

# Inferring Graphics Programs from Images

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1

## 2 1 Introduction

3 How could an agent go from noisy, high-dimensional perceptual input to a symbolic, abstract object,  
 4 like a computer program? Here we consider this problem within a graphics program synthesis domain.  
 5 We develop an approach for converting natural images, such as hand drawings, into executable source  
 6 code for drawing the original image. The graphics programs in our domain draw simple figures like  
 7 those found in machine learning papers (see Fig.1).

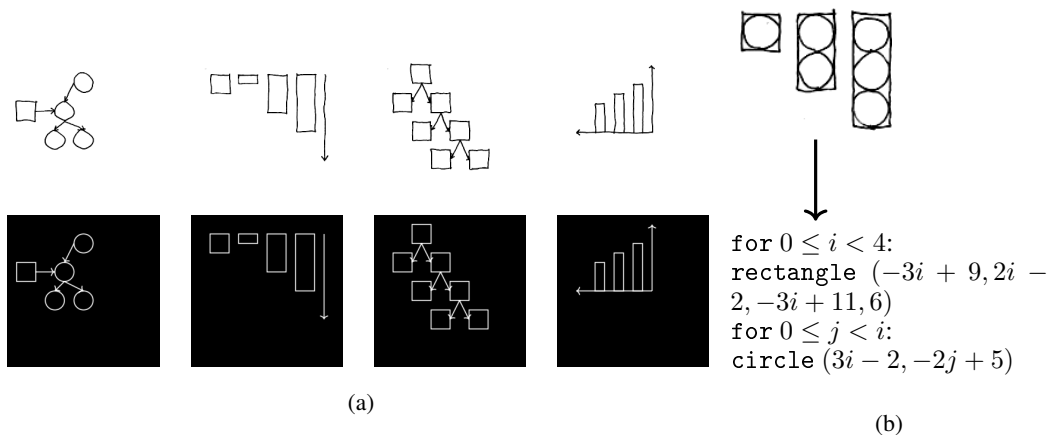


Figure 1: (a): Model learns to convert hand drawings (top) into  $\text{\LaTeX}$  (bottom). (b) Synthesizes high-level *graphics program* from hand drawing.

8 High dimensional perceptual input may seem ill matched to the abstract semantics of a programming  
 9 language. But programs with constructs like recursion or iteration produce a simpler *execution trace*  
 10 of primitive actions; for our domain the primitive actions are drawing commands. Our hypothesis is  
 11 that the execution trace of the program is better aligned with the perceptual input, and that the trace  
 12 can act as a kind of bridge between perception and programs. We test this hypothesis by developing  
 13 a model that learns to map from an image to the execution trace of the graphics program that drew  
 14 it. With the execution trace in hand, we can bring to bear techniques from the program synthesis  
 15 community to recover the latent graphics program. This family of techniques, called *constraint-based*  
 16 *program synthesis* [1], work by modeling a set of possible programs inside of a constraint solver, like  
 17 a SAT or SMT solver [2]. These techniques excel at uncovering high-level symbolic structure, but  
 18 are not well equipped to deal with real-valued perceptual inputs.

19 We develop a hybrid architecture for inferring graphics programs. Our approach uses a deep neural  
20 network infer an execution trace from an image; this network recovers primitive drawing operations  
21 such as lines, circles, or arrows, along with their parameters. For added robustness, we use the  
22 deep network as a proposal distribution for a stochastic search over execution traces. Finally, we  
23 use techniques in the program synthesis community to recover the program from its trace. The  
24 program synthesizer discovers constructs like loops and geometric operations like reflections and  
25 affine transformations. [This paragraph is all about making things a bit more specific, so you really  
26 need more specifics about program synth here.]

27 Each of these three components – the deep network, the stochastic search, the program synthesizer –  
28 confers its own advantages. From the deep network, we get a fast system that can recover plausible  
29 execution traces in about a minute [A minute seems slow to me, for deep net inference. Are you  
30 talking about training time, here, or...?]. From the stochastic search we get added robustness;  
31 essentially, the stochastic search can correct mistakes made by the deep network’s proposals. From  
32 the program synthesizer, we get abstraction: our system recovers coordinate transformations, for  
33 loops, and subroutines, which are useful for downstream tasks and can help correct some mistakes of  
34 the earlier stages. [I wonder if this would work even better as a bulleted list...]

