# ZeroC: A Neuro-Symbolic Model for Zero-shot Concept Recognition and Acquisition at Inference Time

NeurIPS 2022

Tailin Wu[1], Megan Tjandrasuwita[2], Zhengxuan Wu[1], Xuelin Yang[1], Kevin Liu[1], Rok Sosic[1], Jure Leskovec[1]

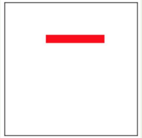
[1] Stanford University
[2] MIT

# Motivation

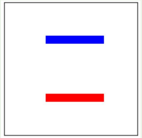
Humans have the remarkable ability to **recognize** and **acquire** novel visual concepts in a zero-shot manner

Suppose we humans have only learned the concept of “line” and relation of “parallel” and “perpendicular”:

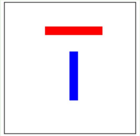
**Prior knowledge:**



“Line” (concept)



“Parallel” (relation)

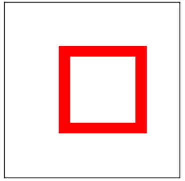


“Perpendicular” (relation)

Zero-shot **recognize** novel (hierarchical) concepts:

**Given:** Symbolic structure of a new concept
E.g.. when told a “rectangle” consists of two pairs of “lines”, the lines within the pairs are “parallel,” and the lines between the pairs are “perpendicular”

**Zero-shot recognition:**



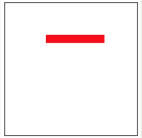
“rectangle”

# Motivation

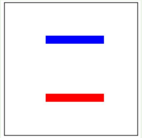
Humans have the remarkable ability to **recognize** and **acquire** novel visual concepts in a zero-shot manner

Suppose we humans have only learned the concept of “line” and relation of “parallel” and “perpendicular”:

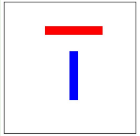
**Prior knowledge:**



“Line” (concept)



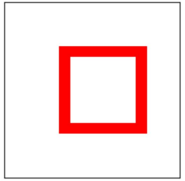
“Parallel” (relation)



“Perpendicular” (relation)

Zero-shot **acquire** novel (hierarchical) concepts:

**Given a single demonstration:**



“rectangle”

**Zero-shot acquire:** Symbolic structure of a new concept

A “rectangle” consists of two pairs of “lines”, the lines within the pairs are “parallel,” and the lines between the pairs are “perpendicular”

## Problem definition and significance:

How can we endow machine learning (ML) models with the capability of zero-shot recognition and acquisition of hierarchical visual concepts?

Having such capability will allow ML models to tackle more complex tasks at inference time, without further training on those specific tasks.

## Why is it hard:

Because machine learning models typically generalize to examples drawn from same/similar distribution as in training. Here we would like the model to generalize to more complex, hierarchical concepts, not seen previously.

## Prior methods:

Only address part of the problem:

- **Visual compositionality:** [1-2] address factors of variation (e.g. color, position, smiling) without hierarchical structures; [3] addresses composition of transformation.
- **Concept** or **relation learning** [4-7]: do not generalize to hierarchical concepts.
- **Zero-shot learning** [8-10]**:** only generalize to new combinations of features (constituent concepts) while neglecting relation structures.

[1] Du et al. NeurIPS 2020
[2] Higgins et al. ICLR 2018
[3] Andreas et al. CVPR 2016
[4] Snell, NeurIPS 2017
[5] Mao et al. ICLR 2019
[6] Kipf et al. ICLR 2018
[7] Shanahan et al. ICML 2020
[8] Romera et al. ICML 2015
[9] Bucher et al. ICCV 2017
[10] Schonfeld et al. CVPR 2019

