wear away. A generalization pool can have multiple generalizations, providing the ability to represent disjunctive concepts. It is similar to K-means plus outliers, but SAGE figures out the number of generalizations automatically, rather than them having to be pre-specified. There is also a working memory version of SAGE, SageWM, where the generalization pools are in working memory and retrieval is based on SME similarity scores, modulated by recency.

We note that this analogy stack supports incremental learning, but today's deep learning systems do not. Thus, all the experiments used here for testing are batch settings.

There have been multiple models of analogy in cognitive science (e.g. Dumas et al., 2008; Hummel & Holyoak, 1997; Mitchell 1993). Unfortunately, these models do not have the representational capacity to handle the kinds of vision tasks described here.

We note that a different notion of generalization, arising from cartography, has been used with sketch maps (e.g. Manivannan et al., 2022). That notion of generalization refers to the representation decisions made by the map creator, in terms of level of detail used to capture aspects of the world. Different maps created by different people can involve different strategies, or the same strategies applied in different ways. Like analogical processing, analyzing sketch maps relies on finding 1:1 correspondences, which can be achieved by using representation techniques designed for that domain (Manivannan et al., 2024). Such techniques might well be useful in analogical learning from sketch maps. We note that they seem less helpful in analogical learning with hybrid primal sketch encodings because they are more uniform.

## 4. The hybrid primal sketch

The key idea of Marr's (1982) Primal Sketch was that vision has a symbolic component. Edges were viewed as a simple primitive symbolic representation, also having continuous properties such as location and contrast, being grounded in parallel visual processing elements. These descriptions were used in a variety of downstream processes. For example, visual structure involving texture would be found by performing grouping operations based on statistical properties of the edges and continuity of form constraints. Marr's group focused on both trying to understand the purpose of visual computations and how they were implemented biologically. While some of the original hypotheses of the Primal Sketch have not worked out, particularly in terms of the biological picture,<sup>1</sup> the overall flow of operations and the construction of intermediate representations – something like edges, regions that provide clues to surfaces in the world, abductive interpretation processes to accumulate local pieces into hypotheses about the perceived world – remain of interest. That said, many efforts in computer vision carve the processing up somewhat differently. For example, recognizers trained using deep learning directly on images provide labels for some region of the image (either a bounding box or a mask approximating the outline of the object being hypothesized). Are such recognizers biologically plausible? We leave that question aside here, since our interest is more practical: How can we use the advances in computer vision to build cognitive systems that can see? Our assumption is that at some point human visual processing does produce symbolic descriptions,<sup>2</sup> but we are agnostic as to where and how. Our concerns lie upstream: The computation of relational representations of shapes and scenes so that analogical learning can be used in visual processing. Even though visual stimuli are plentiful for most organisms, it still behooves them to learn incrementally, in a data-efficient manner. Since mistakes are inevitable in learning, organisms lucky enough to have language can discuss and refine their visual concepts, if they are inspectable. The kinds of qualitative visual and spatial representations developed in qualitative reasoning research are, we hypothesize, among the contents that can be inspected by people and hence should be inspectable by our machines, if they are to participate in discussions of them. In other words, at the level of inspectability, we want human-like representations and processing, whereas for computations below that level, as long as they produce

<sup>1</sup> See Palmer (1999) for an excellent discussion.

<sup>2</sup> In our view, symbolic representations can and do include metric properties, e.g. coordinates, lengths, etc. See Forbus (2019), Chapter 14 for details, including evidence for psychological and neural plausibility.

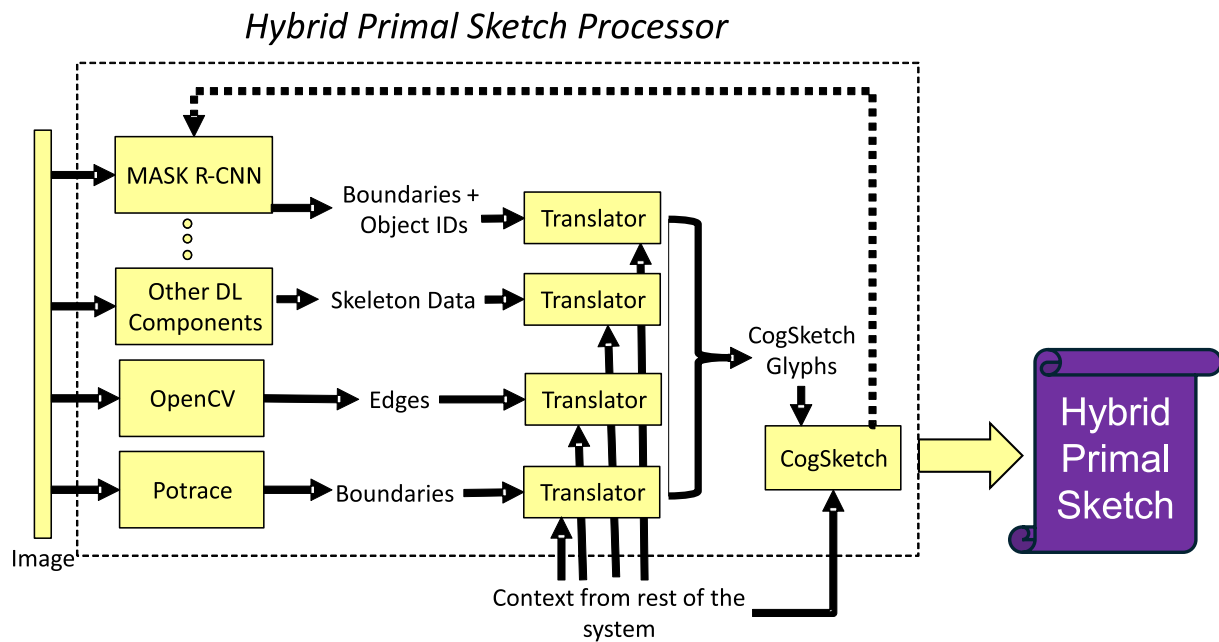


Fig. 2. Architecture of the Hybrid Primal Sketch Processor. The Hybrid Primal Sketch is a hybrid qualitative/quantitative symbolic description.

functionally reasonable representations, that is all that matters.