## 35 2 Related work

36 Our work bears resemblance to the Attend-Infer-Repeat (AIR) system, which learns to decompose an  
37 image into its constituent objects [3]. AIR learns an iterative inference scheme which infers objects  
38 one by one and also decides when to stop inference; this is similar to our approach’s first stage, which  
39 parses images into program execution traces. Our approach further produces interpretable, symbolic  
40 programs which generate those execution traces. The two approaches also differ in their architectures  
41 and training regimes: AIR learns a recurrent auto-encoding model via variational inference, whereas  
42 our parsing stage learns an autoregressive-style model from randomly-generated (execution trace,  
43 image) pairs. Finally, while AIR was evaluated on multi-MNIST images and synthetic 3D scenes, we  
44 focus on parsing and interpreting hand-drawn sketches.

45 Our image-to-execution-trace parsing architecture builds on prior work on controlling procedural  
46 graphics programs [4]. Given a program which generates random 2D recursive structures such as  
47 vines, that system learns a structurally-identical “guide program” whose output can be directed, via  
48 neural networks, to resemble a given target image. We adapt this method to a different visual domain  
49 (figures composed of multiple objects), using a broad prior over possible scenes as the initial program  
50 and viewing the execution trace through the guide program as a symbolic parse of the target image.  
51 We then show how to synthesize higher-level programs from these execution traces.

52 In the computer graphics literature, there have been other systems which convert sketches into  
53 procedural representations. One uses a convolutional network to match a sketch to the output of a  
54 parametric 3D modeling system [5]. Another uses convolutional networks to support sketch-based  
55 instantiation of procedural primitives within an interactive architectural modeling system [6]. Both  
56 systems focus on inferring fixed-dimensional parameter vectors. In contrast, we seek to automatically  
57 learn a structured, programmatic representation of a sketch which captures higher-level visual patterns.

58 Prior work has also applied sketch-based program synthesis to authoring graphics programs. In  
59 particular, Sketch-n-Sketch presents a bi-directional editing system in which direct manipulations  
60 to a program’s output automatically propagate to the program source code [7]. We see this work  
61 as complementary to our own: programs produced by our method could be provided to a Sketch-n-  
62 Sketch-like system as a starting point for further editing.

63 [Do you also want to cite your own work on “Unsupervised Learning by Program Synthesis” here?]

## 64 3 Neural architecture for inferring drawing execution traces

65 We developed a deep network architecture for efficiently inferring a execution trace,  $T$ , from an  
66 image,  $I$ . Our model constructs the trace one drawing command at a time. When predicting the next  
67 drawing command, the network takes as input the target image  $I$  as well as the rendered output of  
68 previous drawing commands. Intuitively, the network looks at the image it wants to explain, as well  
69 as what it has already drawn. It then decides either to stop drawing or proposes another drawing

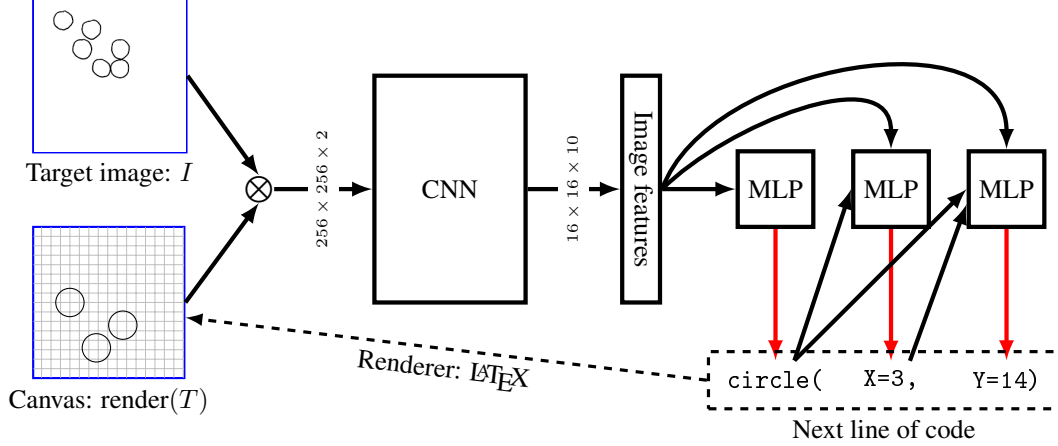


Figure 2: Our neural architecture for inferring the execution trace of a graphics program from its output. **Blue**: network inputs. **Black**: network operations. **Red**: samples from a multinomial. Typewriter font: network outputs. Renders snapped to a  $16 \times 16$  grid, illustrated in gray. [Thoughts on improving this figure: (1) Convnet diagrams typically show the sequence of layers, if possible (space might not permit it here, but those thin arrows just aren’t doing it for me).]

70 command to add to the execution trace; if it decides to continue drawing, the predicted primitive is  
 71 rendered to its “canvas” and the process repeats.

