Bridging Neural and Symbolic Representations with Transitional Dictionary Learning

Junyan Cheng, Peter Chin

Thayer School of Engineering, Dartmouth College

ICLR 2024

Motivation

- **Objective:** Learn a representation that embeds both the compressive power of neural embeddings and the structural information in symbols.
- **Key Insights:**
  - *“Primitive symbols”* emerged in the human brain during the evolution from low-level neural perception to high-level symbols.
  - Symbols represent the entities or concepts that are most frequently reused and composed with each other.
- **Central Question:** Can we learn such a transitional representation that introduces structural information to neural embeddings?

Transitional Representation

- Neural predicate logical representation of an image x: $\Omega_x = \rho_1^1(\cdot) \wedge \rho_2^1(\cdot) \wedge \cdots \wedge \rho_1^2(\cdot,\cdot) \wedge \rho_2^2(\cdot,\cdot) \wedge \cdots$.
  - $\rho_j^i \in D^i$ is an i-ary predicate from i-ary dictionary D^i storing structural information.
  - Each $\cdot$ is the embedding of a visual part r_i storing high-dimensional information.
- **Transitional representation** $R = \{r_i\}_{i=1}^{N_P} = f(x;\theta)$, where each $r_i \in \mathbb{R}^d$ are grounded by dictionaries D parameterized by θ.
- $\boldsymbol{\theta = \underset{\theta}{\mathbf{argmin}} \sum_i^N \epsilon\left(g\left(R^i;\theta\right), x^i\right) + \alpha d_S\left(g_\theta\left(R^i\right), x^i\right)}$, g reconstruct the input with R, $\Omega_\mathrm{R} = g_\theta(R)$, ϵ is the reconstruction error, d_S is the "semantic distance".

Compositions as Subwords

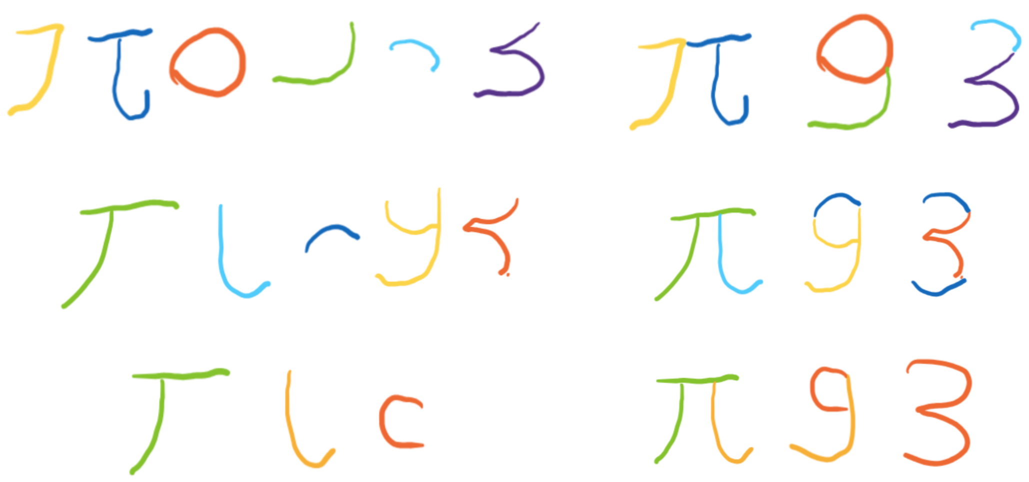
- From our definition, symbols are the compositions that can be **frequently reused and composed**, which means $d_S\left(g_\theta\left(R^i\right), x^i\right) \propto -P(\Omega_{R^i}|x^i, \theta)$.
- Thus, we need to solve $\boldsymbol{\mathrm{argmax}_\theta L = \sum_{i=1}^{N} logP(\Omega_{R^i}|x^i, \theta)}$, we reduce this target to a similar problem, **subword tokenization**:

Corpus hug, pug, pun

Dataset 

	Dictionary (subwords)	Parse (by Viterbi)
Iteration 1	h, u, p, n, g	hug, pug, pun
Iteration 2	h, u, pu, n, g	hug, pug, pun
Iteration 3	hu, pu, n, g	hug, pug, pun

Dictionary (prototypes) Parse (by NN)