Thus, the Hybrid Primal Sketch consists of an ensemble of visual components (see Fig. 2). But what is the common currency that these components produce, which serves as an input to the more inspectable levels of processing? Our hypothesis is that the notion of *glyph* in open-domain sketch understanding (e.g. CogSketch; Forbus et al., 2011) provides that output format. In sketch understanding, a glyph is a visual entity consisting of digital ink and a conceptual label, produced via someone drawing. In some datasets, visual stimuli are encoded in the vector graphics format SVG, which can be directly imported into CogSketch. For bitmaps of drawings, there are algorithms which can produce such vector representations automatically (i.e. Potrace, Section 2.1), and hence can be treated as glyphs by CogSketch. For natural images, the bounding boxes or masks produced by recognizers (e.g. MASK-RCNN, Section 2.1) plus the label they provide can be treated as if they were glyphs. Thus, the elements of the ensemble produce glyphs for CogSketch, which then analyzes them in the same ways that it analyzes human-drawn sketches.

CogSketch was designed as a model of high-level human vision. Prior work has shown that it can be used to model human visual problem solving (Forbus & Lovett, 2022) and as a component in multimodal learning by reading (Chang 2016) and modeling conceptual change (Friedman et al., 2018). The use of analogy in Sketch Worksheets, a deployed educational software system built on CogSketch (Forbus et al., 2017a; Forbus et al., 2018) provides additional evidence that the representations it produces make sense to people and can be used robustly with a broad range of inputs. The candidate inferences produced by SME are used in Sketch Worksheets to provide feedback to students, based on analogical comparison of student sketches and instructor sketches.

The results in subsequent sections provide more evidence for this claim, in terms of learning visual concepts from images and Kinect video.

Given an image and a task context, which algorithms should be used? Currently, this is a procedural parameter set by the experimenter, who sets up a pipeline custom-built for that experiment, using the capabilities of the Companion cognitive architecture (Forbus & Hinrichs, 2017) to streamline the process (i.e. HTN plans provide a form of scripting, and Worker agents enable parallel processing to speed up encoding). We are moving to automatic control of processing by the Companion itself, as discussed in future work.

Table 1  
MNIST results.

Method	Training Size (per digit)	Accuracy
Analogy	10 × 1 epoch	54.9 %
Analogy	100 × 1 epoch	76.24
Analogy	500 × 1 epoch	85.03 %
LeNet-5	1,500 × 20 epochs	98.3 %
LeNet-5	6,000 × 20 epochs	99.2 %

## 5. Experiments with the hybrid primal sketch

We briefly summarize some prior experiments and describe an experiment in progress, to provide evidence for the utility of this framework. We start with sketch recognition, then turn to visual relationship detection, to illustrate its capabilities for handling images. We summarize experiments using Kinect video, to show how these ideas can be extended beyond single images. Finally, we discuss an experiment in progress on diagram understanding, since it provides a testbed for exploring the roles of conceptual understanding in scene understanding.

### 5.1. Sketch recognition

The classic analogical learning pipeline works as follows: Encode the stimuli to be learned and add it to a SAGE generalization pool for that concept. At any time, test items can be tried with MAC/FAC against the union of the generalization pools, with the classification of the test item being the label associated with the pool the top retrieval came from. Running this pipeline on the classic MNIST dataset (LeCun et al., 1998) provides an illustration. Encoding was performed by resizing the original image to be below 300 pixels, then blurred and filtered to black and white. Potrace and Zhang-Suen's thinning algorithm were used to generate SVG for CogSketch. CogSketch broke the SVG down into edges and edge cycles, producing a geon representation (Biederman, 1987), and computed a variety of properties for segments and geons (see (Chen, Rabkina, McLure, & Forbus, 2019) for details). We compared our

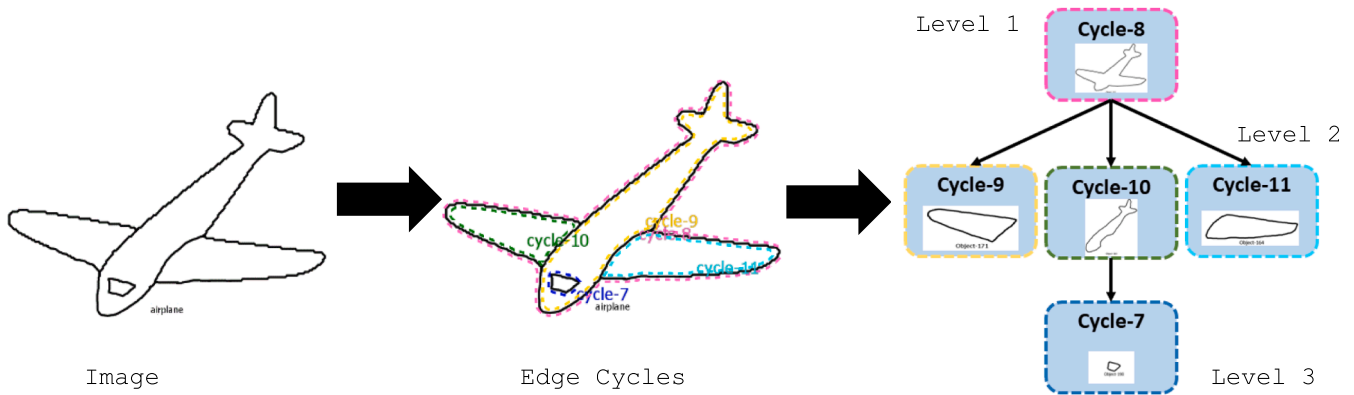


Fig. 3. Automatic hierarchical decomposition of a sketch based on edge cycles.

performance to LeNet-5, the classic baseline. To test data efficiency, we started with many fewer training samples.<sup>3</sup> Table 1 illustrates the results. We note that with fewer than 1,000 training examples, LeNet-5 performs at chance, thereby demonstrating the data-efficiency of analogical learning. Moreover, while LeNet-5 achieves quite high accuracy, it requires 20 passes through the entire set of training data, whereas Analogy requires only one. We note that adding more examples beyond 500 did not substantially improve the analogy results, so they asymptote at a lower accuracy than LeNet-5 achieves. There are two ways that the analogy results might be improved. First, there is no guarantee that the encoding strategy we used is optimal, and so enabling the system to do its own search for better encodings is an interesting possibility. The second way would be to use near-miss learning (McLure et al., 2011), since confusability can be an issue (see Chen, Rabkina, McLure, & Forbus, 2019 for details).