72 Figure 2 illustrates this architecture. We first pass a  $256 \times 256$  target image and a rendering of the  
 73 trace so far to a convolutional network – these two inputs are represented as separate channels for  
 74 the convnet. Given the features extracted by the convnet, a multilayer perceptron then predicts a  
 75 distribution over the next drawing command to add to the trace. We predict the drawing command  
 76 token-by-token, and condition each token both on the image features and on the previously generated  
 77 tokens. For example, the network first decides to emit the `circle` token conditioned on the image  
 78 features, then it emits the  $x$  coordinate of the circle conditioned on the image features and the `circle`  
 79 token, and finally it predicts the  $y$  coordinate of the circle conditioned on the image features, the  
 80 `circle` token, and the  $x$  coordinate. [There are some more details that are important to provide  
 81 about this architecture, though possibly in an Appendix: the functional form(s) of the probability  
 82 distributions over tokens, the network layer sizes, which MLPs share parameters, etc.]

83 The distribution over the next drawing command factorizes:

$$\mathbb{P}_{\theta}[t_1 t_2 \cdots t_K | I, T] = \prod_{k=1}^K \mathbb{P}_{\theta}[t_k | f_{\theta}(I, \text{render}(T)), \{t_j\}_{j=1}^{k-1}] \quad (1)$$

84 where  $t_1 t_2 \cdots t_K$  are the tokens in the drawing command,  $I$  is the target image,  $T$  is an execution trace,  
 85  $\theta$  are the parameters of the neural network, and  $f_{\theta}(\cdot, \cdot)$  is the image feature extractor (convolutional  
 86 network). The distribution over execution traces factorizes as:

$$\mathbb{P}_{\theta}[T | I] = \prod_{n=1}^{|T|} \mathbb{P}_{\theta}[T_n | I, T_{1:(n-1)}] \times \mathbb{P}_{\theta}[\text{STOP} | I, T] \quad (2)$$

87 where  $|T|$  is the length of execution trace  $T$ , and the STOP token is emitted by the network to signal  
 88 that the execution trace explains the image.

89 We train the network by sampling execution traces  $T$  and target images  $I$  for randomly generated  
 90 scenes, and maximizing (2) wrt  $\theta$  by gradient ascent. Training does not require backpropagation across  
 91 the entire sequence of drawing commands: drawing to the canvas ‘blocks’ the gradients, effectively  
 92 offloading memory to an external visual store. In a sense, this model is like an autoregressive variant  
 93 of AIR [3] without attention.

94 This network suffices to “derender” images like those shown in Figure 3. We can perform a beam  
 95 search decoding to recover what the network thinks is the most likely execution trace for images

<code>circle(<math>x, y</math>)</code>	Circle at $(x, y)$
<code>rectangle(<math>x_1, y_1, x_2, y_2</math>)</code>	Rectangle with corners at $(x_1, y_1)$ & $(x_2, y_2)$
<code>LINE(<math>x_1, y_1, x_2, y_2</math>, arrow <math>\in \{0, 1\}</math>, dashed <math>\in \{0, 1\}</math>)</code>	Line from $(x_1, y_1)$ to $(x_2, y_2)$ , optionally with an arrow and/or dashed
<code>STOP</code>	Finishes execution trace inference

Table 1: The deep network in (2) predicts drawing commands, shown above.

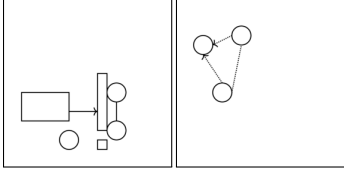


Figure 3: Network is trained to infer execution traces for randomly generated figures like the two shown above.

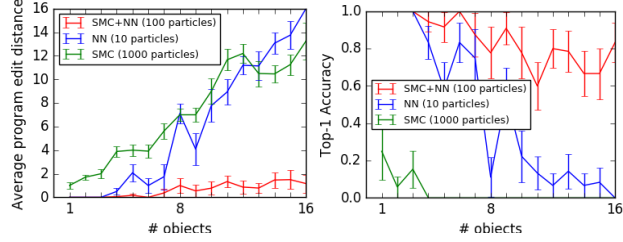


Figure 4: Using the model to parse latex output. The model is trained on diagrams with up to 8 objects. As shown above it generalizes to scenes with many more objects. Neither the stochastic search nor the neural network are sufficient on their own.

like these, recovering traces maximizing  $\mathbb{P}_\theta[T|I]$ . But, if the network makes a mistake (predicts an incorrect line of code), it has no way of recovering from the error. In order to derender an image with  $n$  objects, it must correctly predict  $n$  drawing commands – so its probability of success will decrease exponentially in  $n$ , assuming it has any nonzero chance of making a mistake. For added robustness as  $n$  becomes large, we treat the neural network outputs as proposals for a SMC sampling scheme. For the SMC sampler, we use pixel wise distance as a surrogate for a likelihood function. The SMC sampler is designed to produce samples from the distribution  $\propto L(I|\text{render}(T))\mathbb{P}_\theta[T|I]$ , where  $L(\cdot|\cdot) : \text{image}^2 \rightarrow \mathcal{R}$  uses the distance between two images as a proxy for a likelihood.