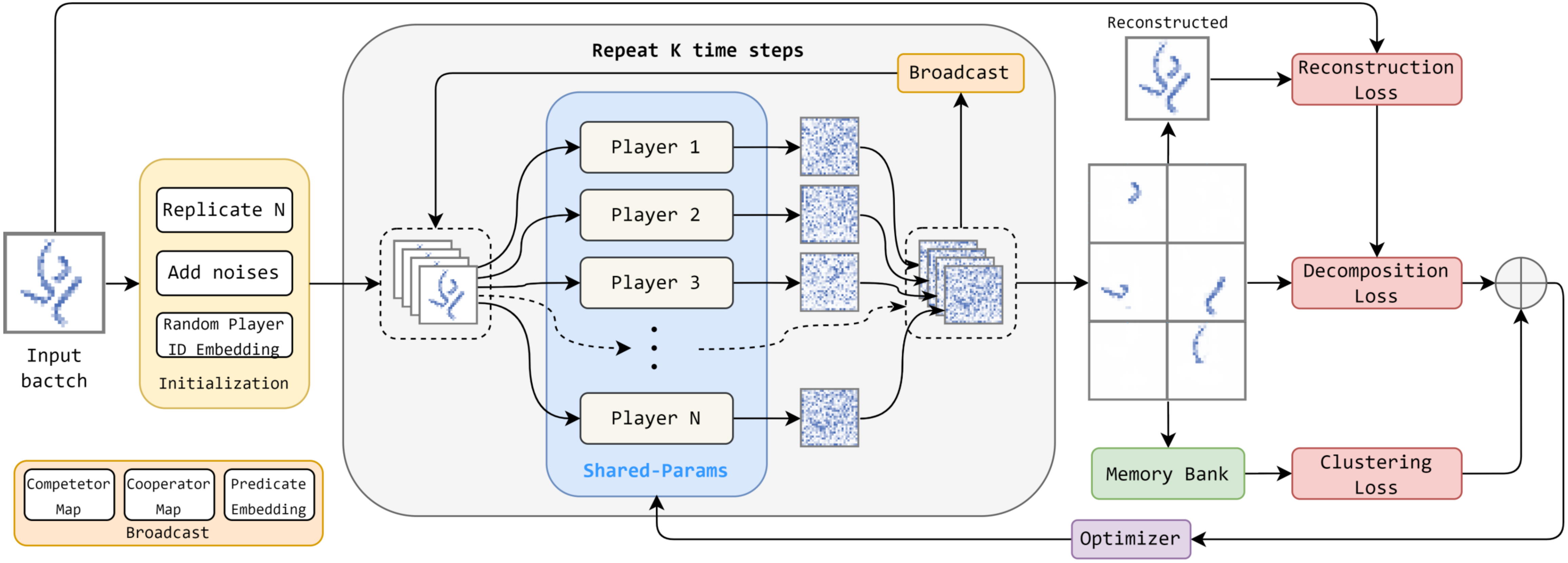


Transitional Dictionary Learning

- Kudo (ACL, 2018) proposed to use an **EM algorithm** that maximizes the likelihood of the corpus with a Unigram Language Model (ULM).
- **In our method:** Likelihood $L = \sum_{i=1}^{N} \sum_{j=1}^{N_A} \log P(\Omega_{R^i} | x^i, \theta)$, where N_A is the number of arities considered, calculated for different arities:
  - **1-ary:** Uses the 1-gram model same as ULM $\log P\left(\Omega_{R^i} \middle| x^i, \theta\right) = \sum_{k=1}^{N_P} \log P(r_k^i)$.
  - **N-ary:** Not considered in ULM. Uses joint probability, not N-gram models (which assume sequences). For 2-ary: $\log P\left(\Omega_{R^i} \middle| x^i, \theta\right) = \sum_{p=1}^{N_P} \sum_{q=1}^{N_P} \log P(r_p^i, r_q^j)$.
- **Transitional Dictionary Learning (TDL)**: Optimize the multi-ary EM algorithm above while minimizing reconstruction error as the constraint.