It has long been observed that human visual representations seem to involve hierarchical structure (e.g. Marr 1982; Palmer 1999). One reason why hierarchical representations are used is to manage complexity of processing – the SME algorithm is  $O(n^2 \log(n))$ , so there is a cost to larger representations. Consequently, we have developed a novel extension to analogical learning, *Part-based Hierarchical Analogical Learning* (PHAL), that uses hierarchical structure to improve recognition. The encoding process involves decomposing a sketch into multiple levels of description. It starts with the outer edge cycle and moves inward (Fig. 3). The decomposition process stops at level three because there are rarely more levels in hand-drawn sketches. Instead of one generalization pool per concept, there are three, one per level. Encoding involves decomposing a stimulus into edge cycles up to three levels, using CogSketch to encode the shapes and relationships between parts, and adding them to the generalization pool for that concept and level.<sup>4</sup> Classification of a novel stimulus starts with the same encoding and decomposition process to construct descriptions at all three levels. The classification process starts with Level 1 (coarsest), picking the K concepts whose average scores are the highest. Then, only within the K concepts found for the coarsest level, repeat at Level 2, keeping the top Q categories. Finally, within the Q concepts found for Level 2, use MAC/FAC to determine the top V candidates at Level 3. Evidence from all three levels is combined, along with a bias that each part probe should

Table 2  
Results on CBO dataset.

Method	Accuracy
CNN	5.26 %
ResNet50	10.53 %
ResNet50 (pretrained)	21.05 %
Analogy	29.47 %
PHAL	37.37 %

retrieve a distinct item from its gpool (i.e. it is preferable that the left wing and right wing of an aircraft match to different generalizations or outliers). We note that increasing the number of levels would significantly impact efficiency, by increasing the number of cases. The excessive granularity also led to less interpretable representations, hence three levels provided a practical and cognitively plausible balance between expressiveness and efficiency.

PHAL has been tested on two datasets to date (Chen, Forbus, Srinivasan, Chhaya, & Usher, 2023). The first is the TU Berlin dataset (Eitz et al., 2012), consisting of 20,000 hand-drawn sketches, with 80 sketches per category and covering 250 categories. PHAL achieves 69.85 % accuracy, which is competitive compared to baselines. The only systems which outperform it either do data augmentation to produce more training data or involve massive pre-training before fine-tuning on the TU Berlin dataset itself. By contrast, PHAL only uses the training data once, versus multiple epochs for the deep learning systems. On this dataset, we used 10 CPUs to encode sketches and just one to perform the hierarchical analogical learning – No GPUs or TPUs required, thereby demonstrating training efficiency as well.

To focus further on data-efficiency, we developed the Coloring Book Object dataset (Chen, Rabkina, McLure, & Forbus, 2019). It consists of images from open-license children’s coloring books, with only ten examples for each of 19 categories (animals and everyday objects). Being from coloring books, the images cover a wide range – some of the animals have hats, for example, and are drawn in a variety of styles and poses. We reproduce the results from Chen, Forbus, Srinivasan, Chhaya, & Usher (2023) in Table 2. The CNN was using the LeNet architecture, and both it and ResNet50, when trained from scratch, were at chance. Pretrained ResNet50 was better than chance, but was beaten by traditional analogical learning. PHAL does best of all, handily beating traditional analogy. This provides additional evidence for data-efficient learning via analogy, especially when exploiting hierarchical structure.

We note that convolutional neural networks and transformers can capture hierarchical features, typically progressing from low-level to high-level representations. However, these features are often entangled and lack the human-like interpretable structures needed for explainability. They also are subject to hallucinations, unlike our methods.

<sup>3</sup> MNIST has 60,000 training images and 10,000 test images. Some deep learning systems generate more training data by resizing, resampling, and other manipulations.

<sup>4</sup> A subtlety involves texture, e.g. the pepperoni on a pizza or the patterns on the back of a turtle. As explained in Chen, Forbus, Srinivasan, Chhaya, & Usher (2023), the edges and edge cycles within a description are added to a special generalization pool, with those assimilated into a generalization being replaced by a summary of that generalization. Thus texture is also detected via analogy.

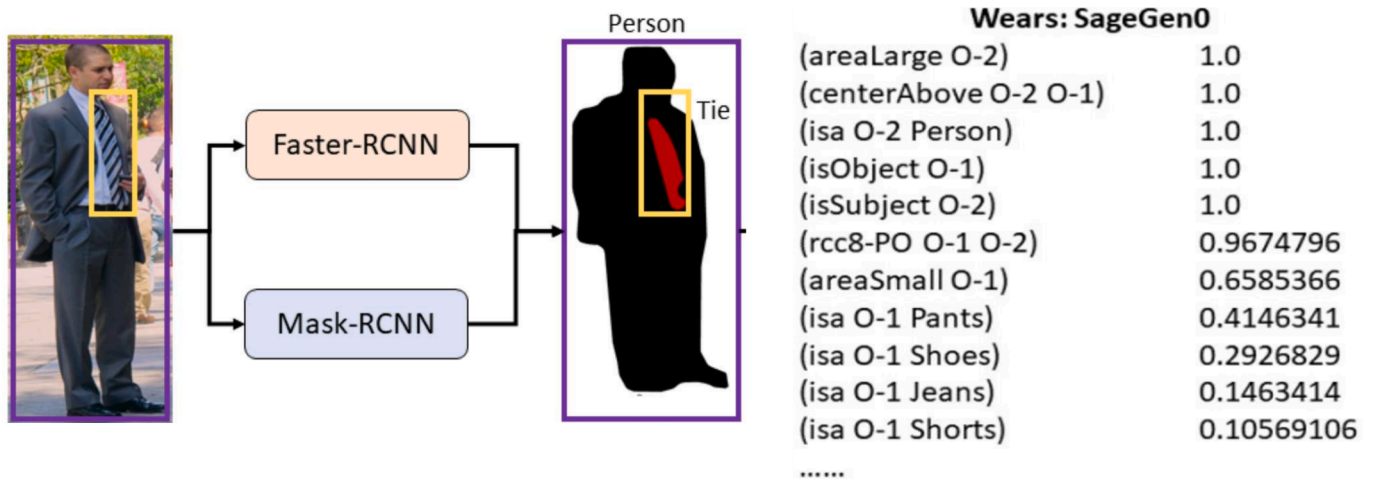


Fig. 4. An example of “wears”, the bounding boxes for the entities, and one of the generalizations that SAGE constructs for the concept.