Figure 4 compares the neural network with SMC against the neural network by itself or SMC by itself. Only the combination of the two passes a critical test of generalization: when trained on images with  $\leq 8$  objects, it successfully parses scenes with many more objects than the training data.

### 3.1 Generalizing to hand drawings

A practical application of our neural network is the automatic conversion of hand drawings into a subset of  $\text{\LaTeX}$ . We train the model to generalize to hand drawings by introducing noise into the renderings of the training target images. We designed this noise process to introduce the kinds of variations found in hand drawings (figure 6). Our neurally-guided SMC procedure used pixel-wise distance as a surrogate for a likelihood function ( $L(\cdot|\cdot)$  in section 3). But pixel-wise distance fares poorly on hand drawings, which never exactly match the model’s renders. So, for hand drawings, we *learned* a surrogate likelihood function. Our learned  $L(\cdot|\cdot)$  is a convolutional network that we train to predict the distance between two traces conditioned upon their renderings. Formally we train our likelihood surrogate to approximate:

$$-\log L(\text{render}(T_1)|\text{render}(T_2)) \approx |T_1 - T_2| + |T_2 - T_1| \quad (3)$$

We drew 100 figures by hand; see figure ???. These were drawn reasonably carefully but not perfectly. Because our model assumes that objects are snapped to a  $16 \times 16$  grid, we made the drawings on graph paper.

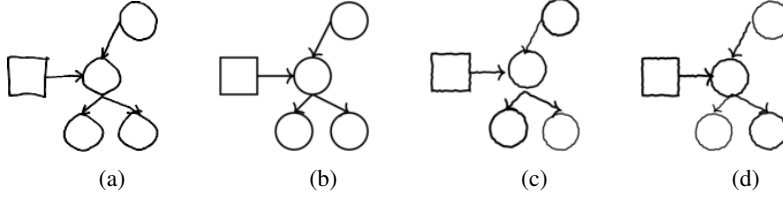


Figure 5: (a): a hand drawing. (b): Rendering of the parse our model infers for (a). We can generalize to hand drawings like these because we train the model on images corrupted by a noise process designed to resemble the kind of noise introduced by hand drawings - see (c) & (d) for noisy renderings of (b).

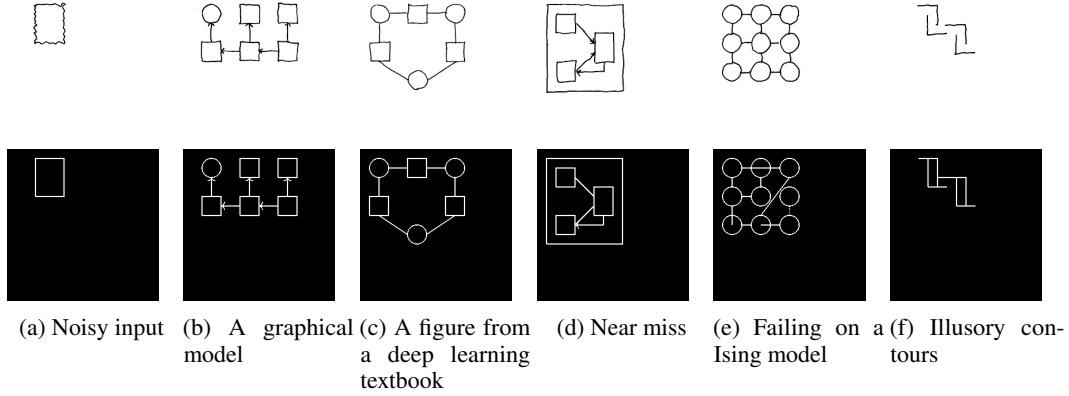


Figure 6: Example drawings above model outputs. See also Fig. 1

## 120 4 Synthesizing graphics programs from execution traces

121 Although the execution trace of a graphics program describes the parts of a scene, it fails to encode  
 122 higher-level features of the image, such as repeated motifs, symmetries or reflections. A *graphics*  
 123 *program* better describe structures like these, and we now take as our goal to synthesize simple  
 124 graphics programs from their execution traces.