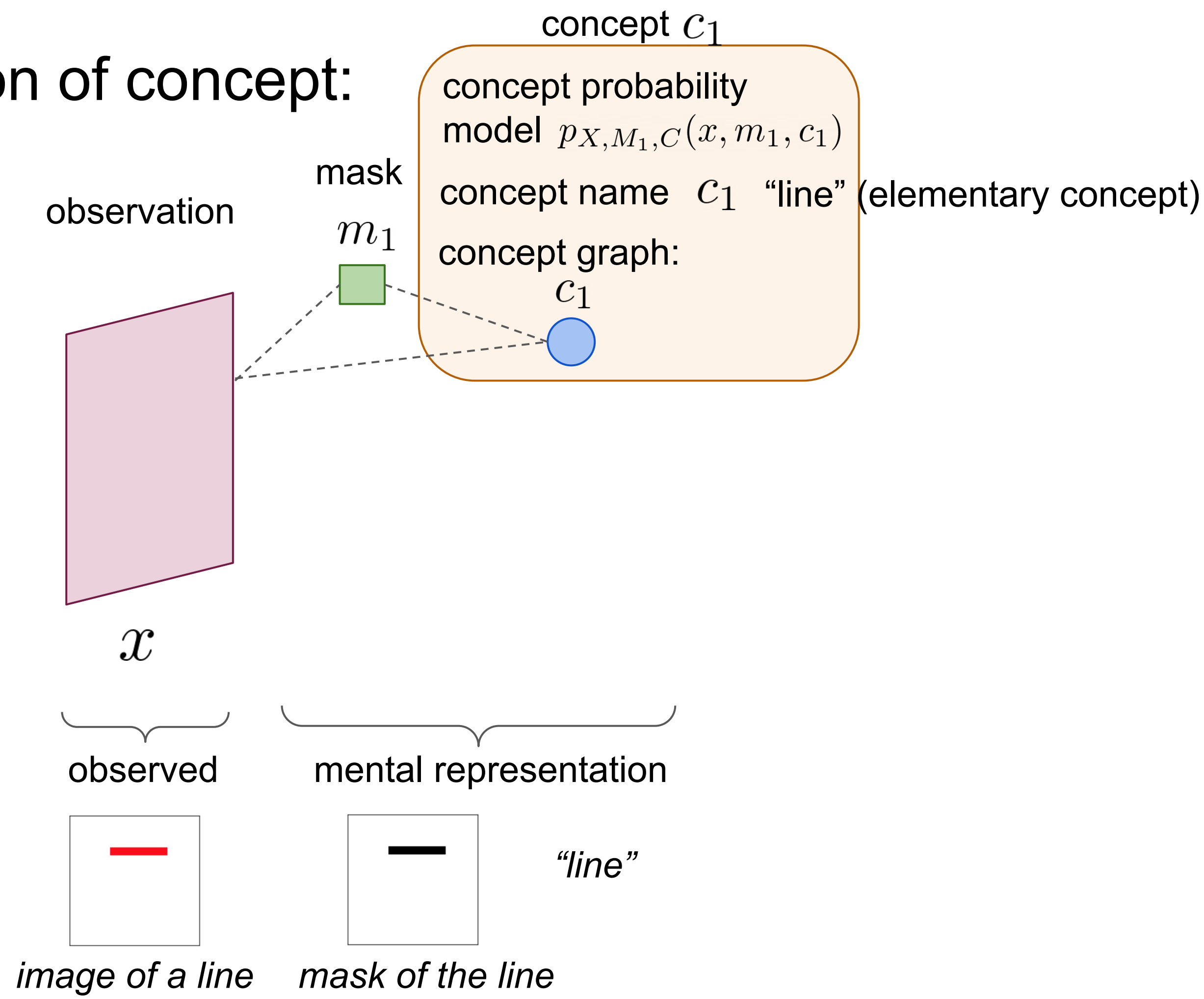
# Our contribution:

In this work, we introduce Zero-shot Concept Recognition and Acquisition (ZeroC) to address this problem.

