

# PLAD: Learning to Infer Shape Programs with Pseudo-Labels and Approximate Distributions

R. Kenny Jones  
Brown University

Homer Walke  
UC Berkeley

Daniel Ritchie  
Brown University

## Abstract

*Inferring programs which generate 2D and 3D shapes is important for reverse engineering, editing, and more. Training models to perform this task is complicated because paired (shape, program) data is not readily available for many domains, making exact supervised learning infeasible. However, it is possible to get paired data by compromising the accuracy of either the assigned program labels or the shape distribution. Wake-sleep methods use samples from a generative model of shape programs to approximate the distribution of real shapes. In self-training, shapes are passed through a recognition model, which predicts programs that are treated as ‘pseudo-labels’ for those shapes. Related to these approaches, we introduce a novel self-training variant unique to program inference, where program pseudo-labels are paired with their executed output shapes, avoiding label mismatch at the cost of an approximate shape distribution. We propose to group these regimes under a single conceptual framework, where training is performed with maximum likelihood updates sourced from either Pseudo-Labels or an Approximate Distribution (PLAD). We evaluate these techniques on multiple 2D and 3D shape program inference domains. Compared with policy gradient reinforcement learning, we show that PLAD techniques infer more accurate shape programs and converge significantly faster. Finally, we propose to combine updates from different PLAD methods within the training of a single model, and find that this approach outperforms any individual technique.*

## 1. Introduction

Having access to a procedure which generates a visual datum reveals its underlying structure, facilitating high-level manipulation and editing by a person or autonomous agent. Thus, inferring such programs from visual data is an important problem. In  $\mathbb{R}^2$ , inferring shape programs has applications in the design of diagrams, icons, and other 2D graphics. In  $\mathbb{R}^3$ , it has applications in reverse engineering of

CAD models, procedural modeling for 3D games, and 3D structure understanding for autonomous agents.

We formally define shape program inference as obtaining a latent program  $\mathbf{z}$  which generates a given observed shape  $\mathbf{x}$ . We model  $p(\mathbf{z}|\mathbf{x})$  with deep neural networks that train over a distribution of real shapes in order to amortize the cost of shape program inference on unseen shapes (e.g. a test set). This is a challenging problem: it is a structured prediction problem whose output is high-dimensional and can feature both discrete and continuous components (i.e. program control flow vs. program parameters). Nevertheless, learning  $p(\mathbf{z}|\mathbf{x})$  becomes tractable provided that one has access to paired  $(\mathbf{X}, \mathbf{Z})$  data (i.e. a dataset of shapes and the programs which generate them) [34]. Unfortunately, while shape data is increasingly available in large quantities [2], these shapes do not typically come with their generating program.

To circumvent this data problem, researchers have typically *synthesized* paired data by generating synthetic programs and pairing them with the shapes they output [25, 31]. However, as there is typically significant distributional mismatch between these synthetic shapes and “real” shapes from the distribution of interest,  $S^*$ , various techniques must be employed to fine-tune  $p(\mathbf{z}|\mathbf{x})$  models towards  $S^*$ .

A number of these fine-tuning strategies attempt to directly propagate gradients from geometric similarity measures back to  $p(\mathbf{z}|\mathbf{x})$ . When the program executor is not a black-box, it may be possible to do this by implementing a differentiable relaxation of its behavior [16], but this requires knowledge of its functional form. One can also try learning a differentiable proxy of the executor’s behavior [31], but this approximation introduces errors. Moreover, as shape programs typically involve many discrete structural decisions, training such a model end-to-end is usually infeasible in many domains. Thus, many prior works often resort to general-purpose policy gradient reinforcement learning [8, 25], which treats the program executor as a (non-differentiable) black-box. The downside of this strategy is that RL is notoriously unstable and slow to converge.

In this paper, we study a collection of methods that create

(shape, program) data pairs used to train  $p(\mathbf{z}|\mathbf{x})$  models with maximum likelihood estimation (MLE) updates while treating the program executor as a black-box. As discussed, ground-truth (shape, program) pairs are often unavailable, so these techniques must make compromises in how they formulate paired data. In wake-sleep, a generative model  $p(\mathbf{z})$  is trained to convergence on alternating cycles with respect to  $p(\mathbf{z}|\mathbf{x})$ . When training  $p(\mathbf{z}|\mathbf{x})$ , paired data can be created by sampling from  $p(\mathbf{z})$ . Each program label  $\mathbf{z}$  is valid with respect to its associated  $\mathbf{x}$  shape, but there is often a distributional mismatch between the generated set of shapes,  $\mathbf{X}$ , and shapes from the target distribution,  $S^*$ . In self-training, one uses  $p(\mathbf{z}|\mathbf{x})$  to infer latent  $\mathbf{z}$ 's for unlabeled input  $\mathbf{x}$ 's; these  $\mathbf{z}$ 's then become "pseudo-labels" which are treated as ground truth for another round of supervised training. In this paradigm, there is no distributional shift between  $\mathbf{X}$  and  $S^*$ , but each  $\mathbf{z}$  is only an approximately correct label with respect to its paired  $\mathbf{x}$ .

We observe that shape program inference has a unique property that makes it especially well-suited for self-training: the distribution  $p(\mathbf{x}|\mathbf{z})$  is known *a priori*—this is a delta distribution defined by the program executor. When using a model  $p(\mathbf{z}|\mathbf{x})$  to infer a program  $\mathbf{z}$  from some shape  $\mathbf{x}^*$  of interest, one can use this executor to produce a shape  $\mathbf{x}$  that is consistent with the program  $\mathbf{z}$ : in the terminology of self-training,  $\mathbf{z}$  is guaranteed to be the "correct label" for  $\mathbf{x}$ . However, similar to wake-sleep, formulating  $\mathbf{X}$  as shape executions produced by model inferred programs can cause a distributional shift between  $\mathbf{X}$  and  $S^*$ . Since this variant of self-training involves executing the inferred latent program  $\mathbf{z}$ , we call this procedure latent execution self-training (LEST).

As all of the aforementioned fine-tuning regimes use either Pseudo-Labels or Approximate Distributions to formulate (shape, program) pairs, we group them under a single conceptual framework: PLAD. We evaluate PLAD methods experimentally, using them to fine-tune shape program inference models in multiple shape domains: 2D and 3D constructive solid geometry (CSG), and assembly-based modeling with ShapeAssembly, a domain-specific language for structures of manufactured 3D objects [14]. We find that PLAD training regimes offer substantial advantages over the de-facto approach of policy gradient reinforcement learning, achieving better shape reconstruction performance while requiring significantly less computation time. Further, we explore combining training updates from a mixture of PLAD methods, and find that this approach leads to better performance compared with any individual method. Code for our method and experiments can be found at <https://github.com/rkjones4/PLAD>.

In summary, our contributions are:

1. Proposing the PLAD conceptual framework to group a family of related self-supervised learning techniques for shape program inference.

2. Introducing latent execution self-training, a PLAD method, to take advantage of the unique properties of the shape program inference problem.
3. Experiments across multiple 2D and 3D shape program inference domains, demonstrating that (i) fine-tuning under PLAD regimes outperforms policy gradient reinforcement learning and (ii) combining PLAD methods is better than any individual technique.

## 2. Related Work

Program synthesis is a broad field that has employed many techniques throughout its history. A program synthesizer takes as input a domain-specific language (DSL) and a specification; it outputs a program in the DSL that meets the specification. Machine learning has been used to improve performance on program synthesis tasks by e.g. performing a neurally-guided search over all possible programs or letting a recurrent network predict program text directly [1, 5, 7, 23, 30, 35].

In this work, we are interested in the sub-problem of *shape program inference*, which is a type of *visual program induction* problem [4]. In our case, the specification is a visual representation of an object which the inferred program's output must geometrically match. Here, we discuss prior work that has attacked this problem, organized by methodology used to learn  $p(\mathbf{z}|\mathbf{x})$ ; see Table 1 for an overview. As discussed, a common practice in this prior work is to start with a model that has been pretrained on synthetically generated (shape, program) pairs with supervised learning, and then perform fine-tuning towards a distribution of interest [8, 9, 20, 25, 31].

**Policy Gradient Reinforcement Learning** The most general method for fine-tuning a pretrained  $p(\mathbf{z}|\mathbf{x})$  is reinforcement learning: treating  $p(\mathbf{z}|\mathbf{x})$  as a policy network and using policy gradient methods [33]. The geometric similarity of the inferred program's output to its input is the reward function; the program executor  $p(\mathbf{x}|\mathbf{z})$  can be treated as a (non-differentiable) black-box. CSG-Net uses RL for fine-tuning [25, 26], as does other recent work on inferring CSG programs from input geometry [8]. While CSG-Net has been improved to allow it to converge without supervised pretraining [38], not starting from the supervised model results in worse performance. The main problem with policy gradient RL is its instability due to high variance gradients, leading to slow convergence. Like RL, PLAD methods treat the program executor as a black-box, but as we show experimentally, they converge faster and achieve better reconstruction performance.