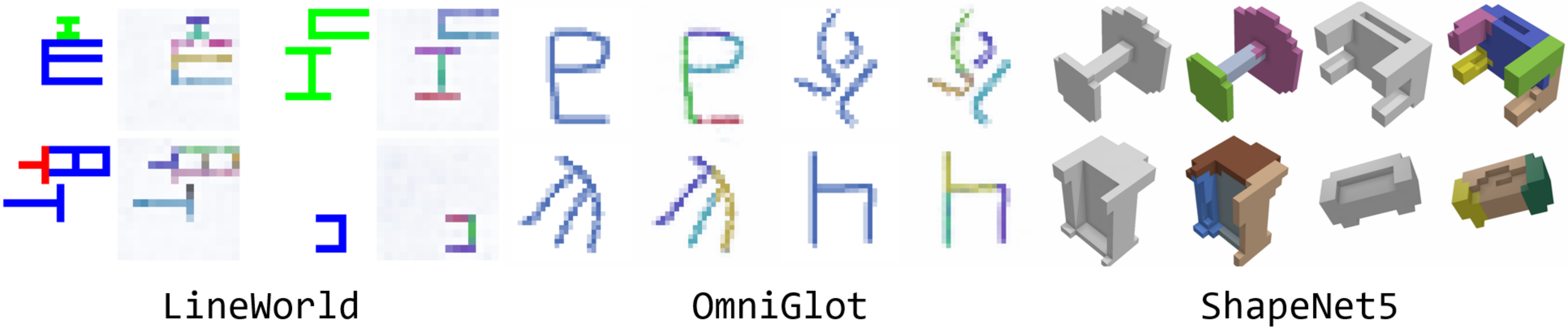
Implementing TDL for Vision Data

- We assess our TDL framework within the *abstract visual object* settings:
  - Train a model to **generate the visual parts** to reconstruct the input.
  - Meanwhile, **clustering** the generated parts to learn prototype dictionaries.



Unsupervised Learning Experiment

- We use three abstract compositional visual objects datasets:



- **LineWorld**: 50K images of 1~3 non-overlapping shapes made up of parallel or perpendicular lines generated by babyARC engine (Wu et al., NeurIPS, 2022).
- **OmniGlot** (Lake et al., Science, 2015): 27K handwritten characters.
- **ShapeNet5**: 27K voxelized 3D shapes from ShapeNet (Chang et al., arXiv 1512.03).

Evaluation metrics: Clustering Information Gain

- **Clustering Information Gain (CIG):** assess the learned dictionaries.
  - Mean Clustering Error (or Energy) $MCE = [\sum_{i=1}^{N}\sum_{j=1}^{N_P}(\min_{c \in C}\left\|r_j^i - c\right\|_2)/N_P]/N$
  - $CIG = 1 - MCE_{model}/MCE_{random}$, compare to a random dictionary with no information, $CIG = 1$ means all clusters are concentrated in their centroids, $CIG = 0$ means data points are evenly scattered, higher CIG means higher cohesive.