## 5.2. Visual relationship detection

Sketches are of course simpler than natural images. How does the HPS approach fare on natural images? To explore this, we used a standard Visual Relationship Detection dataset (Lu et al., 2016), which consists of 5,000 images involving 100 object categories and 70 predicates. As described in (Chen & Forbus, 2021), given ground truth bounding boxes and categories (the PREDT task), analogical learning over HPS representations had 52.38 for both recall@50 and recall@100, putting it in the middle of the pack compared to prior systems, and hence competitive. We note that, as usual, the analogical learner only examined each training item once, while the deep learning systems required between seven and 30 epochs, thereby demonstrating training efficiency. (As before, the analogy learner only used CPU machines, whereas the deep learning systems use GPUs or TPUs.) While we have not been able to find recall@1 for other systems, the analogy learner had 32.26 for recall@1.<sup>5</sup> We also used analogical learning in the RELDT task, using off-the-shelf components (e.g. Faster-RCNN), see Chen & Forbus (2021) for details. Again, our results were competitive, with recall = 16.12 @50, 18.41 @ 100, and 8.91 @1 (see Chen & Forbus, 2021 for details).

In addition to being data and training-efficient, analogical learning produces inspectable models. Because the visual relationships computed by CogSketch are inspired by human vision, the generalizations produced tend to be interpretable. Fig. 4 shows an example of an image illustrating the “wears” relation, the bounding boxes computed for it, and one of the generalizations that SAGE constructs from the training data. (A generalization pool is a disjunctive concept description, consisting of different generalizations and outliers.) For example, in this generalization that which is being worn overlaps the person, is small compared to the size of the person, and is most likely pants, shoes, jeans, or shorts. This illustrates how the combination of the symbolic relational descriptions and probabilities produced by SAGE are inspectable.

## 5.3. Human action recognition

Seeing images in isolation is a simplification of vision, with video being a closer approximation since organisms exist in dynamic environments. Here we examine learning from Kinect video, a special case focused on human actions. Kinect sensors produce skeleton descriptions

<sup>5</sup> The problem with @N measures is that they don’t provide insight to what is most commonly the long-term desired behavior, i.e. returning the right answer, or @1. We suggest that @1 measures should be reported whenever @N where N > 1 is used, to better gauge progress.

of human motion, based on data from depth cameras, with 3D coordinates at every joint. We developed a pipeline that generates qualitative representations of such video in the form of video *sketch graphs*, a sequence of snapshots, akin to panels in a comic strip. Each snapshot describes the motion that is happening over some interval of time corresponding to a qualitative state. This temporal decomposition is accomplished with QSRLib (Gatsoulis et al., 2016), a library of qualitative spatial relations and calculi that takes streams of coordinates in, and segments them based on changes in the qualitative spatial relationships being tracked. The original 3D coordinates are projected into a front and side view, over which qualitative spatial relationships are computed for temporal segmentation.

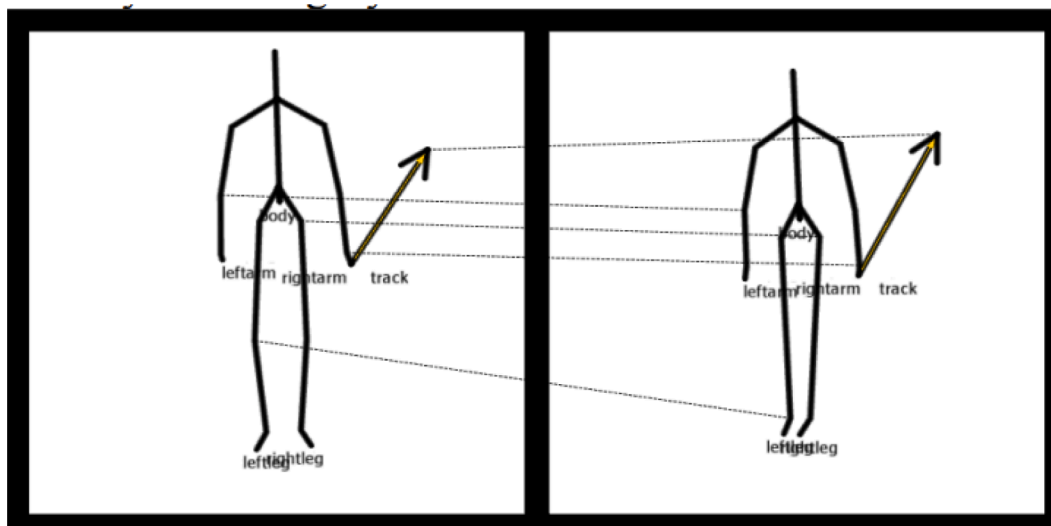
There are multiple data sets of videos of simple human actions, like swiping left, clapping, throwing, jogging, walking, etc. Here we used Florence 3D Action (Seidenari et al., 2013), UTD-MHAD (Chen et al., 2015), and UTKinect-Action3D (Xia et al., 2012) datasets. As reported in (Chen & Forbus, 2018), this approach was competitive with others, but did not perform as well with actions involving substantial noise. However, unlike the other approaches, ours provides inspectable models that can be used to explain a system’s answers, as Fig. 5 illustrates.