**Differentiable Executor** If the functional form of the program executor  $p(\mathbf{x}|\mathbf{z})$  is known and differentiable, then the gradient of the reward with respect to the parameters of  $p(\mathbf{z}|\mathbf{x})$  can be computed, making policy gradient unnecessary. Shape programs are typically not fully differentiable, as

Method	Models	Black-Box $p(\mathbf{x} \mathbf{z})$ ?	$X = S^*$	Low variance, unbiased gradients
Policy gradient RL	$p(\mathbf{z} \mathbf{x})$	✓	✓	X
Differentiable executor	$p(\mathbf{z} \mathbf{x})$	X	✓	✓
Variational Bayes	$p(\mathbf{z} \mathbf{x}), p(\mathbf{z})$	X	✓	✓
Wake-sleep, EM	$p(\mathbf{z} \mathbf{x}), p(\mathbf{z})$	✓	X	✓
Self-training	$p(\mathbf{z} \mathbf{x})$	✓	✓	X
LEST	$p(\mathbf{z} \mathbf{x})$	✓	X	✓

Table 1. Comparison of different methods for fine-tuning  $p(\mathbf{z}|\mathbf{x})$ , in terms of the models that must be trained, if they treat the program executor as black-box, if their distribution of training shapes matches the distribution real shapes ( $X = S^*$ ), and if their loss gradients are unbiased with low-variance. The last three rows describe methods that fall under the PLAD framework.

they often involve discrete choices (e.g. which type of primitives to create). UCSGNet uses a differentiable relaxation to circumvent this issue [16]. Other work trains a differentiable network to approximate the behavior of the program executor [31], which introduces errors. PLAD regimes do not require the program executor to be differentiable, yet they perform better than other approaches (e.g. policy gradient RL) that share this desirable property.

**Generative Model Learning** Shape program inference has also been explored in the context of learning a generative model  $p(\mathbf{x}, \mathbf{z})$  of programs and the shapes they produce. The most popular approach for training such models is variational Bayes, in particular the variational autoencoder [18]. This method simultaneously trains a generative model  $p(\mathbf{x}, \mathbf{z})$  and a recognition model  $p(\mathbf{z}|\mathbf{x})$  by optimizing a lower bound on the marginal likelihood  $p(\mathbf{x})$ . When the  $\mathbf{z}$ ’s are shape programs, the program executor is  $p(\mathbf{x}|\mathbf{z})$ , so learning the generative model reduces to learning a prior over programs  $p(\mathbf{z})$ . Training such models with gradient descent requires that the executor  $p(\mathbf{x}|\mathbf{z})$  be differentiable. When this is not possible, the wake-sleep algorithm is a viable alternative [13]. This approach alternates training steps of the generative and recognition models, training one on samples produced by the other. Recent work has used wake-sleep for visual program induction [10, 12]. If one trains the generative model and the inference model to convergence before switching to training the other, this is equivalent to expectation maximization (viewed as alternating maximization [22]).

**Self-Training** Traditionally, self-training has been employed in weakly-supervised learning paradigms to increase the predictive accuracy of simple classification models [21, 24, 36]. Recently, renewed interest in self-training-inspired data augmentation approaches have demonstrated empirical performance improvements for neural models in domains such as large-scale image classification, machine translation, and speech recognition [11, 15, 39]. But while self-training has been shown to yield practical gains for some domains, for others it can actually lead to worse perfor-

mance, as training on too many incorrect pseudo-labels can cause learning to degrade [3, 28]. For self-training within the PLAD framework, the assigned pseudo-label for each example changes during fine-tuning whenever the inference model discovers a program that better explains the input shape; similar techniques have been proposed for learning programs that perform semantic parsing under the view of iterative maximum likelihood [19]. To our knowledge, self-training has not been applied for fine-tuning visual program inference models, likely because it is somewhat unintuitive to view a program as a “label” for a visual datum.

### 3. Method

In this section, we describe the PLAD framework: a conceptual grouping of fine-tuning methods for shape program inference models. Our formulation assumes three inputs: a training dataset of shapes from the distribution of interest,  $S^*$ , a program inference model,  $p(\mathbf{z}|\mathbf{x})$ , and a program executor that converts programs into shapes,  $E$ . Throughout this paper, we assume that the  $p(\mathbf{z}|\mathbf{x})$  passed as input has undergone supervised pretraining on a distribution of synthetically generated shapes. Methods within the PLAD framework return a fine-tuned  $p(\mathbf{z}|\mathbf{x})$  specialized to the distribution of interest from which  $S^*$  was sampled.

We depict the PLAD procedure both algorithmically and pictorially in Figure 1. To fine-tune  $p(\mathbf{z}|\mathbf{x})$  towards  $S^*$ , PLAD methods iterate through the following steps: (1) use  $p(\mathbf{z}|\mathbf{x})$  to find visually similar programs to  $S^*$ , (2) construct a dataset of shape and program pairs  $(\mathbf{X}, \mathbf{Z})$  using the inferred programs, and (3) fine-tune  $p(\mathbf{z}|\mathbf{x})$  with maximum likelihood estimation updates on batches from  $(\mathbf{X}, \mathbf{Z})$ . Through successive iterations, these steps bootstrap one another, forming a virtuous cycle: improvements to  $p(\mathbf{z}|\mathbf{x})$  create  $(\mathbf{X}, \mathbf{Z})$  pairs that more closely match the statistics of  $S^*$ , and training on better  $(\mathbf{X}, \mathbf{Z})$  pairs specializes  $p(\mathbf{z}|\mathbf{x})$  to  $S^*$ .

Methods that fall within the PLAD framework differ in how the paired  $(\mathbf{X}, \mathbf{Z})$  data is created within each round. We detail this process for wake-sleep (Section 3.1), self-training

---

**Input:**  $(S^*, p(z|x), E)$   
**Output:**  $p(z|x)$  fine-tuned on  $S^*$

```

 $P^{BEST} \leftarrow \{\}$ 
for Number of Rounds do
  // Update Best Programs
   $P^{BEST} \leftarrow \text{inferProgs}(p(z|x), S^*, P^{BEST})$ 
  // Create Training Data
  if Self-Train then
     $Z \leftarrow P^{BEST}$ 
     $X \leftarrow S^*$ 
  else if LEST then
     $Z \leftarrow P^{BEST}$ 
     $X \leftarrow \{E(z) \mid z \in Z\}$ 
  else if Wake-Sleep then
     $p(z) \leftarrow \text{trainGenerative}(P^{BEST})$ 
     $Z \leftarrow \text{sample}(p(z), |S^*|)$ 
     $X \leftarrow \{E(z) \mid z \in Z\}$ 
  end if
  // Train inference model
   $p(z|x) \leftarrow \text{trainMLE}(X, Z)$ 
end for

```

---

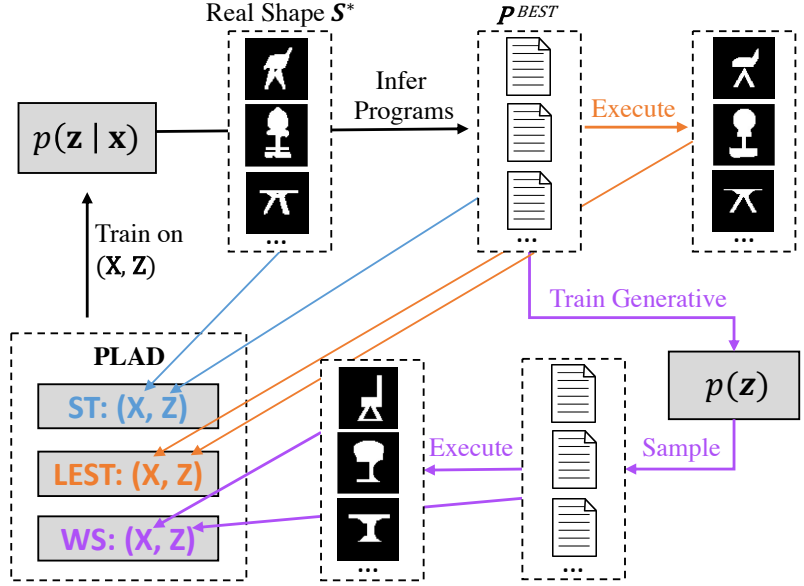


Figure 1. (Left) Pseudocode for fine-tuning shape program inference models,  $p(z|x)$ , towards a shape distribution of interest,  $S^*$ , with Pseudo-Labels and Approximate Distributions (PLAD). PLAD methods iterate through three steps: infer programs for  $S^*$  with  $p(z|x)$ , create a dataset of  $(X, Z)$  shape-program pairs, and train  $p(z|x)$  on batches from  $(X, Z)$ . Self-training, latent execution self-training, and wake-sleep differ in how  $(X, Z)$  is constructed. (Right) A visual illustration of the algorithm’s dataflow.