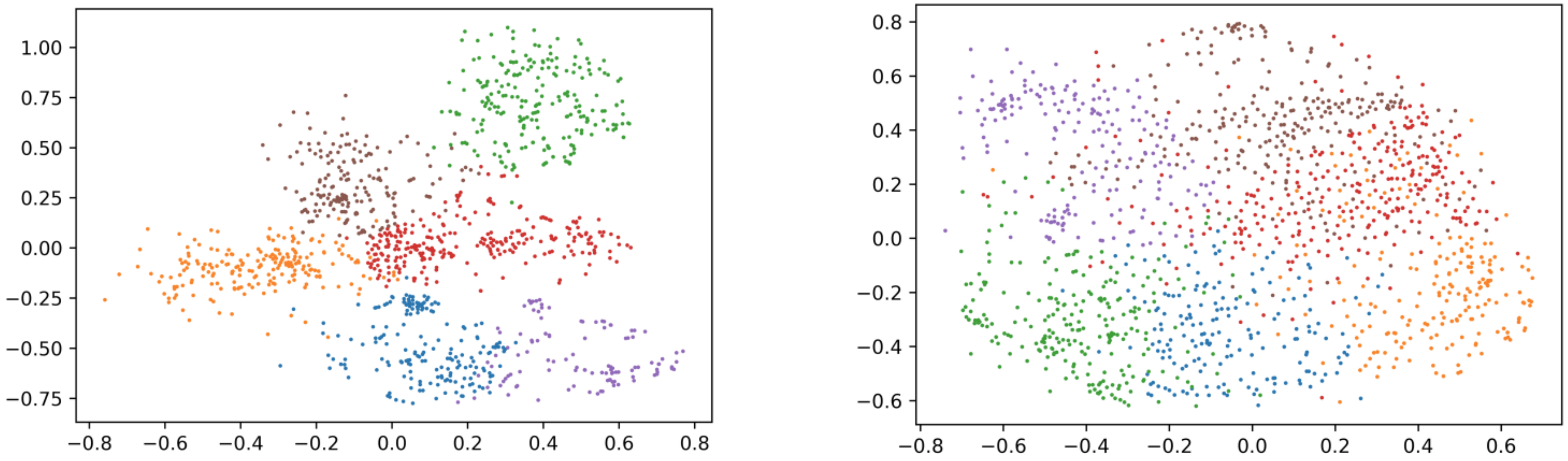
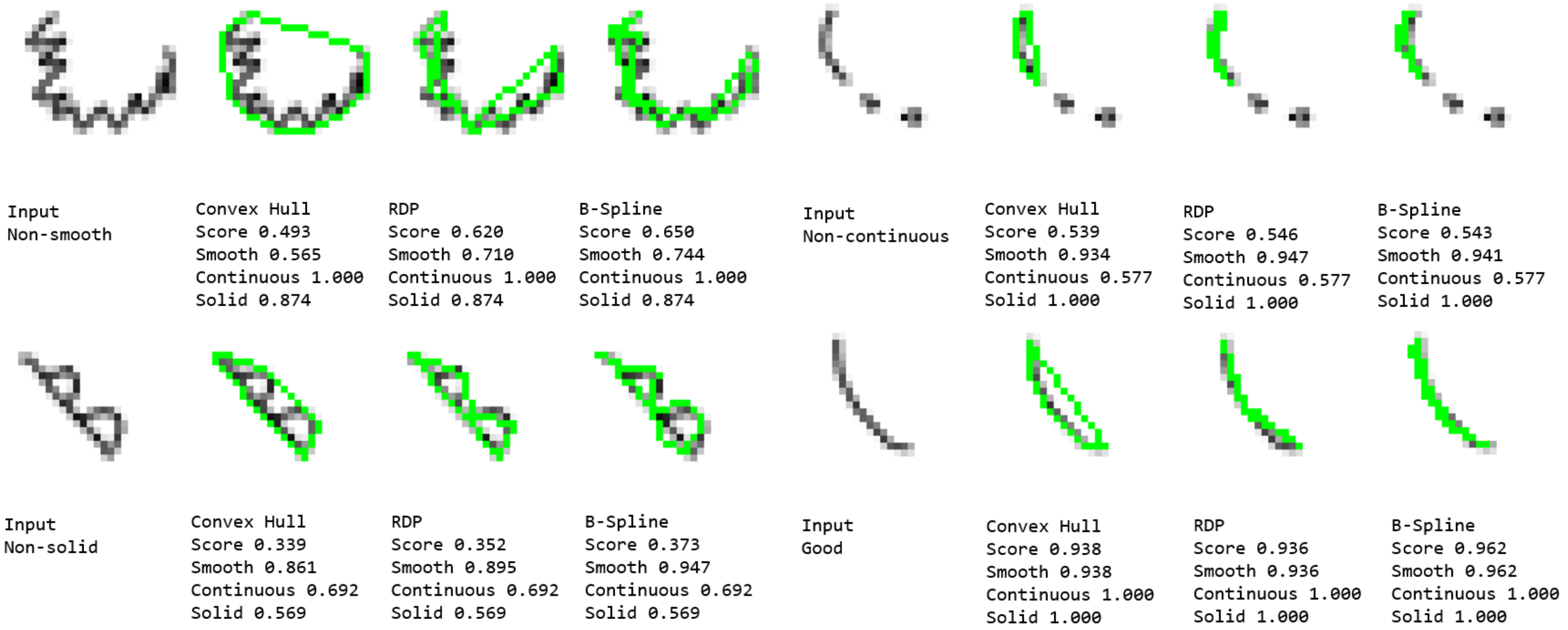


Figure 3: t-SNE for the latent space of 1 (left) and 2-ary (right) representations in LineWorld test set.