## 6. In progress: understanding diagrams

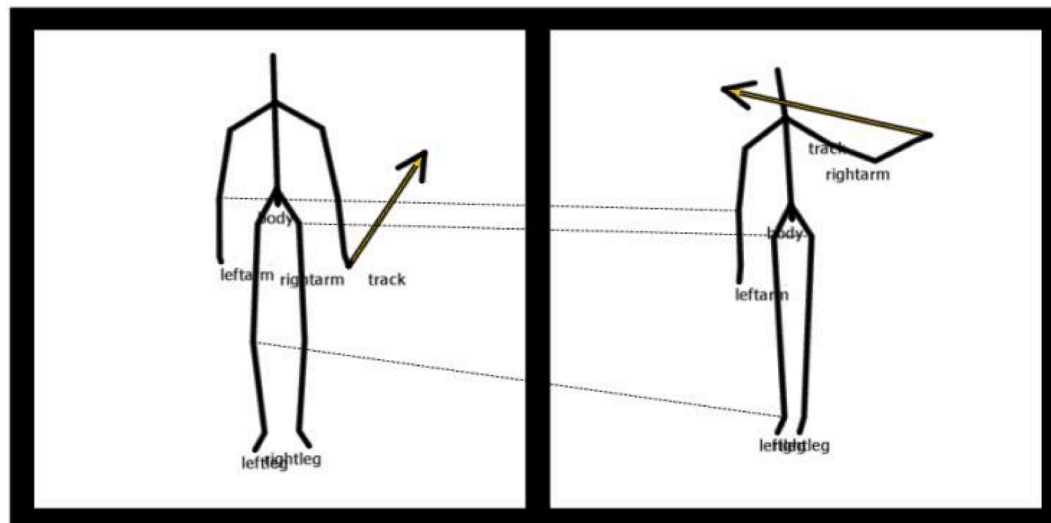
Learning from instructional media, like text with diagrams, is an important task for bootstrapping AI systems. This involves understanding conventions used for depicting things in diagrams and sketches. For example, the area inside a circle intended to depict the Earth is solid, while the area inside a circle depicting an orbit is not. Lockwood et al. (2008) casts this as the problem of *conceptual segmentation*, i.e. assigning conceptual interpretations to regions and edges within a diagram or sketch. We are currently building on this approach as a further way to extend the HPS in the direction of broader conceptual interpretations, beyond just object recognition or simple visual relationship detection. We are using as a testbed the AI2 Diagrams dataset (Kembhavi et al., 2016) which focuses on types of diagrams commonly used in science educational materials.

The first challenge is to generate an initial visual segmentation as the starting point for conceptual segmentation. Here we are relying on Meta’s SAM (Kirillov et al., 2023), which is capable of carving up diagrams into useful intermediate regions with reasonable accuracy. Fig. 6 illustrates.

Conceptual segmentation involves determining what regions are and



### Perfect Qualitative Match



### Qualitative Mismatch

Fig. 5. Sketch graphs can provide explanations about classification decisions, since qualitative mismatches reduce similarity.

how they relate to each other,<sup>6</sup> using a process of abductive reasoning. Abduction can be facilitated by having appropriate intermediate relationships. Based on its success in handling sketches in STEM domains (Forbus et al., 2017a), we think CogSketch has a set of distinctions particularly suited for diagrams. CogSketch has three kinds of glyphs: (1) *Entity glyphs* represent objects or categories, like fish or butterflies. All of the examples used in this paper so far are entity glyphs. (2) *Relation glyphs* represent binary relationships between other glyphs, such as that one animal type consumes another. Typically, these are drawn by arrows but need not be. CogSketch uses proximity to conjecture which entities might be the two arguments to the relation. (3) *Annotation glyphs* represent additional information about another glyph,

e.g. the flow rate of CO<sub>2</sub> between the atmosphere and ocean or a direction of motion. Fig. 6 shows an example of a diagram from the dataset, the segmentation found by SAM, and the use of proximity to connect a relation glyph (i.e. one of the arrows) to entity glyphs. We are using off-the-shelf OCR to detect text labels. Visual overlap between a SAM region (or set of regions) and an OCR-detected region provides an explanation of those original SAM regions as a text label. To what does the label apply? Again, visual proximity is a common heuristic. Note that all three types of glyphs can be labeled, although relation glyphs and annotation glyphs often aren't, with conceptual context being needed to disambiguate them. For example, an arrow in a life cycle diagram often represents sequentially in stages, whereas an arrow in a food web diagram often represents a consumption relationship.

While this research is still in progress, we summarize some pilot results. Using the AI2 Foodweb dataset (Krishnamurthy et al., 2016), which includes a diverse set of 490 foodweb examples, we used the Hybrid Primal Sketch Processor to analyze these diagrams. In this

<sup>6</sup> It can also involve determining that some regions do not have relevant conceptual interpretations, as some diagrams include decorations intended to hold reader interest.

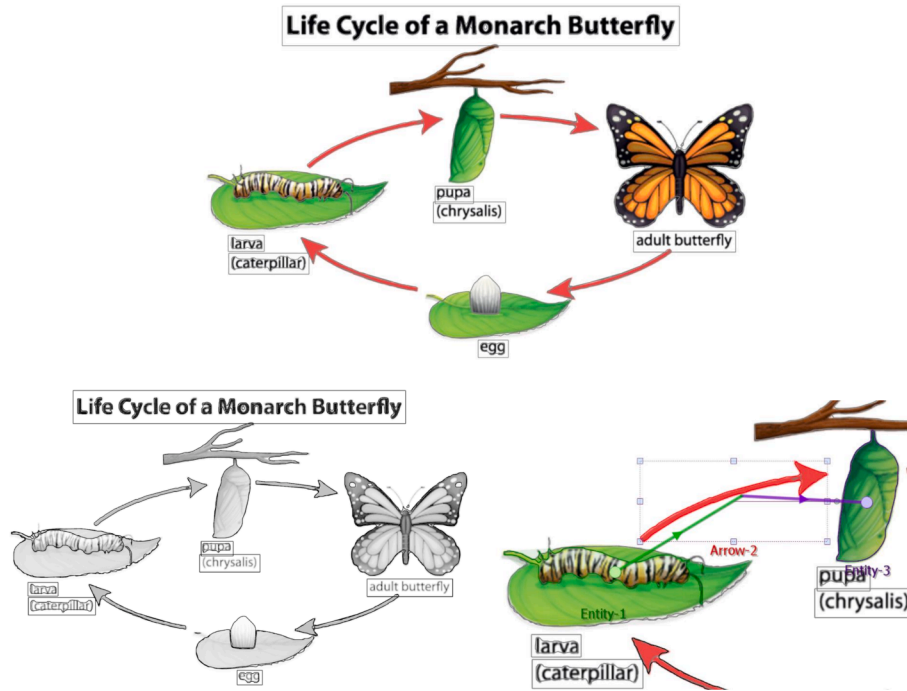


Fig. 6. Example of a life cycle diagram. The top is the original diagram. The bottom left is the segmentation constructed by SAM. The bottom right illustrates how CogSketch interprets a relation glyph by selecting nearby glyphs as its arguments.