(Section 3.2), and latent execution self-training (Section 3.3). In Section 3.4 we explain our program inference procedure (inferProgs, Fig 1). Finally, in Section 3.5 we discuss how a single  $p(z|x)$  can be fine-tuned by multiple PLAD methods.

### 3.1. Wake-Sleep $(X, Z)$ Construction

Wake-sleep uses a generative model,  $p(z)$  to construct  $(X, Z)$ . In our implementation, we choose to model  $p(z)$  as a variational auto-encoder (VAE) [18], where the encoder consumes visual data and the decoder outputs a program. To create data for  $p(z)$ , we take the current best programs discovered for  $S^*$ ,  $P^{BEST}$ , and execute each program to form a set of shapes  $X^G$ .  $p(z)$  is then trained on pairs from  $(X^G, P^{BEST})$  in the typical VAE framework. Note that the design space for  $p(z)$  is quite flexible, for instance,  $p(z)$  can be trained without access to  $X^G$  if implemented with a program encoder.

Once  $p(z)$  has converged, we use it to sample  $|S^*|$  programs by decoding normally distributed random vectors. This set of programs becomes  $Z$ , and  $X$  is formed by executing each program in  $Z$ . In this set of  $(X, Z)$  programs,  $Z$  is always the correct label for  $X$ , so the gradient estimates during  $p(z|x)$  training will be low-variance and unbiased. However,  $X$  is not guaranteed to be close to  $S^*$ , it is only

an approximate distribution. Note though, that as  $P^{BEST}$  better approximates  $S^*$ , the distributional mismatch should become smaller, as long as the generative model has enough capacity to properly model  $p(z)$ .

### 3.2. Self-Training $(X, Z)$ Construction

Self-training constructs  $(X, Z)$  by assigning labels from the current best program set,  $P^{BEST}$ , to shape instances from  $S^*$ . Formally,  $X \leftarrow S^*$  and  $Z \leftarrow P^{BEST}$ . This framing maintains the nice property that  $X = S^*$ , so there will never be distributional mismatch between these two sets. The downside is that unless programs from  $P^{BEST}$  exactly recreate their paired shapes from  $S^*$  when executed, we know that the pseudo-labels from  $P^{BEST}$  are ‘incorrect’. From this perspective, we can consider gradient estimates that come from such  $(X, Z)$  pairs to be biased. However, as we will show experimentally, when  $X$  forms a good approximation to  $S^*$ , sourcing gradient estimates from these pseudo-labels leads to strong reconstruction performance.

### 3.3. LEST $(X, Z)$ Construction

A unique property of shape program inference is that the distribution  $p(x|z)$  is readily available in the form of the program executor  $E$ . We leverage this property to propose



LEST, a variant of self-training that does not create mismatch between pseudo-labels and their associated visual data. Similar to the self-training paradigm, LEST first constructs  $\mathbf{Z}$  as the current best program set,  $P^{\text{BEST}}$ . Then, differing from self-training, LEST constructs  $\mathbf{X}$  as the executed version of each program in  $\mathbf{Z}$ . By construction, the labels in  $\mathbf{Z}$  will now be correct for their paired shapes in  $\mathbf{X}$ . The downside is that, like wake-sleep, LEST may introduce a distributional mismatch between  $\mathbf{X}$  and  $S^*$ . But once again, as  $P^{\text{BEST}}$  better approximates  $S^*$ , the mismatch between the two distributions will decrease.

### 3.4. Inferring Programs with $p(z|x)$

During each round of fine-tuning, PLAD methods rely on  $p(z|x)$  to infer programs that approximate  $S^*$ . We propose to train PLAD methods such that the best matching inferred programs for  $S^*$  are maintained across rounds. Specifically, we construct a data structure  $P^{\text{BEST}}$  that maintains a program for each training shape in  $S^*$ . In this way, as more iterations are run,  $P^{\text{BEST}}$  always forms a closer approximation to  $S^*$ . There is an alternative framing, where  $P^{\text{BEST}}$  is reset each epoch, but we show experimental results in the supplemental material that this can lead to worse generalization.

To update  $P^{\text{BEST}}$  each round, we employ an inner-loop search procedure. For each shape in  $S^*$ ,  $p(z|x)$  suggests high-likelihood programs, and the  $P^{\text{BEST}}$  entry is updated to keep the program whose execution obtains the highest similarity to the input shape; the specific similarity metric varies by domain. While there are many ways to structure this inner-loop search, we choose beam-search, as we find it offers a good trade-off between speed and performance. Experimentally, we demonstrate that PLAD methods are capable of performing well even as the time spent on inner-loop search is varied (Section 4.4).

### 3.5. Training $p(z|x)$ with multiple PLAD methods

As detailed in the preceding sections, the main difference between PLAD approaches is in how they construct the  $(\mathbf{X}, \mathbf{Z})$  dataset used for fine-tuning  $p(z|x)$ . However, there is no strict requirement that these different  $(\mathbf{X}, \mathbf{Z})$  distributions be kept separate. We explore how  $p(z|x)$  behaves under fine-tuning from multiple PLAD methods, such as combining LEST and self-training. We implement these mixtures on a per-batch basis. Before  $p(z|x)$  training, we construct distinct  $(\mathbf{X}, \mathbf{Z})$  distributions for each method in the combination. Then, during training, each batch is randomly sampled from one of the  $(\mathbf{X}, \mathbf{Z})$  distributions. We experimentally validate the effectiveness of this approach in the next section.

## 4. Results

We evaluate a series of methods on their ability to fine-tune shape program inference models across multiple domains. We describe the different domains in Section 4.1 and

details of our experimental design in Section 4.2. In Section 4.3, we compare the reconstruction accuracy of each method, and study how they are affected by varying the time spent on inner-loop search (Section 4.4) and the size of the training set (Section 4.5). Finally, we explore the convergence speed of each method in Section 4.6.

### 4.1. Shape Program Domains

We run experiments across three shape program domains: 2D Constructive Solid Geometry (CSG), 3D CSG, and ShapeAssembly. Details can be found in the supplemental.

In CSG, shapes are created by declaring parametric primitives (e.g. circles, boxes) and combining them with Boolean operations (union, intersection, difference). CSG inference is non-trivial: as CSG uses non-additive operations (intersection, difference), inferring a CSG program does not simply reduce to primitive detection. For 2D CSG, we follow the grammar defined by CSGNet [25], using 400 shape tokens that correspond to randomly placed circles, triangles and rectangles on a 64 x 64 grid. For 3D CSG, we use a grammar that has individual tokens for defining primitives (ellipsoids and cuboids), setting primitive attributes (position and scales), and the three Boolean operators. Attributes are discretized into 32 bins.

ShapeAssembly is designed for specifying the part structure of manufactured 3D objects. It creates objects by declaring cuboid part geometries and assembling those parts together via attachment and symmetry operators. Our grammar contains tokens for each command type and parameter value; to handle continuous values, we discretize them into 32 bins.

### 4.2. Experimental Design

**Fine-Tuning Methods** We compare the ability of the following training schemes to fine-tune a model on a specific domain of interest:

- **SP:**  $p(z|x)$  with supervised pretraining.
- **RL:** Fine-tuning with REINFORCE.
- **WS:** Fine-tuning with wake-sleep.
- **ST:** Fine-tuning with self-training.
- **LEST:** Fine-tuning with latent execution self-training.
- **LEST+ST:** combining LEST and ST.
- **LEST+ST+WS:** combining LEST, ST and WS.

**Shape Datasets** Fine-tuning methods learn to specialize  $p(z|x)$  against a distribution of real shapes  $S^*$ . For each domain, we construct a dataset of shapes  $S^*$ , and split it into train, validation, and test sets. We perform early-stopping with respect to the validation set. For 2DCSG, we use the CAD dataset from CSGNet [25], which consists of front and side views of chairs, desks, and lamps from the Trimble 3D warehouse. We split the dataset into 10K shapes for training, 3K shapes for validation, and 3K shapes for testing. For 3D CSG and ShapeAssembly, we use CAD shapes from the

Method	2D CSG CD ↓	3D CSG IoU ↑	ShapeAssembly IoU ↑
Supervised Pretraining (SP)	1.580	41.0	37.6
REINFORCE (RL)	1.097	53.4	50.8
Wake-Sleep (WS)	1.118	67.4	57.2
Self-Training (ST)	0.841	67.3	61.3
LEST	0.976	69.8	56.5
LEST+ST	0.829	70.8	66.0
LEST+ST+WS	<b>0.811</b>	<b>74.3</b>	<b>66.4</b>

Table 2. Test-set reconstruction performance across multiple shape program inference domains. The top row contains results for the pretrained  $p(\mathbf{z}|\mathbf{x})$  model fine-tuned by the other methods. For 2D CSG the metric is Chamfer distance (CD, lower is better). For 3D CSG and ShapeAssembly the metric is intersection over union (IoU, higher is better). Individual PLAD methods outperform RL, and combining PLAD methods achieves the best performance across all domains (LEST+ST+WS).

chair, table, couches, and benches categories of ShapeNet; voxelizations are provided by [6]. We split the dataset into 10K shapes for training, 1K shapes for validation, and 1K shapes for testing.