Evaluation metrics: Heuristic Shape Score

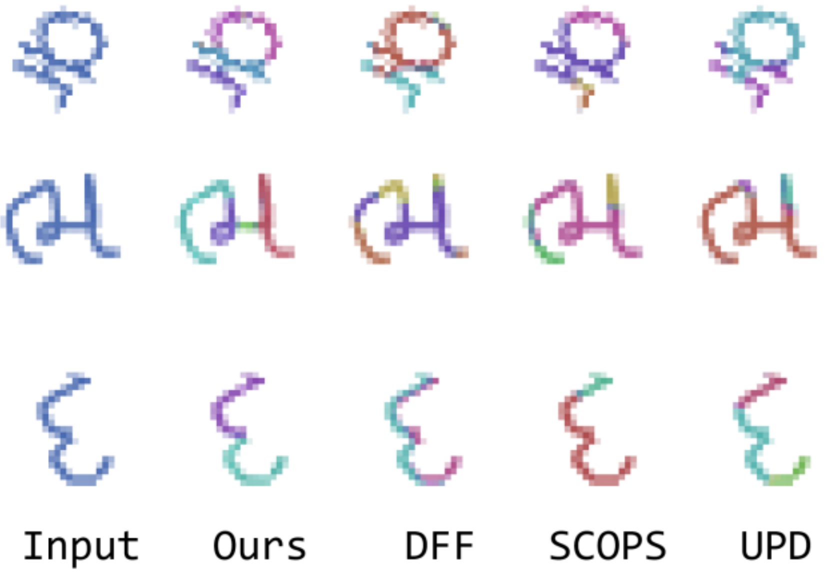
- **Heuristic Shape Score (SP):** evaluates the generated visual parts in three dimensions based on whether the shapes are natural and meet human intuition:
  - **Solidity**: there are no holes inside a part.
  - **Smoothness** the surfaces or contours of the part are smooth.
  - **Continuity**: the shape is not segmented and is an integral whole.



Results for Unsupervised Learning

- We compare to 3 unsupervised part segmentation baselines:
  - **DFF** (Collins et al, ECCV, 2018) use Non-negative Matrix Factorization (NMF) on activation map of last convolution layer of a pretrained backbone (e.g., VGG-19).
  - **SCOPS** (Hung et al, CVPR, 2019) and **UPD** (Choudhury, NeurIPS, 2021) learn to produce k-channels heatmap of parts with self-supervised learning.
- "RL" tune an unsupervised learning model with Shape Score as reward.
- "AE" is a reference auto-encoder as a baseline for the reconstruction error.

	LineWorld			**OmniGlot**			**ShapeNet5**		
	IoU	CIG	SP	MAE	CIG	SP	IoU	CIG	SP
AE	97.7	-		0.9	-		85.1	-	
DFF	-	33.1	38.3	-	36.9	33.3	-	20.1	19.2
SCO.	-	35.7	42.4	-	38.6	38.9	-	23.1	24.3
UPD	-	36.3	42.8	-	42.8	37.4	-	25.4	22.6
Ours	94.3	**58.0**	**82.6**	1.8	**68.5**	**77.6**	79.8	**54.6**	**60.1**
w/o RL	93.7	57.0	71.9	2.0	65.1	68.0	78.8	52.9	54.4



Transfer Learning Experiments

- Finetune unsupervised learning pre-trained models on two downstream tasks:
  - **LW-G**: predict the part mask (e.g., lines), and pair-wise relation annotations (e.g., perpendicular and parallel) from the babyARC engine, contains 7K samples.
  - **OG-G**: predict ground-truth strokes from OmniGlot, contains 5.8K samples that are not used in unsupervised learning.

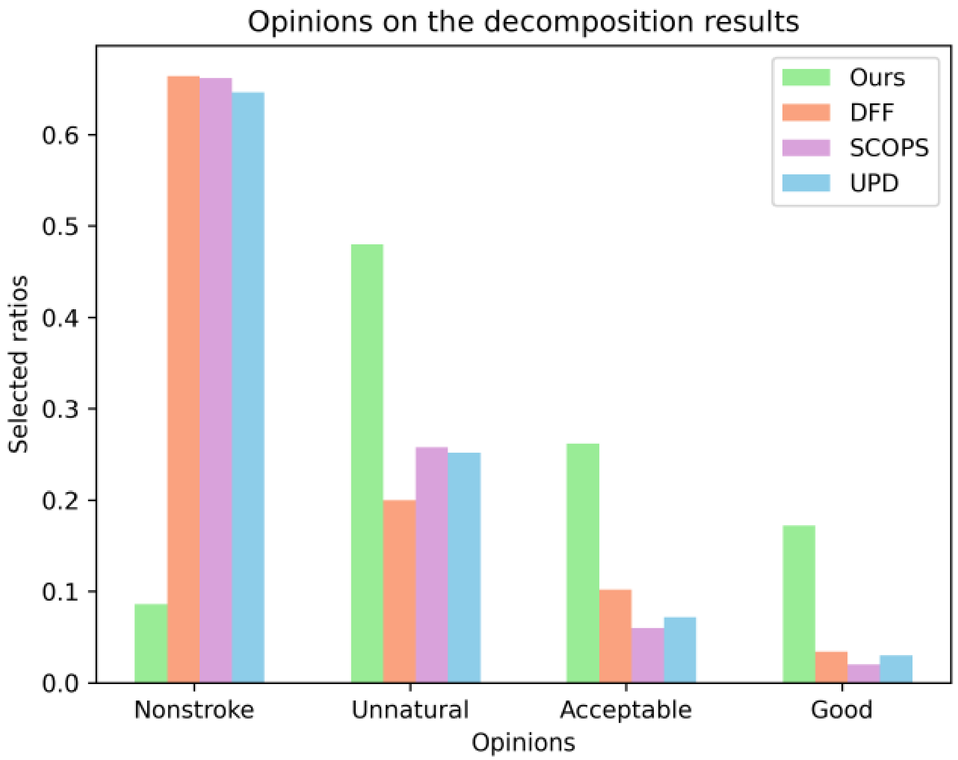
	LW-G		**OG-G**
	IoU	Acc.	IoU
AE	-		-
DFF	43.1	28.8	42.8
SCO.	46.8	26.4	46.9
UPD	46.2	28.7	48.9
Ours	**78.4**	**74.8**	**75.9**
w/o RL	78.2	74.3	75.1