Table 3

Results on consumption relation extractions. Three variants are tested for each VLM. + UNITS means including recognized text labels. + Ours means including extracted relations from our pipeline.

Method	Size	Precision	Recall	F1
Ours	<1B	57.8 %	54.4 %	56.0 %
LLaVA-CoT	11B	62.5 %	47.0 %	53.7 %
+ UNITS		60.4 %	48.1 %	53.5 % (−0.2 %)
+ Ours		53.8 %	58.5 %	56.1 % (+2.4 %)
CogVLM 2	19B	58.6 %	34.9 %	43.7 %
+ UNITS		52.6 %	32.2 %	40.0 % (−3.7 %)
+ Ours		56.4 %	52.1 %	54.1 % (+10.4 %)
LLaVA-Next	34B	36.5 %	19.2 %	25.2 %
+ UNITS		46.2 %	34.2 %	39.3 % (+14.1 %)
+ Ours		46.8 %	60.3 %	52.7 % (+27.5 %)
Qwen2-VL-72B	72B	75.4 %	51.7 %	61.4 %
+ UNITS		73.7 %	54.1 %	62.4 % (+1.0 %)
+ Ours		74.7 %	69.4 %	71.9 % (+10.5 %)
Gemini-1.5-flash	–	77.9 %	65.9 %	71.4 %
+ UNITS		79.4 %	69.6 %	74.2 % (+2.8 %)
+ Ours		75.0 %	75.9 %	75.4 % (+4.0 %)
Gemini-1.5-pro	–	74.8 %	68.6 %	71.6 %
+ UNITS		75.6 %	70.7 %	73.1 % (+1.5 %)
+ Ours		73.7 %	76.0 %	74.9 % (+3.3 %)
Gemini-2.0-flash	–	80.9 %	72.8 %	76.7 %
+ UNITS		81.5 %	72.0 %	76.5 % (−0.2 %)
+ Ours		78.2 %	76.4 %	77.3 % (+0.6 %)
GPT-4o-mini	–	67.7 %	41.5 %	51.4 %
+ UNITS		63.8 %	43.9 %	52.0 % (+0.6 %)
+ Ours		66.9 %	66.3 %	66.6 % (+15.2 %)
GPT-4o	–	74.2 %	60.5 %	66.7 %
+ UNITS		76.1 %	63.4 %	69.2 % (+2.5 %)
+ Ours		82.0 %	70.6 %	75.9 % (+9.2 %)

ensemble, UNITS (Kil et al., 2023) detects all occurrences of text in the diagram, providing both the text bounding boxes and the recognized text strings, which are converted to annotation glyphs. OWLv2 (Minderer et al., 2023) detects instances of specified types of objects and outputs their bounding boxes. We use it as the first-pass arrow detection by specifying “arrow”. SAM (Kirillov et al., 2023) produces segments of

arrows from bounding boxes, which are converted to relation glyphs. We also use SAM to carve up the entire image and segment all general objects, remove those that overlap with the detected textual labels and arrows by UNITS and OWLv2, filter further for arrows (some segments might represent arrows not detected by OWLv2) with geometrical heuristics, and use the remaining segments as organism illustrations, which are converted to entity glyphs. A third source of arrow detection is added specifically for the kind of thin arrows that are often ignored by OWLv2 detection and full-image SAM segmentation. We find line segments from high-quality edge detection with TEED (Soria Poma et al., 2023), followed by Probabilistic Hough Transform (Galambos et al., 1999), and check whether these line segments form thin arrows based on spatial and geometrical heuristics. All three sources of arrows are converted to relation glyphs.

With all three types of glyphs, the connections between them (the source and target entities of a consumption relation and the label tied to the entity illustration) are determined by visual proximity. Analyzing these connections gives predictions of the prey-predator, which are matched against the ground truth as a measure of our approach’s effectiveness. To compare our training-free pipeline, we used pre-trained large Vision-Language Models (VLM) like LLaVA-NeXT (Liu et al., 2024), CogVLM 2 (Wang et al., 2023), and Qwen2-VL (Wang et al., 2024a) as they achieved impressive zero-shot capabilities for a range of vision tasks, including diagram understanding. Some closed-source VLMs, including the GPT-4o series (OpenAI, 2024) and the Gemini family (Google, 2024), and a VLM variant trained with visual Chain-of-Thought, LLaVA-CoT (Xu et al., 2024) are also compared. The result is shown in Table 3.

Our approach uses deep learning modules of less than 1B parameters combined but achieves competitive performance to VLMs of many times the parameter size. This parameter efficiency might imply that the explicit visual analysis we performed could better capture the spatial understanding that VLMs seem to be struggling with (Wang et al., 2024b). In another experiment, we combine the consumption relation predictions from our pipeline into the textual prompt of VLMs as additional reference, which leads to improved performance across all VLMs we used (Table 3). Moreover, the performance gains are much greater

than those with only recognized text (by UNITS) as additional references, which confirms that the improved performance is not due to simple augmentation with text recognition. The common pattern of improved VLM performance is that the precision slightly changes, but the recall substantially increases, which might be due to the top-down constraints from VLM's background knowledge. Overall, our pipeline and VLMs seem to be complementary in understanding the diagram and extracting consumption relations.

In Lockwood et al. (2008), conceptual segmentation operated over sketches drawn using CogSketch, so each glyph represented an intended visual entity. Automatic segmentation invariably has errors, so merging and splitting regions is needed. A tradeoff we will explore involves the relative roles of conceptual and visual constraints in the interpretation process. The top-down expectations imposed by text captions (e.g. "Life Cycle of..." in Fig. 6) are quite strong, for instance. We are expanding CogSketch's ontology with more concepts for depiction (e.g. visual containers to aggregate groups or display an entity in its setting), and setting up visual routines to interpret novel diagrams.