**Model Architectures** For all experiments, we model  $p(\mathbf{z}|\mathbf{x})$  in an encoder-decoder framework, although the particular architectures vary by domain. In all cases, the encoder is a CNN that converts visual data into a latent variable, and the decoder is an auto-regressive model that decodes the latent variables into a sequence of tokens. For 2D CSG, we use the same  $p(\mathbf{z}|\mathbf{x})$  architecture as CSGNet. A CNN consumes  $64 \times 64$  binary mask shape images to produce a latent code that initializes a GRU-based recurrent decoder. For 3D CSG and ShapeAssembly, we use a 3D CNN that consumes  $32 \times 32 \times 32$  occupancy voxels. This CNN outputs a latent code that is attended to by a Transformer decoder network [32]; this network also attends over token sequences in a typical auto-regressive fashion.

**Supervised Pretraining** Before fine-tuning,  $p(\mathbf{z}|\mathbf{x})$  undergoes supervised pretraining on synthetically generated programs until it has converged on that set. For 2D CSG, we follow CSGNet’s approach. For 3D CSG, we construct valid programs by (i) sampling a set of primitives within the allotted grid (ii) identifying potential overlaps (iii) constructing a binary tree of boolean operations using these overlaps. For ShapeAssembly, we propose programs by sampling random grammar expansions according to the language’s typing system. We then employ a validation step where a program is rejected if any of its part are not the sole occupying part of at least 8 voxels (to discourage excessive part overlaps). For 3D CSG and ShapeAssembly we sample 2 million synthetic programs and train until convergence on a validation set of 1000 programs. Full details provided in the supplemental.

### 4.3. Reconstruction Accuracy

We evaluate the performance of each fine-tuning method according to reconstruction accuracy: how closely the output of a shape program matches the input shape from which it

was inferred, on a held out set of test shapes. The specific metric varies by domain. For 2D CSG, we follow CSGNet and use Chamfer Distance (CD), where lower distances indicate more similar shapes. For 3D CSG and ShapeAssembly, we use volumetric intersection over union (IoU).

For each domain, we run each fine-tuning method to convergence, starting with the same  $p(\mathbf{z}|\mathbf{x})$  model that has undergone supervised pretraining. The reward for RL models follows the similarity metric in each domain: CD for 2D CSG; IoU for 3D CSG and ShapeAssembly. For PLAD fine-tuning methods, the similarity metric in each domain determines which program is kept during updates to  $P^{\text{BEST}}$ . At inference time, when evaluating the reconstruction performance of each  $p(\mathbf{z}|\mathbf{x})$ , we employ a beam search procedure, decoding multiple programs in parallel, and choosing the program that achieves the highest similarity to the target shape. We use a beam size of 10, unless otherwise stated.

We present quantitative results in Table 2. Looking at the middle four rows, the two self-training variants (ST and LEST) outperform RL as a fine-tuning method in all the domains we studied. The wake-sleep variant (WS) also outperforms RL for both 3D CSG and ShapeAssembly. These are more challenging domains with larger token spaces, posing difficulties for policy gradient fine-tuning. As demonstrated by the last two rows, further improvement can be had by combining multiple methods: for each domain, the best performance is achieved by LEST+ST+WS. In fact, for 2DCSG, the test-set reconstruction accuracy achieved by LEST+ST+WS (0.811) outperforms previous state-of-the-art results, CSGNet (1.14) [25] and CSGNetStack (1.02) [26], for paradigms where the executor is treated as a black-box.

Mixing updates from multiple PLAD methods is beneficial because, in this joint paradigm, each method can cover the other’s weaknesses. For instance, employing ST ensures that some samples of  $\mathbf{X}$  are sourced from  $S^*$ , and employing LEST ensures that some samples of  $\mathbf{X}$  have paired  $\mathbf{Z}$  programs which are exact labels. We present qualitative results in Figure 3. The reconstructions from the PLAD combina-

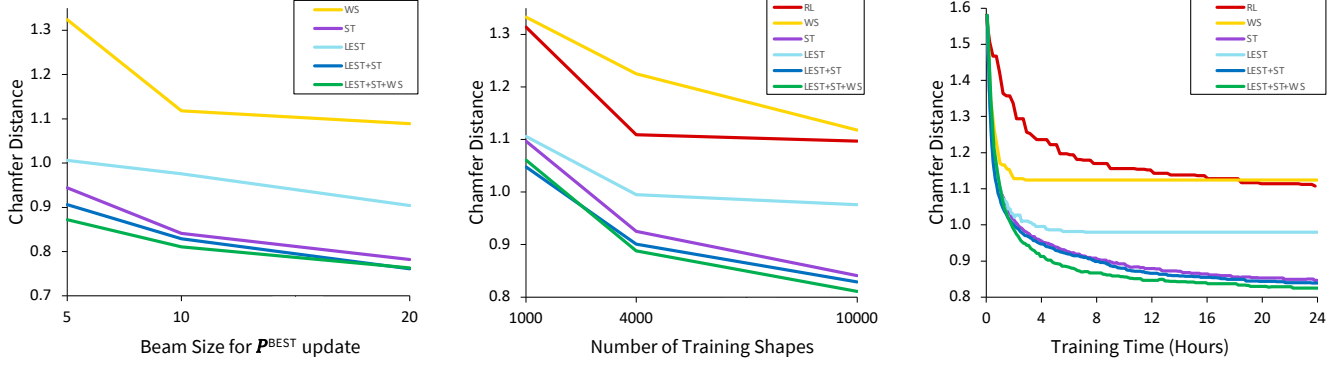


Figure 2. Experiments exploring properties of PLAD methods on 2D CSG. On the X-axis we plot the beam size used during the  $P^{\text{BEST}}$  update (*Left*), the number of training shapes (*Middle*), and the training time (*Right*). The Y-axis of each plot measures reconstruction accuracy on test-set shapes.

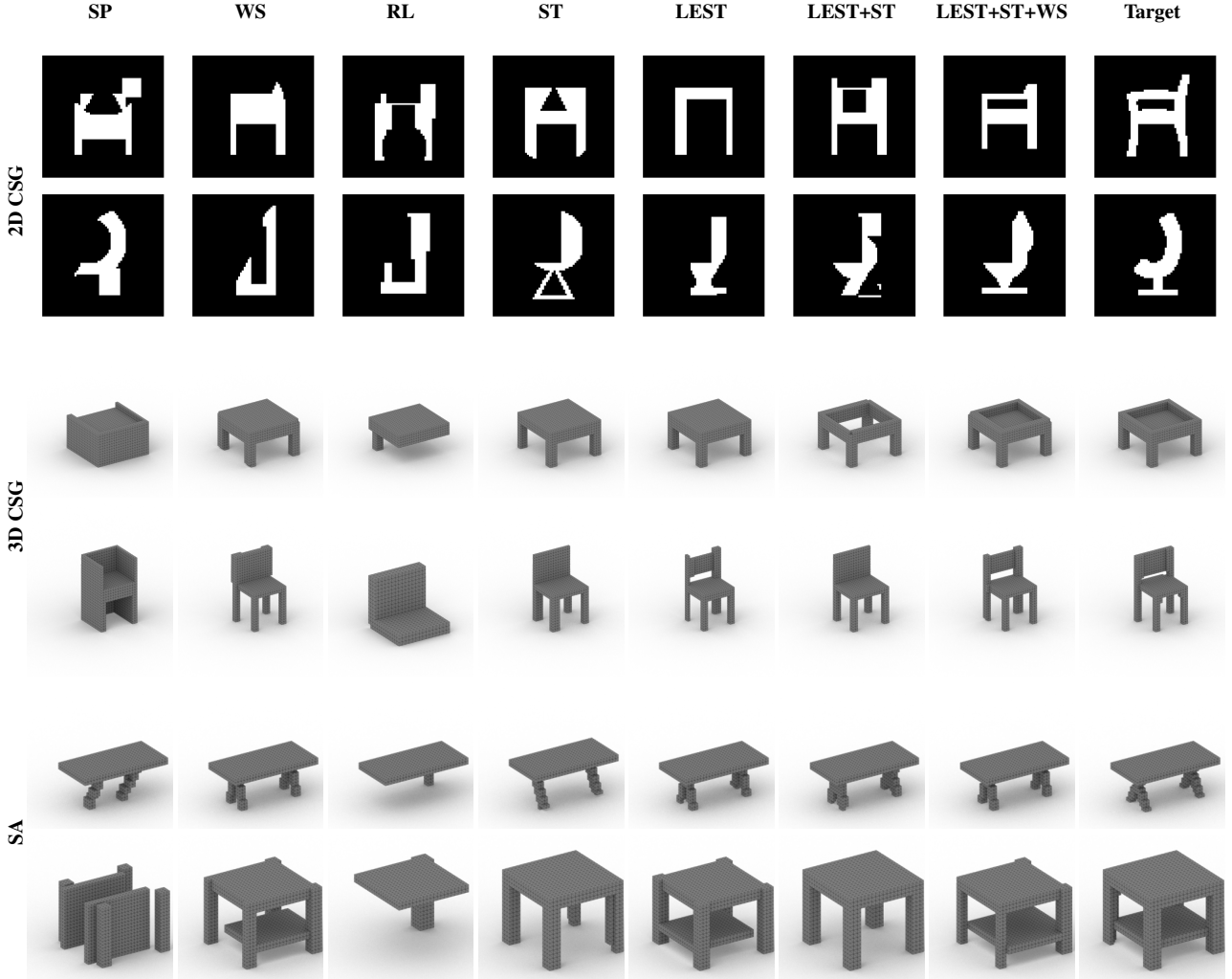


Figure 3. Qualitative comparisons of shape programs inferred for test-set shapes made by different fine-tuning methods for 2D CSG (*Top*), 3D CSG (*Middle*), and ShapeAssembly (*Bottom*). We provide additional qualitative results in the supplemental.