125 We constrain the space of allowed programs by writing down a context free grammar over a space of  
 126 programs. Although it might be desirable to synthesize programs in a Turing-complete language like  
 127 Lisp or Python, a more tractable approach is to specify what in the program languages community  
 128 is called a Domain Specific Language (DSL). Our DSL (Table 2) encodes prior knowledge of what  
 129 graphics programs tend to look like.

Program	→	Command; ...; Command
Command	→	circle(Expression, Expression)
Command	→	rectangle(Expression, Expression, Expression, Expression)
Command	→	LINE(Expression, Expression, Expression, Expression, Boolean, Boolean)
Command	→	for( $0 \leq \text{Var} < \text{Expression}$ ) { Program }
Command	→	REFLECT(Axis) { Program }
Expression	→	$Z * \text{Var} + Z$
Var	→	A free (unused) variable
Z	→	an integer
Axis	→	$X = Z$
Axis	→	$Y = Z$

Table 2: Grammar over graphics programs. We allow loops (for), vertical/horizontal reflections (REFLECT), and affine transformations ( $Z * \text{Var} + Z$ ).

130 Given the DSL and a trace  $T$ , we want to recover a program that both evaluates to  $T$  and, at the same  
 131 time, is the “best” explanation of  $T$ . For example, we might prefer more general programs or, in the  
 132 spirit of Occam’s razor, prefer shorter programs. We wrap these intuitions up into a cost function  
 133 over programs, and seek the minimum cost program consistent with  $T$ :

$$\text{program}(T) = \arg \min_{\substack{p \in \text{DSL} \\ p \text{ evaluates to } T}} \text{cost}(p) \quad (4)$$

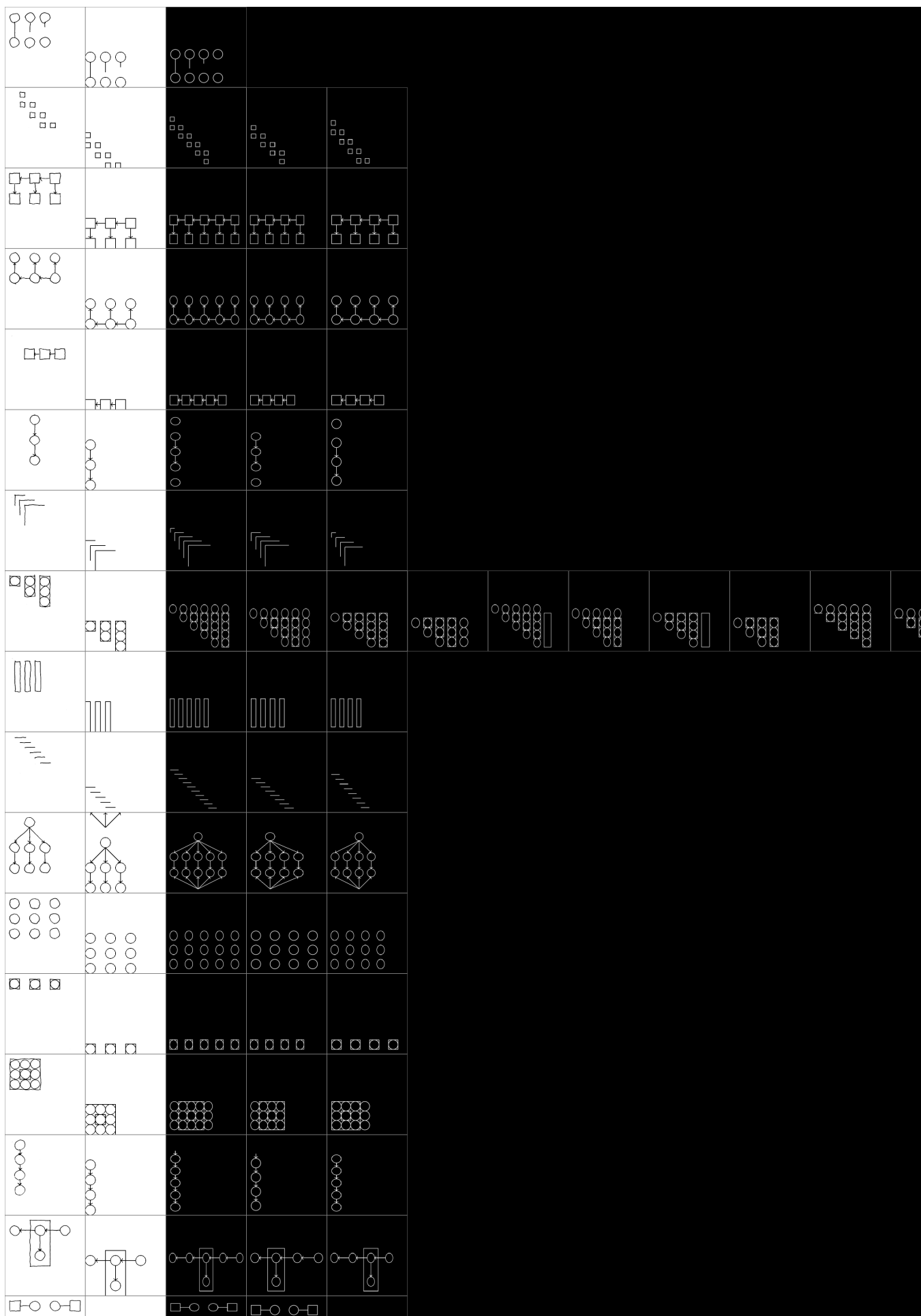
134 We define the cost of a program to be the number of statements it contains, where a statement is a  
 135 “Command” in Table 2.