## 7. Conclusions and future work

Given the complexity of perception, finding the appropriate intermediate representations is crucial. This paper argues that CogSketch's glyph representation provides a notion of visual entity that is useful for perception beyond just sketch understanding. The ability to decompose shapes within a glyph, and represent relationships between glyphs, enables analogical learning to be used for recognition of entities and properties of scenes. The evidence presented here supports our hypothesis in sketch recognition (Section 4.1), natural scenes (i.e. the visual-relation detection task of Section 4.2), and video of human actions (Section 4.3). We note that performance depends on the ability of the low-level vision algorithms to deliver reasonable quality results for the type of images being analyzed, with natural images remaining the most difficult. As computer vision research on recognizers and segmenters improves, the quality of the hybrid primal sketch should improve, and hence so should algorithms that rely upon it. By using human-like qualitative relational representations, the hybrid primal sketch representations are inspectable and hence debuggable, providing clues to developers and/or learning algorithms as to where improvements in the ensemble are needed.

We plan three lines of future work. The first is to expand the exploration of diagram understanding, since it stress-tests the integration of conceptual and visual knowledge. The second is to make the ensemble operations more automatic. Like visual routines (Ullman 1984) and spatial routines (Lovett, 2012), we believe there is learnable procedural knowledge used to guide the application of ensemble operations to new situations and tasks. Uncovering the nature of these orienting computations is high on our agenda. Finally, we plan to continue to expand the components in the ensemble to tackle more complex problems involving natural imagery and video.

## CRedit authorship contribution statement

**Kenneth D. Forbus:** Writing – review & editing, Writing – original draft, Supervision, Software, Project administration, Methodology, Funding acquisition, Conceptualization. **Kezhen Chen:** Writing – review & editing, Writing – original draft, Validation, Software, Investigation, Data curation, Conceptualization. **Wangcheng Xu:** Writing – review & editing, Visualization, Software, Investigation, Data curation, Conceptualization. **Madeline Usher:** Writing – review & editing, Visualization, Validation, Software, Data curation.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

## Acknowledgements

This research was supported by the Machine Learning, Reasoning, and Intelligence Program of the US Office of Naval Research, Grant N00014-23-1-2294.

## Data availability

All data sets are already available from their authors

## References

- Amorapant, P., Kranjec, A., Bromberger, B., Lehet, M., Widick, P., Woods, A., & Chatterjee, A. (2012). Language, perception, and the schematic representation of spatial relations. *Brain and Language*, 120, 226–236.
- Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological review*, 94(2), 115.
- Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 679–698.
- Chang, M. D. (2016). *Capturing Qualitative Science Knowledge with Multimodal Instructional Analogies*. Evanston, Illinois: Northwestern University, Department of Electrical Engineering and Computer Science. Doctoral dissertation.
- Chen, C., Jafari, R., and Kehtarnavaz, N. 2015. UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor. Image Processing (ICIP), 2015 IEEE International Conference on (pp. 168-172). IEEE.
- Chen, K. and Forbus, K.D. (2018). Action Recognition from Skeleton Data via Analogical Generalization over Qualitative Representations. *Proceedings of AAAI 2018*.
- Chen, K., Rabkina, I., McLure, M., & Forbus, K. (2019) Human-like Sketch Object Recognition. *Proceedings of AAAI 2019*.
- Chen, K., Forbus, K., Gentner, D., Hespos, S., & Anderson, E. (2020). Simulating Infant Visual Learning by Comparison: An Initial Model. *Proceedings of CogSci 2020*.
- Chen, K. & Forbus, K. (2021) Visual Relation Detection using Hybrid Analogical Learning. *Proceedings of AAAI 2021*.
- Chen, K., Forbus, K., Srinivasan, B., Chhaya, N., & Usher, M. (2023). Sketch Recognition via Part-based Hierarchical Analogical Learning. Proceedings of IJCAI 2023, Macao, China.
- Cohn, A. G., Bennett, B., Gooday, J., & Gotts, N. M. (1997). Qualitative spatial representation and reasoning with the region connection calculus. *Geoinformatica*, 1, 275–316.
- Doumas, L., Hummel, J., & Sandhofer, C. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, 115, 1–43.
- Eitz, M., Hayes, J., & Alexa, M. (2012). How do humans sketch objects? *ACM Transactions on Graphics*.
- Forbus, K. (2019). *Qualitative Representations: How People Reason and Learn about the Continuous World*. MIT Press.
- Forbus, K., Gentner, D., & Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 131–205.
- Forbus, K., & Lovett, A. (2021). Same/different in visual reasoning. *Current Opinion in Behavioral Sciences*, 37, 63–68. <https://doi.org/10.1016/j.cobeha.2020.09.008>
- Forbus, K. D., & Hinrichs, T. (2017). *Analogy and Qualitative Representations in the Companion Cognitive Architecture*. AI Magazine.
- Forbus, K. D., Usher, J., Lovett, A., Lockwood, K., & Wetzel, J. (2011). CogSketch: Sketch understanding for cognitive science research and for education. *Topics in Cognitive Science*, 3(4), 648–666.
- Forbus, K. D., Chang, M., McLure, M., & Usher, M. (2017). The cognitive science of sketch worksheets. *Topics in Cognitive Science*. <https://doi.org/10.1111/tops.12262>
- Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending SME to handle large-scale cognitive modeling. *Cognitive Science*, 1152–1201.
- Forbus, K.D., Garnier, B., Tikoff, B., Marko, W., Usher, M. & McLure, M. (2018). Sketch Worksheets in STEM Classrooms: Two Deployments. Deployed Application Prize paper. *Proceedings of IAAI 2018*.
- Friedman, S., Forbus, K., & Sherin, B. (2018). Representing, running, and revising Mental Models: A Computational Model. *Cognitive Science*, 1110–1145. <https://doi.org/10.1111/cogs.12574>
- Galambos, C., Kittler, J., & Matas, J. (1999). Progressive probabilistic Hough transform for line detection. *Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Vol. 1)*, 554–560.
- Gatsoulis, Y., Alomari, M., Burbridge, C., Doudrup, C., Duckworth, P., Lightbody, P., Hanheide, M., Hawes, N., Hogg, D.C. and Cohn, A.G. 2016. QSRlib: a software library for online acquisition of Qualitative Spatial Relations from Video. *QR 2016*.
- Gemini Team, Google. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. ArXiv, abs/2403.05530.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155–170.
- He, K.; Gkioxari, G.; Dollár, Piotr.; and Girshick, R. 2017. Mask R-CNN. arXiv: 1703.06870.
- Huang, W., Liu, H., Guo, M., & Gong, N. (2024) Visual Hallucinations of Multimodal Large Language Models. <https://doi.org/10.48550/arXiv.2310.03744>.