tion methods better reflect the input shapes, reinforcing the quantitative trends.

#### 4.4. Inner-loop Search Time

PLAD methods make use of  $P^{\text{BEST}}$  to generate  $(\mathbf{X}, \mathbf{Z})$  datasets that train  $p(\mathbf{z}|\mathbf{x})$ . To study how time spent on inner-loop search affects each technique, we ran an experiment using different beam sizes to update  $P^{\text{BEST}}$ . We present results in Figure 2, left. On the X-axis we plot beam size; on the Y-axis we plot test-set reconstruction Chamfer distance. Unsurprisingly, spending more time on the inner-loop search leads to better performance; finding better programs for training shapes improves test time generalization. That said, across all beam sizes, we find that it is always best to train under a combination of PLAD methods; the LEST+ST and the LEST+ST+WS variants always outperform any individual fine-tuning scheme. Note, RL is not included in this experiment because REINFORCE, as defined, has no inner-loop search mechanism. In this way, PLAD provides an additional control lever, where time spent on inner-loop search modulates a trade-off between convergence speed and test-set reconstruction performance.

#### 4.5. Number of Training Shapes from $S^*$

All the fine-tuning methods make use of a training distribution of shapes that are sampled from  $S^*$ . For some domains, the size of available samples from  $S^*$  may be limited. We run an experiment on 2D CSG to see how different fine-tuning methods are affected by training data size. We present the results of this experiment in Figure 2, middle. We plot the number of training shapes on the X-axis and test set reconstruction accuracy on the Y-axis. All fine-tuning methods improve as the training size of  $S^*$  increases, but once again, combining multiple PLAD methods leads to the best performance in all regimes. This study also demonstrates the sample efficiency of PLAD combinations: LEST+ST and LEST+ST+WS trained on 1,000 shapes achieve better test set generalization than RL trained on 10,000 shapes.

#### 4.6. Convergence Speed

Beyond reconstruction accuracy, we are also interested in the convergence properties of a fine-tuning method. Policy gradient RL is notoriously unstable and slow to converge, which is undesirable. For 2D CSG, we record the convergence speed of each method and present these results in Figure 2, right. We plot reconstruction accuracy (Y-axis) as a function of training wall-clock time (X-axis); all timing information was collected on a machine with a GeForce RTX 2080 Ti GPU and an Intel i9-9900K CPU. All PLAD techniques converge faster than policy gradient RL. For instance, RL took 36 hours to reach its converged test-set CD of 1.097, while LEST matched this performance at 1.1 hours

(32x faster) and LEST+ST matched this performance at 0.85 hours (42x faster).

## 5. Conclusion

We presented the PLAD framework to group a family of techniques for fine-tuning shape program inference models with Pseudo-Labels and Approximate Distributions. Within this framework, we proposed LEST: a self-training variant that creates a shape distribution  $\mathbf{X}$  approximating the real distribution  $S^*$  by executing inferred latent programs. Experiments on 2D CSG, 3D CSG, and ShapeAssembly demonstrate that PLAD methods achieve better reconstruction accuracy and converge faster than policy gradient RL, the current standard approach for black-box fine-tuning. Finally, we found that combining updates from multiple PLAD methods outperforms any individual technique.

While fine-tuning  $p(\mathbf{z}|\mathbf{x})$ , PLAD methods construct  $(\mathbf{X}, \mathbf{Z})$  sets approximating the statistics of  $S^*$ , specializing  $p(\mathbf{z}|\mathbf{x})$  towards  $S^*$ . As a consequence,  $p(\mathbf{z}|\mathbf{x})$  may not generalize as well to shapes outside of  $S^*$ ; we explore this phenomenon in the supplemental material. Training a general-purpose inference model for all shapes expressible under the grammar is an interesting line of future work.

While our work focuses on reconstruction quality, producing programs with ‘good’ structure matters just as much, if the program is to be used for editing tasks. Currently, the synthetic pretraining data is the only place where knowledge about what constitutes “good program structure” can be injected. Such knowledge must be expressed in procedural form, which may be harder to elicit from domain experts than declarative knowledge (i.e. “a good program has these properties” vs. “this is how you write a good program”). Finding efficient ways to elicit and inject such knowledge is an important future direction.

Finally, we believe that ideas from the PLAD framework are applicable to a broader class of program inference problems than those we evaluated in this paper. In principle, these approaches can be used to train an approximate inference model  $p(\mathbf{z}|\mathbf{x})$  for any domain in which (1) an executor  $p(\mathbf{x}|\mathbf{z})$  is available and (2) the executed output  $\mathbf{x}$  takes the form of some concrete artifact which can be encoded via neural network (image, audio, text, etc.). For example, one could imagine using PLAD techniques to infer graphics shader programs which produce certain textures or audio synthesizers which sound like certain real-world instruments.

## Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions. This work was funded in parts by NSF award #1941808 and a Brown University Presidential Fellowship. Daniel Ritchie is an advisor to Geopipe and owns equity in the company. Geopipe is a start-up that is



developing 3D technology to build immersive virtual copies of the real world with applications in various fields, including games and architecture

## References

- [1] Rudy Bunel, Matthew Hausknecht, Jacob Devlin, Rishabh Singh, and Pushmeet Kohli. Leveraging grammar and reinforcement learning for neural program synthesis. *ICLR*, 2018. [2](#)
- [2] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. *arXiv:1512.03012*, 2015. [1](#)
- [3] Eugene Charniak. Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI’97/IAAI’97, page 598–603. AAAI Press, 1997. [3](#)
- [4] Siddhartha Chaudhuri, Daniel Ritchie, Jiajun Wu, Kai Xu, and Hao Zhang. Learning Generative Models of 3D Structures. *Computer Graphics Forum*, 2020. [2](#)
- [5] Xinyun Chen, Chang Liu, and Dawn Song. Execution-guided neural program synthesis. In *International Conference on Learning Representations*, 2019. [2](#)
- [6] Zhiqin Chen, Kangxue Yin, Matthew Fisher, Siddhartha Chaudhuri, and Hao Zhang. Bae-net: Branched autoencoder for shape co-segmentation. *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. [6](#)
- [7] Jacob Devlin, Jonathan Uesato, Surya Bhupatiraju, Rishabh Singh, Abdel-rahman Mohamed, and Pushmeet Kohli. Robustfill: Neural program learning under noisy i/o. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, page 990–998. JMLR.org, 2017. [2](#)
- [8] Kevin Ellis, Maxwell Nye, Yewen Pu, Felix Sosa, Josh Tenenbaum, and Armando Solar-Lezama. Write, execute, assess: Program synthesis with a repl. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. [1](#), [2](#)
- [9] Kevin Ellis, Daniel Ritchie, Armando Solar-Lezama, and Josh Tenenbaum. Learning to Infer Graphics Programs from Hand-Drawn Images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. [2](#)
- [10] Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sable-Meyer, Luc Cary, Lucas Morales, Luke Hewitt, Armando Solar-Lezama, and Joshua B. Tenenbaum. Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning, 2020. [3](#)
- [11] Junxian He, Jiatao Gu, Jiajun Shen, and Marc’Aurelio Ran-zato. Revisiting self-training for neural sequence generation. In *International Conference on Learning Representations*, 2020. [3](#)
- [12] Luke B. Hewitt, Tuan Anh Le, and Joshua B. Tenenbaum. Learning to learn generative programs with memoised wake-sleep, 2020. [3](#)
- [13] Geoffrey E Hinton, Peter Dayan, Brendan J Frey, and Radford M Neal. The “wake-sleep” algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995. [3](#)
- [14] R. Kenny Jones, Theresa Barton, Xianghao Xu, Kai Wang, Ellen Jiang, Paul Guerrero, Niloy J. Mitra, and Daniel Ritchie. Shapeassembly: Learning to generate programs for 3d shape structure synthesis. *ACM Transactions on Graphics (TOG), Siggraph Asia 2020*, 39(6):Article 234, 2020. [2](#), [10](#)
- [15] Jacob Kahn, Ann Lee, and Awni Hannun. Self-training for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7084–7088. IEEE, 2020. [3](#)
- [16] Kacper Kania, Maciej Zieba, and Tomasz Kajdanowicz. Ucs-net – unsupervised discovering of constructive solid geometry tree. In *arXiv*, 2020. [1](#), [3](#)
- [17] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR 2015*, 2015. [12](#)
- [18] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014. [3](#), [4](#)
- [19] Chen Liang, Jonathan Berant, Quoc Le, Kenneth D Forbus, and Ni Lao. Neural symbolic machines: Learning semantic parsers on freebase with weak supervision. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 23–33, 2017. [3](#)
- [20] Yunchao Liu, Zheng Wu, Daniel Ritchie, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. Learning to Describe Scenes with Programs. In *International Conference on Learning Representations (ICLR)*, 2019. [2](#)
- [21] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA, June 2006. Association for Computational Linguistics. [3](#)
- [22] Radford M. Neal and Geoffrey E. Hinton. A new view of the em algorithm that justifies incremental and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1993. [3](#)
- [23] Emilio Parisotto, Abdel-rahman Mohamed, Rishabh Singh, Lihong Li, Dengyong Zhou, and Pushmeet Kohli. Neuro-symbolic program synthesis. *ICLR*, 2017. [2](#)
- [24] H. Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965. [3](#)
- [25] Gopal Sharma, Rishabh Goyal, Difan Liu, Evangelos Kalogerakis, and Subhransu Maji. CSGNet: Neural Shape Parser for Constructive Solid Geometry. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [1](#), [2](#), [5](#), [6](#), [10](#)
- [26] Gopal Sharma, Rishabh Goyal, Difan Liu, Evangelos Kalogerakis, and Subhransu Maji. Neural shape parsers for constructive solid geometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. [2](#), [6](#)
- [27] Ryan Spring and Anshumali Shrivastava. Scalable and sustainable deep learning via randomized hashing. In *Proceedings of the 23rd ACM SIGKDD International Conference*