136 The constrained optimization problem in equation 4 is intractable in general, but there exist efficient-  
 137 in-practice tools for finding exact solutions to program synthesis problems like these. We use the  
 138 state-of-the-art Sketch tool [1]. Describing Sketch’s program synthesis algorithm is beyond the scope  
 139 of this paper; see supplement. At a high level, Sketch takes as input a space of programs, along with  
 140 a specification of the program’s behavior and optionally a cost function. It translates the synthesis  
 141 problem into a constraint satisfaction problem, and then uses a SAT solver to find a minimum cost  
 142 program satisfying the specification. In exchange for not having any guarantees on how long it will  
 143 take to find a minimum cost solution, it comes with the guarantee that it will always find a globally  
 144 optimal program.

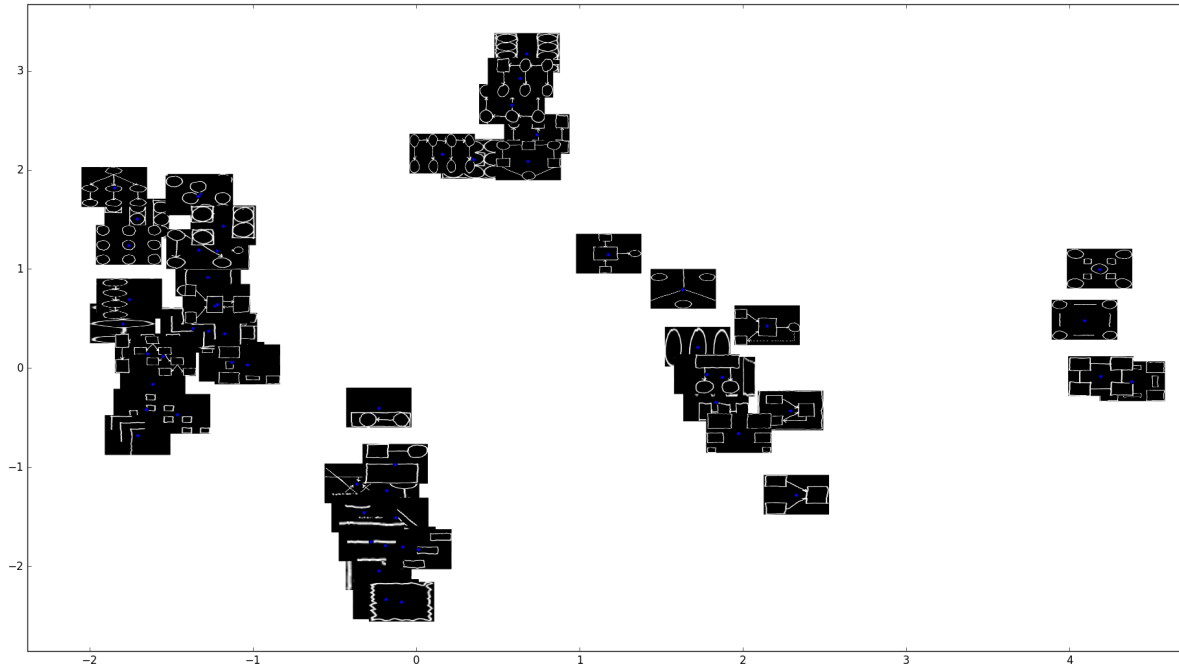
145 Why synthesize a graphics program, if the execution trace already suffices to recover the objects in  
 146 an image? Within our domain of hand-drawn figures, graphics program synthesis has several uses:

#### 147 **4.1 Extrapolating figures**

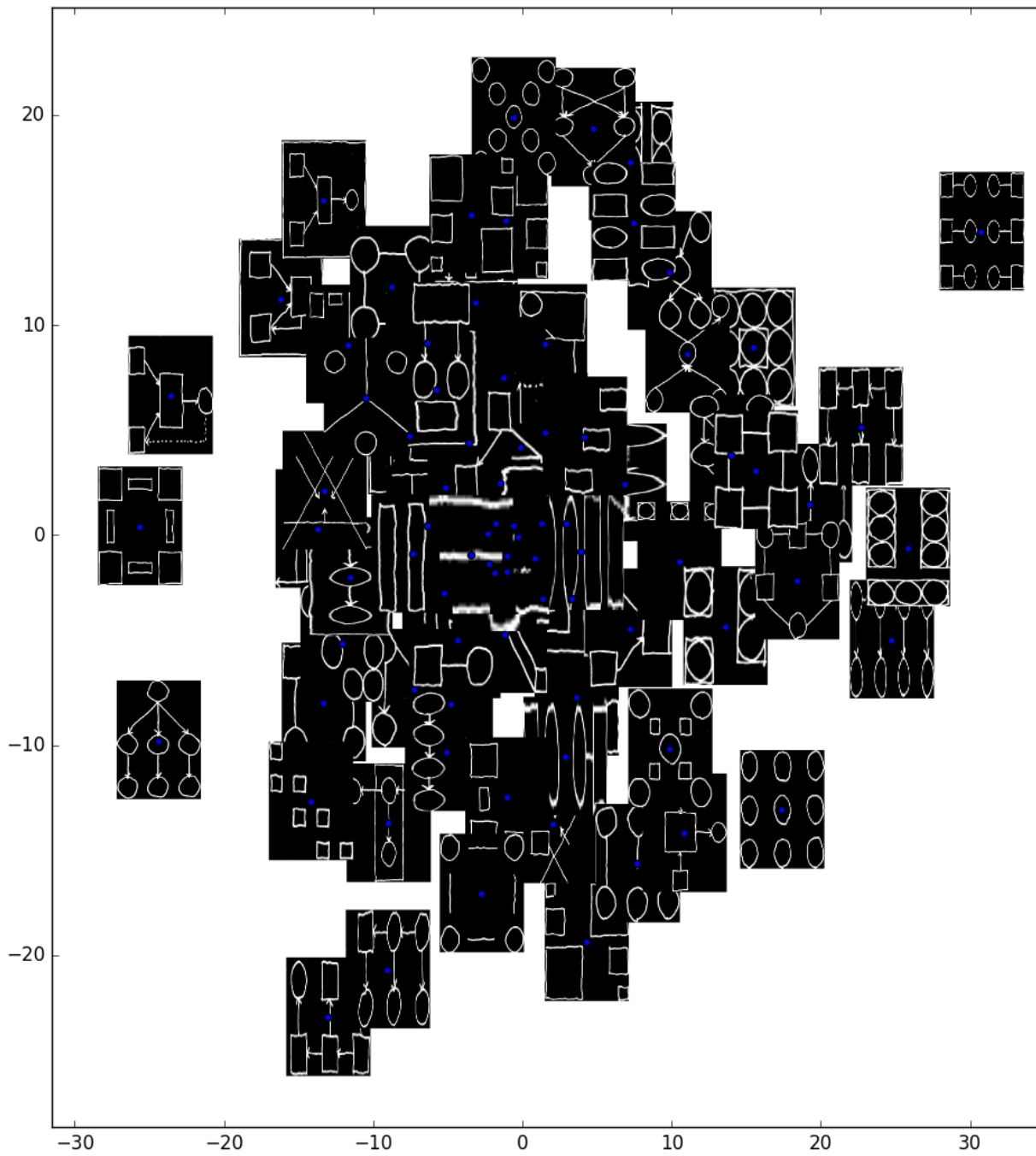
148 Having access to the source code of a graphics program facilitates coherent, high-level edits to the  
 149 figure generated by that program. For example, we can change all of the circles to squares, were  
 150 make all of the lines be dashed. We can also **extrapolate** figures by increasing the number of times  
 151 that loops are executed.