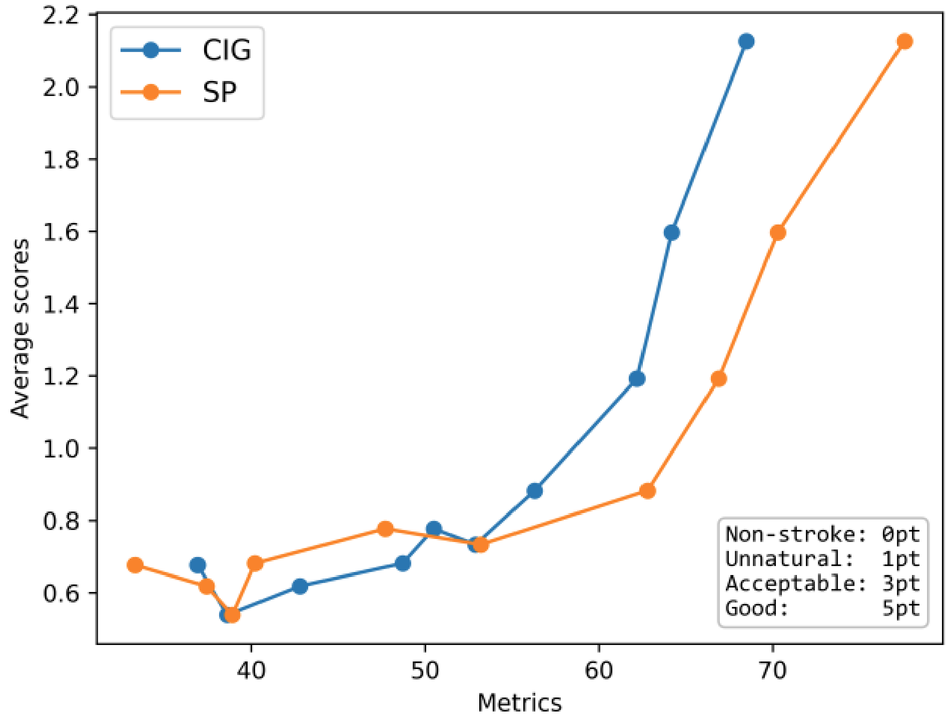
- Few-shot learning on unseen classes from ShapeGlot (Achlioptas et al., ICCV, 2019) with our model, each class has 230~550 samples, "PT" is pre-training.

	Bed			**Lamp**			**Sofa**			**Table**		
	IoU	CIG	SP	IoU	CIG	SP	IoU	CIG	SP	IoU	CIG	SP
w/ PT	67.3	48.1	52.9	61.1	42.1	49.1	62.2	46.8	45.2	68.3	50.1	54.6
w/o PT	18.1	19.0	13.2	18.3	19.9	14.6	21.5	18.9	19.8	19.9	22.1	17.9

Human Evaluation

- We further hire human annotators to rate the decomposition results in the OmniGlot test set from ours and baselines:
  - Our methods provides much more valid strokes.
  - The proposed methods show a consensus to the human evaluation.





Thank you!