on *Knowledge Discovery and Data Mining*, pages 445–454, 2017. [13](#)

- [28] Mark Steedman, Miles Osborne, Anoop Sarkar, Stephen Clark, Rebecca Hwa, Julia Hockenmaier, Paul Ruhlén, Steven Baker, and Jeremiah Crim. Bootstrapping statistical parsers from small datasets. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, Apr. 2003. Association for Computational Linguistics. [3](#)
- [29] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019. [13](#)
- [30] Shao-Hua Sun, Hyeonwoo Noh, Sriram Somasundaram, and Joseph Lim. Neural program synthesis from diverse demonstration videos. In *Proceedings of the 35th International Conference on Machine Learning*, 2018. [2](#)
- [31] Yonglong Tian, Andrew Luo, Xingyuan Sun, Kevin Ellis, William T. Freeman, Joshua B. Tenenbaum, and Jiajun Wu. Learning to Infer and Execute 3D Shape Programs. In *International Conference on Learning Representations (ICLR)*, 2019. [1](#), [2](#), [3](#)
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. [6](#), [12](#)
- [33] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8, 1992. [2](#)
- [34] Karl D. D. Willis, Yewen Pu, Jieliang Luo, Hang Chu, Tao Du, Joseph G. Lambourne, Armando Solar-Lezama, and Wojciech Matusik. Fusion 360 gallery: A dataset and environment for programmatic cad construction from human design sequences. *ACM Transactions on Graphics (TOG)*, 40(4), 2021. [1](#)
- [35] Jiajun Wu, Joshua B Tenenbaum, and Pushmeet Kohli. Neural scene de-rendering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [2](#)
- [36] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA, June 1995. Association for Computational Linguistics. [3](#)
- [37] Haoran You, Chaojian Li, Pengfei Xu, Yonggan Fu, Yue Wang, Xiaohan Chen, Richard G Baraniuk, Zhangyang Wang, and Yingyan Lin. Drawing early-bird tickets: Towards more efficient training of deep networks. *arXiv preprint arXiv:1909.11957*, 2019. [13](#)
- [38] Chenghui Zhou, Chun-Liang Li, and Barnabas Poczos. Unsupervised program synthesis for images using tree-structured lstm, 2020. [2](#)
- [39] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D. Cubuk, and Quoc V. Le. Rethinking pre-training and self-training, 2020. [3](#)

## A. Details of Domain Grammars

**2D CSG** We follow the grammar from CSGNet [25]. This grammar contains 3 Boolean operations (intersect, union, subtract), 3 primitive types (square, circle, triangle), and parameters to initialize each primitive (L and R tuples). Please refer to the CSGNet paper for details.

$$\begin{aligned}
 S &\rightarrow E; \\
 E &\rightarrow EET \mid P(L, R); \\
 T &\rightarrow \textit{intersect} \mid \textit{union} \mid \textit{subtract}; \\
 P &\rightarrow \textit{square} \mid \textit{circle} \mid \textit{triangle}; \\
 L &\rightarrow [8 : 8 : 56]^2; \quad R \rightarrow [8 : 4 : 32].
 \end{aligned}$$

**3D CSG** We design our own grammar for 3D CSG similar in spirit to the grammar of CSGNet. While CSGNet does contain a 3D CSG grammar, we find that it overly discretizes the possible spacing and positioning of primitives. Therefore in our grammar, we allow each primitive to be parameterized at the same granularity as the voxel grid (32 bins). In this way, each primitive takes in 6 parameters (instead of 2 parameter tuples), where the 6 parameters control the position and scaling of the primitive.

$$\begin{aligned}
 S &\rightarrow E; \\
 E &\rightarrow EET \mid P(F, F, F, F, F, F); \\
 T &\rightarrow \textit{intersect} \mid \textit{union} \mid \textit{subtract}; \\
 P &\rightarrow \textit{cuboid} \mid \textit{ellipsoid}; \\
 F &\rightarrow [1 : 32]
 \end{aligned}$$

**ShapeAssembly** ShapeAssembly is a domain-specific language for creating structures of 3D Shapes [14]. It creates structures by instantiating parts (*Cuboid* command), and then attaching parts to one another (*attach* command). It further includes macro operators that capture higher-order spatial patterns (*squeeze*, *reflect*, *translate* commands). To remain consistent with our CSG experiments, we further modify the grammar such that all continuous parameters are discretized.

$S \rightarrow BBoxBlock; ShapeBlock;$   
 $BBoxBlock \rightarrow \text{bbox} = \text{Cuboid}(1.0, x, 1.0)$   
 $ShapeBlock \rightarrow PBlock; ShapeBlock \mid None$   
 $PBlock \rightarrow c_n = \text{Cuboid}(x, x, x); ABlock; SBlock$   
 $ABlock \rightarrow Attach \mid Attach; Attach \mid Squeeze$   
 $SBlock \rightarrow Reflect \mid Translate \mid None$   
 $Attach \rightarrow \text{attach}(cube_n, f, uv, uv)$   
 $Squeeze \rightarrow \text{squeeze}(cube_n, cube_n, face, uv)$   
 $Reflect \rightarrow \text{reflect}(axis)$   
 $Translate \rightarrow \text{translate}(axis, m, x)$   
 $f \rightarrow \text{right} \mid \text{left} \mid \text{top} \mid \text{bot} \mid \text{front} \mid \text{back}$   
 $axis \rightarrow X \mid Y \mid Z$   
 $x \in [1, 32]/32.$   
 $uv \in [1, 10]^2/10.$   
 $n \in [0, 10]$   
 $m \in [1, 4]$

## B. Details of Synthetic Pretraining

**2DCSG** We follow the synthetic pretraining steps from CSGNet and directly use their released pretrained model weights. Please refer to their paper and code for further details.

**3DCSG** We generate synthetic programs for 3D CSG with the following procedure. First, we sample  $K$  primitives, where  $K$  is randomly chosen between 2 and 12. To sample a primitive, we sample a center position within the voxel space, and then we sample a scale, such that the scale is constrained so that the primitive will not extend past the borders of the voxel grid. We then find if the bounding boxes of any two primitives overlap in space (using the position and scale of each primitive). We then construct a binary tree of Boolean operations by randomly merging the  $K$  primitives together, until only one group remains. Each Boolean operation merges two primitive groups into a single primitive group. The type of semantically valid Boolean operation depends on the overlaps between primitives of the two groups. When a group of primitives  $A$  and a group of primitives  $B$  is merging: union is always a valid operation, difference is a valid operation if each primitive in group  $B$  shares an overlap with some primitive in group  $A$ , and intersection is a valid operation if each primitive in group  $A$  shares an overlap with some primitive in group  $B$  and each primitive in group  $B$  shares an overlap with some primitive in group  $A$ . We can then unroll this binary tree of boolean operations

into a sequence of tokens from the CSG grammar, forming a synthetic program. We sample 2,000,000 synthetic programs according to this procedure, that are used during supervised pretraining, and we sample another 1000 synthetic programs that we use as a validation set. We pretrain our model for 40 epochs, where each epoch takes around 1.5 hours to complete. At this check-point, the model had converged to a reconstruction IoU of .90 on both train and validation synthetic data.