- Hummel, J., & Holyoak, K. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104, 427–466.
- Kandaswamy, S. and Forbus, K. (2012). Modeling Learning of Relational Abstractions via Structural Alignment. *Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci)*. Sapporo, Japan.
- Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajjishirzi, H., & Farhadi, A. (2016). A Diagram is Worth a Dozen Images. ArXiv, abs/1603.07396.
- Krenski, K. & Selinger, P. (2003) Potrace. <http://potrace.sourceforge.net/>.
- Kil, T. H., Kim, S., Seo, S., Kim, Y., & Kim, D. (2023). Towards unified scene text spotting based on sequence generation. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15223–15232).
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, C., Gustafson, L., Xiao, T., Whitehead, S. Berg, A., Lo, W., Dollar, P. Girschick, 2023. Segment Anything. arXiv:2304.0264, <https://doi.org/10.48550/arXiv.2304.02643>.
- Krishnamurthy, J., Tafjord, O., & Kembhavi, A. (2016). Semantic Parsing to Probabilistic Programs for Situated Question Answering. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 160–170).
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *IEEE*, 2278–2324.
- Liu, H., Li, C., Li, Y., Lee, Y. (2023) Improved Baselines with Visual Instruction Tuning. <https://doi.org/10.48550/arXiv.2310.03744>.
- Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., & Lee, Y.J. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.
- Lockwood, K., Lovett, A., Forbus, K., Deghani, M., & Usher, J. (2008). Automatic Interpretation of Depiction Conventions in Sketched Diagrams. *Proceedings of the Eurographics Workshop on Sketch-Based Interfaces and Modeling*.
- Lovett, A. (2012). *Spatial Routines for Sketches: A Framework for Modeling Spatial Problem-solving*. Evanston, Illinois: Northwestern University, Department of Electrical Engineering and Computer Science. Doctoral dissertation.
- Lowet, A., Firestone, C., & Scholl, B. (2018). Seeing structure: Shape skeletons modulate perceived similarity. *Attent Percept Psychophysics*, 80, 1278–1289.
- Lu, C., Krishna, R., Bernstein, M., Li F.F., 2016. Visual Relationship Detection with Language Priors. *European Conference on Computer Vision*, Springer, Cham.
- Lu, J., Rao, J., Chen, K., Guo, X., Zhang, Y., Sun, B., Yang, C., & Yang, J. (2024). Evaluation and Enhancement of Semantic Grounding in Large Vision-Language Models. AAAI.
- Marr, D. (1982; 2010) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press.
- Manivannan, C., Krukar, J., & Schwering, A. (2022). Spatial generalization in sketch maps: A systematic classification. *Journal of Environmental Psychology*, 83, Article 101851.
- Manivannan, C., Krukar, J., & Schwering, A. (2024). An algorithmic approach to detect generalization in sketch maps from sketch map alignment. *PLoS One*, 19(6). <https://doi.org/10.1371/journal.pone.0304696>
- McLure, M. D., Friedman, S. E., Lovett, A. and Forbus, K. D. (2011). Edge-cycles: A qualitative sketch representation to support recognition. *Proceedings of the 25th International Workshop on Qualitative Reasoning*. Barcelona, Spain.
- Minderer, M., Gritsenko, A., & Houlsby, N. (2023) Scaling Open-Vocabulary Object Detection. *Proceedings of NeurIPS 2023*.
- Mitchell, M. (1993). *Analogy-making as perception: A computer model*. MIT Press.
- Moratz, R. (2017). Qualitative Spatial reasoning. In S. Shekhar (Ed.), *Encyclopedia of GIS*. Springer International.
- Museros, Lledó, Escrib, M., 2004. A Qualitative Theory for Shape Representation and Matching for Design, pp. 858-862.
- OpenAI. (2024). GPT-4o System Card. ArXiv, abs/2410.21276.
- Palmer, S. (1999). *Vision Science: Photons to Phenomenology*. MIT Press.
- Ullman, S. (1984). Visual routines. *Cognition*, 18, 97–159. [https://doi.org/10.1016/0010-0277\(84\)90023-4](https://doi.org/10.1016/0010-0277(84)90023-4)
- Ren, S.; He, K.; Girshick, R.; Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*.
- Seidenari, L., Varano, V., Berretti, S., Bimbo, A., & Pala, P. (2013). Recognizing actions from depth cameras as weakly aligned multi-part bag-of poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 479–485).
- Soria Poma, X., Li, Y., Rouhani, M., & Sappa, A. D. (2023). Tiny and efficient model for the edge detection generalization. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops* (pp. 1356–1365).
- Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., Xu, J., Xu, B., Li, J., Dong, Y., Ding, M., & Tang, J. (2023). CogVLM: Visual Expert for Pretrained Language Models. ArXiv, abs/2311.03079.
- Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., & Lin, J. (2024). Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. ArXiv, abs/2409.12191.
- Wang, J., Ming, Y., Shi, Z., Vineet, V., Wang, X., & Joshi, N. (2024). Is A Picture Worth A Thousand Words? Delving Into Spatial Reasoning for Vision Language Models. ArXiv, abs/2406.14852.
- Wu, J., Li, X., Xu, S., Yuah, H., Ding, H., Yang, Y., Li, X., Zhang, J., Tong, Y., Jiang, X., Ghanem, B., & Tao, D. (2024). Towards Open Vocabulary Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Xia, L., Chen, C., & Aggarwal, J. (2012). In *View invariant human action recognition using histograms of 3D joints* (pp. 20–27). IEEE.
- Xu, G., Jin, P., Li, H., Song, Y., Sun, L., & Yuan, L. (2024). LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. ArXiv, abs/2411.10440.
- Zhang, H., You, H., Dufter, P. Zhang, B., Chen, C., Chen, H. Fu, T., Wang, W. Chang, S., Gan, Z., & Yang, Y. (2024) Ferret-v2: An Improved Baseline for Referring and Grounding with Large Language Models. arXiv:2404.07973v1.
- Zhang, T., & Suen, Y. (1984). A fast parallel algorithm for thinning digital patterns. *CACM*.