154 Similarity metric in program space:







### 158 4.3 Correcting errors made by the neural network

159 The program synthesizer can help correct errors from the neural network by favoring execution traces  
 160 which lead to more concise or general programs. For example, one generally prefers figures with  
 161 neatly aligned objects over figures whose parts are slightly misaligned – and nice alignment lends  
 162 itself to short programs. Similarly, figures often have repeated parts, which the program synthesizer  
 163 might model as a reflection. So, if one is considering several candidate traces proposed by the neural  
 164 network, we might prefer traces whose best programs have desirable features like being short or  
 165 having iterated structures.

166 Concretely we implemented the following scheme: the neurally guided sampling scheme of section 3  
 167 for image  $I$  produces (samples) candidate traces,  $\mathcal{F}(I)$ . Instead of predicting the most likely trace in  
 168  $\mathcal{F}(I)$  according to the neural network, we can take into account the programs that best explain the  
 169 traces. Writing  $\hat{T}(I)$  for the trace the model predicts for image  $I$ ,

$$\hat{T}(I) = \arg \max_{T \in \mathcal{F}(I)} L(I|\text{render}(T)) \times \mathbb{P}_\beta[\text{program}(T)] \quad (5)$$

170 where  $\mathbb{P}_\beta[\cdot]$  is a prior probability distribution over programs parameterized by  $\beta$ . This is equivalent  
 171 to doing MAP inference in a generative model where the program is first drawn from  $\mathbb{P}_\beta[\cdot]$ , then the  
 172 program is executed deterministically, and then we observe a noisy version of the program’s output,  
 173 where  $L$  is the noise model.

174 Given a corpus of graphics program synthesis problems with annotated ground truth traces (so pairs  
 175 of  $(I, T)$ ), we find a maximum likelihood estimate of  $\beta$ :

$$\beta^* = \arg \max_{\beta} \mathbb{E} \left[ \log \frac{\mathbb{P}_\beta[\text{program}(T)] \times L(I|\text{render}(T))}{\sum_{T' \in \mathcal{F}(I)} \mathbb{P}_\beta[\text{program}(T')] \times L(I|\text{render}(T'))} \right] \quad (6)$$

176 where the expectation is taken both over the model predictions and the  $(I, T)$  pairs in the training  
 177 corpus. We define  $\mathbb{P}_\beta[\cdot]$  to be a log linear distribution  $\propto \exp(\beta \cdot \phi(\text{program}))$ , where  $\phi(\cdot)$  is a feature  
 178 extractor for programs. We can extract a few basic features of a program, like its size or how many  
 179 loops it has, and use these features to help predict whether a trace is the correct explanation for an  
 180 image.

181 [Seems like you’re still fleshing this part out, but I’ll give my feedback anyway: (1) This subsection  
 182 could really use a motivational introduction, e.g. “The program synthesizer can help correct errors/bad  
 183 proposals from the neural network by favoring execution traces which lead to more concise/general  
 184 programs.” (2) The image likelihood function should probably be introduced sooner, when you talk  
 185 about SMC/beam search. (3) Where does  $\theta$  come from? Is it set by hand? Learned? (4) How does  
 186 Equation 4 get used? Is this a modification to the beam search objective / SMC posterior? If so, it’d  
 187 be great to have set up the version without it in an earlier section, and then be able to refer to this as a  
 188 small modification of the previous equation.]



## 6 Preliminary Synthesis results

### References

- [1] Armando Solar Lezama. *Program Synthesis By Sketching*. PhD thesis, EECS Department, University of California, Berkeley, Dec 2008.
- [2] Leonardo De Moura and Nikolaj Bjørner. Z3: An efficient smt solver. In *Tools and Algorithms for the Construction and Analysis of Systems*, pages 337–340. Springer, 2008.
- [3] SM Eslami, N Heess, and T Weber. Attend, infer, repeat: Fast scene understanding with generative models. arxiv preprint arxiv:..., 2016. URL <http://arxiv.org/abs/1603.08575>.
- [4] Daniel Ritchie, Anna Thomas, Pat Hanrahan, and Noah Goodman. Neurally-guided procedural models: Amortized inference for procedural graphics programs using neural networks. In *Advances In Neural Information Processing Systems*, pages 622–630, 2016.
- [5] Haibin Huang, Evangelos Kalogerakis, Ersin Yumer, and Radomir Mech. Shape synthesis from sketches via procedural models and convolutional networks. *IEEE transactions on visualization and computer graphics*, 2017.
- [6] Gen Nishida, Ignacio Garcia-Dorado, Daniel G. Aliaga, Bedrich Benes, and Adrien Bousseau. Interactive sketching of urban procedural models. *ACM Trans. Graph.*, 35(4), 2016.
- [7] Brian Hempel and Ravi Chugh. Semi-automated svg programming via direct manipulation. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, pages 379–390, New York, NY, USA, 2016. ACM.