**ShapeAssembly** We generate synthetic programs for ShapeAssembly with the following procedure. We first sample the number of primitive blocks  $K$  (PBlock), where  $K$  is randomly chosen between 2 and 8; note that the number of cuboids created can be greater than  $K$ , when symmetry operations are applied. Each PBlock is filled in with random samples according to the grammar syntax. First a cuboid is created, then an attach block is applied, then a symmetry block is applied. An attach block can contain either one attach operation, one squeeze operation, or two attach operations. A symmetry block can contain either a reflect operation, a translation operation, or no operation. Command parameters are randomly sampled according to simple heuristics (e.g. reflections are more common than translations) and in order to maintain language semantics (e.g. attaches can only be made to previously instantiated cuboid indices). A final validation step occurs after a complete set of program tokens has been synthetically generated; we execute the synthetic program, and check how many voxels are uniquely occupied by each cuboid in the executed output. If any cuboid uniquely occupies less than 8 voxels, the entire synthetic sample is rejected. We sample 2,000,000 synthetic programs according to this procedure, that are used during supervised pretraining, and we sample another 1000 synthetic programs that we use as a validation set. We pretrain our model for 26 epochs, where each epoch takes around 40 minutes to complete. At this check-point the model had converged to reconstruction IoU of .70 on both train and validation synthetic data.

## C. Experiment Hyperparameters

**3D Experiments** For 3D CSG and ShapeAssembly, we use the following model hyper-parameters.

The encoder for both cases is a 3D CNN that consumes a  $32 \times 32 \times 32$  voxel grid. It has four layers of convolution, ReLU, max-pooling, and dropout. Each convolution layer uses kernel size of 4, stride of 1, padding of 2, with channels (32, 64, 128, 256). The output of the CNN is a  $(2 \times 2 \times 2 \times 256)$  dimensional vector, which we transform into a  $(8 \times 256)$  vector. This vector is then sent through a 3-layer MLP with ReLU and dropout to produce a final  $(8 \times 256)$  vector that acts as an 8-token embedding of the voxel grid.

The decoder for both cases is a Transformer Decoder module [32]. It uses 8 layers and 16 heads, with a hidden dimension size of 256. It attends over the 8-token CNN voxel encoding and up to 100 additional sequence tokens, with an auto-regressive attention mask. We use a learned positional embedding for each sequence position. An embedding layer lifts each token into an embedding space, consumed by the transformer, and a 2-layer MLP converts Transformer outputs into a probability distribution over tokens.

In all cases we set dropout to 0.1 . We use a learning rate of 0.0005 with the Adam optimizer [17] for all training modes, except for RL, where following CSGNet we use SGD with a learning rate of 0.01 . During supervised pre-training we use a batch size of 400. During PLAD method fine-tuning we use batch size of 100. During RL fine-tuning we use a batch size of 4, due to memory limitations (a batch size of 4 takes up 10GB of GPU memory). Early stopping on the validation set is performed to determine when to end each round and when to stop introducing additional rounds. For deciding when to stop introducing additional rounds, we use a patience of 100 epochs. For deciding when to stop each round, we use a patience of 10 epochs. In both cases we employ a patience threshold of 0.001 IoU improvement (e.g. we must see at least this much improvement to reset the patience). Within each round of PLAD training, we check validation set reconstruction performance with a beam size of 3; between rounds of PLAD training we check validation set reconstruction performance with a beam size of 5; final reconstruction performance of converged models is computed with a beam size of 10.

For RL runs, we make a gradient update after every 10 batches, following CSGNet. For runs that involve VAE training (all Wake-Sleep runs), we add an additional module in-between the encoder and the decoder. This module uses an MLP to convert the output of the encoder into a  $128 \times 2$  latent vector (representing 128 means and standard deviations). This module then samples an 128 dimensional vector from a normal distribution described by the means and standard deviations, and further lifts this encoding into the dimension that the decoder expects with a sequence of linear layers. For each round of VAE training, we allow the VAE to update for no more than 100 epochs. We perform early-stopping for VAE training with respect to its loss, where the loss is a combination of reconstruction (cross-entropy on token predictions) and KL divergence, both weighed equally.

**2D Experiments** For 2DCSG, we follow the network architecture and hyper-parameters of CSGNet. All training regimes use a dropout of 0.2 and a batch size of 100. PLAD methods use the Adam optimizer with a learning rate of 0.001. For deciding when to stop introducing additional rounds, we use a patience of 1000 epochs. For deciding when to stop each round, we use a patience of 10 epochs.

$P^{BEST}$ mode	ST	LEST	LEST+ST	LEST+ST+WS
Per round	0.881	1.011	0.853	0.845
All-time	<b>0.841</b>	<b>0.976</b>	<b>0.829</b>	<b>0.811</b>

Table 3. Different ways to update  $P^{BEST}$  data structure. In the "Per round" row, the data structure is cleared in between rounds. In the "All-time" row, the data structure maintains the best program for each input shape across multiple rounds.

In both cases we employ a patience threshold of 0.005 CD improvement. The parameters for the RL runs and VAE training are the same as in the 3D Experiments.

## D. P Best Update mode

During updates to  $P^{BEST}$ , we choose to update each entry in  $P^{BEST}$  according to which inferred program has achieved the best reconstruction similarity with respect to the input shape. The entries of this data structure are maintained across rounds. There is another framing where the entries of this data structure are reset each round, so that the best program for each shape is reset each epoch. This is similar to traditional self-training framing.

We run experiments on 2D CSG with this variant of  $P^{BEST}$  update and present results in Table 3. When the best program is maintained across rounds (All-time, bottom row) each fine-tuning strategy reaches a better converged reconstruction accuracy compared with when the best program is reset after each round (Per round, top row).

## E. Failure to generalize beyond $S^*$

As demonstrated by our experiments, PLAD fine-tuning methods are able to successfully specialize  $p(\mathbf{z}|\mathbf{x})$  towards a distribution of interest  $S^*$ . Unfortunately, this specialization comes at a cost; the fine-tuned  $p(\mathbf{z}|\mathbf{x})$  may actually generalize worse to out of distribution samples. To demonstrate this, we collected a small dataset of 2D icons from the The Noun Project<sup>1</sup>. We tested the shape program inference abilities of the initial  $p(\mathbf{z}|\mathbf{x})$  trained under supervised pretraining (SP) and of the fine-tuned  $p(\mathbf{z}|\mathbf{x})$  trained under PLAD regimes (LEST+ST+WS) and specialized to CAD shapes. We show qualitative examples of this experiment in Figure 4. While both methods fail to accurately represent the 2D icons, fine-tuning  $p(\mathbf{z}|\mathbf{x})$  on CAD shapes lowers the reconstruction accuracy significantly; the SP variant achieves an average CD of 1.9 while the LEST+ST+WS variant achieves a CD of 4.1 Developing  $p(\mathbf{z}|\mathbf{x})$  models capable of out-of-domain generalization is an important area of future research.

<sup>1</sup><https://thenounproject.com>



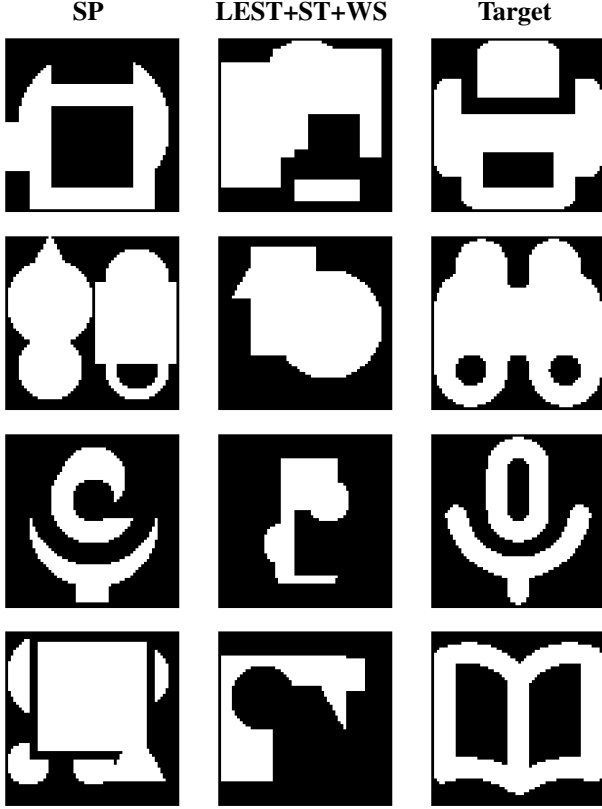


Figure 4. Qualitative examples of inferring 2D CSG programs for 2D icons. Both SP and LEST+ST+WS fail to infer representative programs, but the reconstructions from LEST+ST+WS are even less accurate than those from SP.