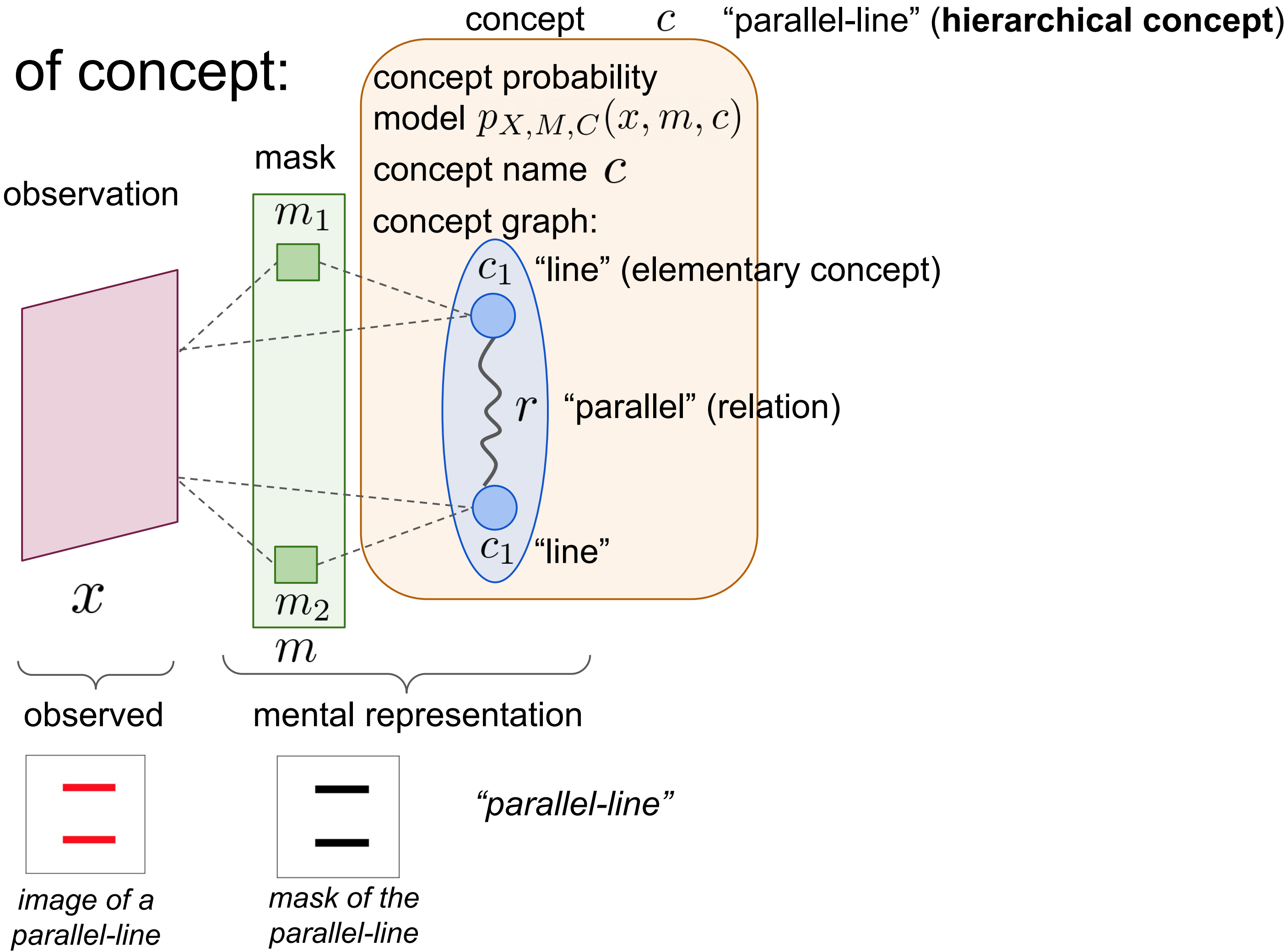
ZeroC represents concepts as **graphs** of constituent concept models (as nodes) and their relations (as edges). It allows a **one-to-one** mapping between a *symbolic graph structure* of a concept and its corresponding *recognition model.*

It (for the first time) allows acquiring new concepts, communicating its graph structure, and applying it to classification and detection tasks (even across domains) at inference time.