## F. Potential Societal Impacts

Fine-tuning our deep neural networks  $p(\mathbf{z}|\mathbf{x})$  requires a relatively large amount of electricity, which can have a significant environmental impact [29]. Reducing the energy consumption of deep learning is an active research area [27, 37]. Notably, PLAD techniques place no restrictions on the inference model, making it easy to adopt more efficient deep learning techniques. Moreover, shape program inference procedures may also allow the reverse engineering of protected intellectual property. Thus, improvements in shape program inference may impact the content and enforcement of copyright law.

## G. Additional Qualitative Results

We present additional qualitative results comparing various fine-tuning methods in Figure 5 (2D CSG), Figure 6 (3D CSG) and Figure 7 (ShapeAssembly).

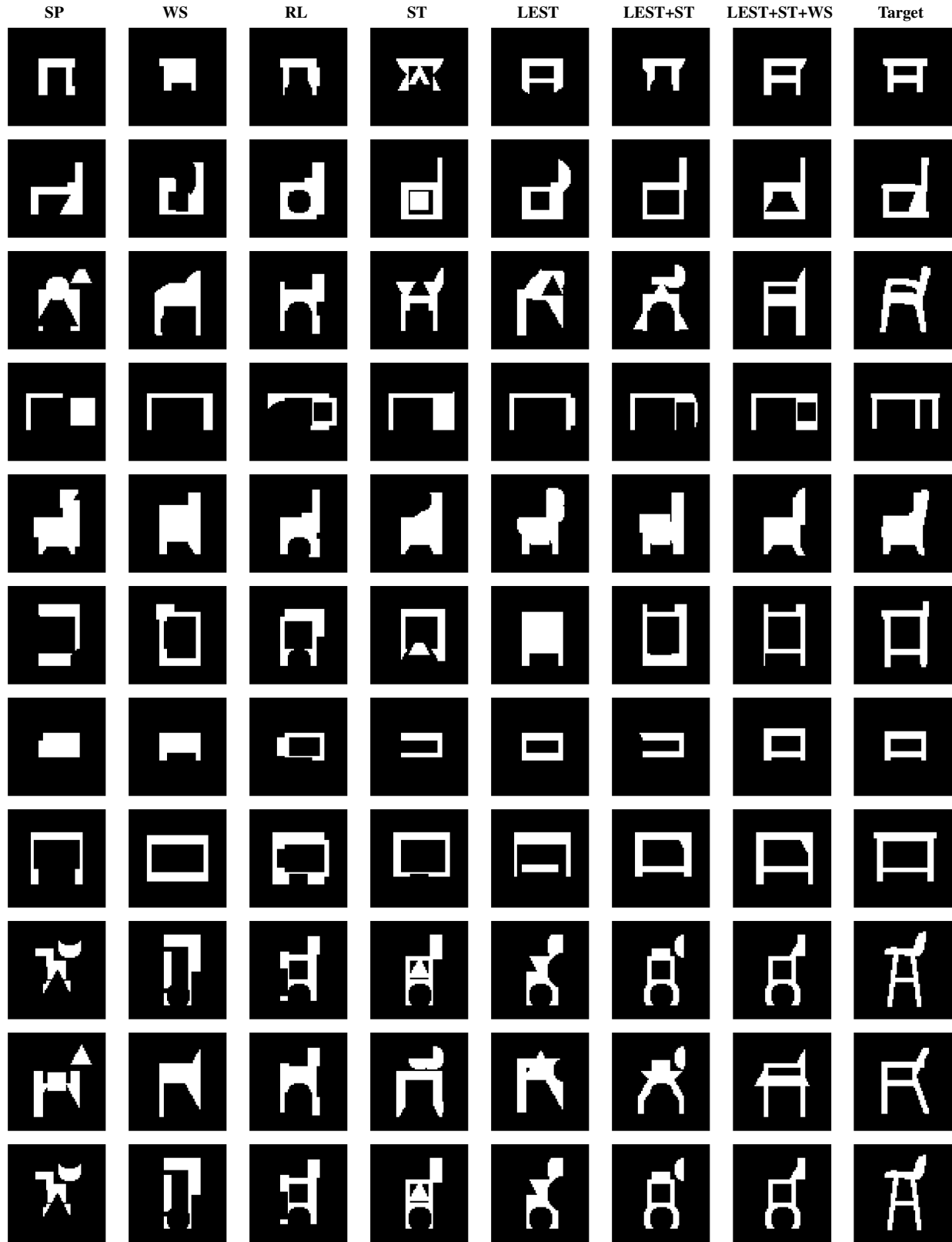


Figure 5. 2DCSG qualitative examples.

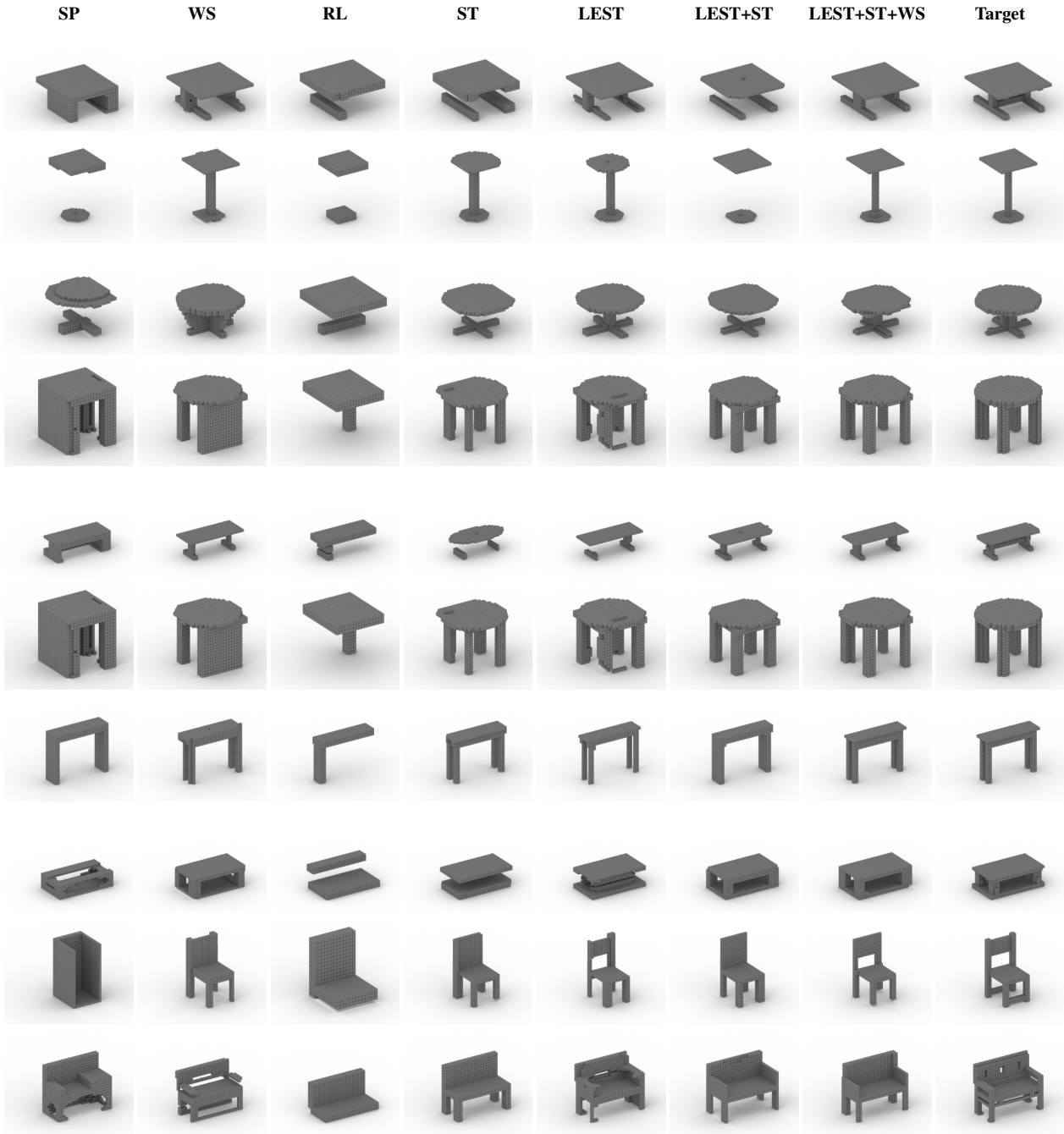




Figure 7. ShapeAssembly qualitative examples.