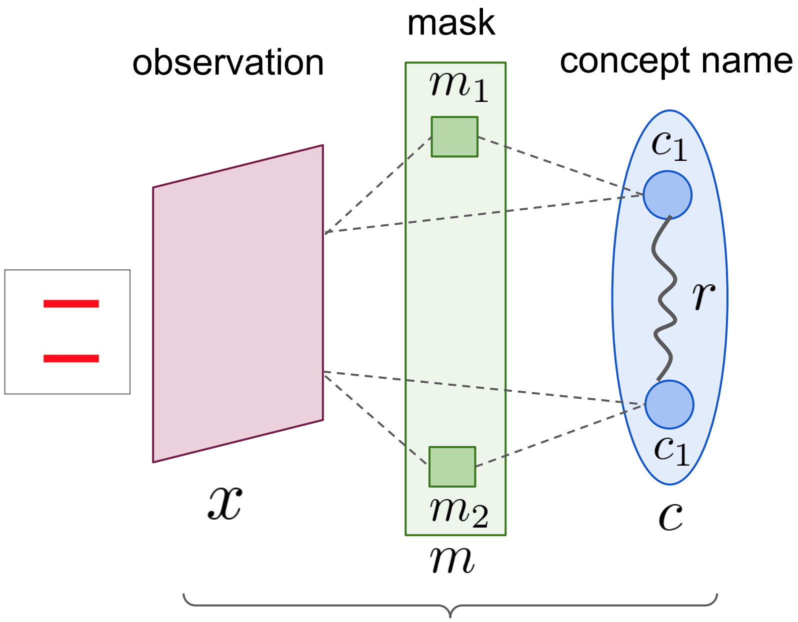
# Illustration of concept:



# Illustration of concept:



## Question: How to compose the probability function of a hierarchical concept?




$$\underbrace{f_{X,M_1,C}(x,m_1,c_1)\cdot f_{X,M,C}(x,m_2,c_1)\cdot f_{X,M_1,M_2,R}(x,m_1,m_2,r)}$$
$$= p_{X,M,C}(x,m,c)$$

Here the f are non-negative functions

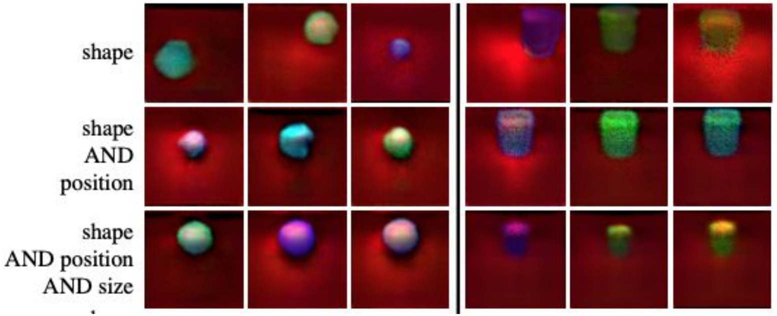
# Energy-based Models (EBM)

The probability function $p_\theta(x)$ can be written in terms of a energy-based model $E_\theta(x)$, where $E_\theta$ maps the input $x$ to a scalar value which we called energy.

$$p_\theta(x) \propto e^{-E_\theta(x)}$$

The benefit of using EBM is that multiplication of probability translates to addition of the energy terms:

$$p_{\theta_1}(x)p_{\theta_2}(x) \propto e^{-(E_{\theta_1}(x)+E_{\theta_2}(x))}$$


Du, Yilun, et al. Compositional Visual Generation with Energy Based Models, *NeurIPS 2020*: 6637-6647.

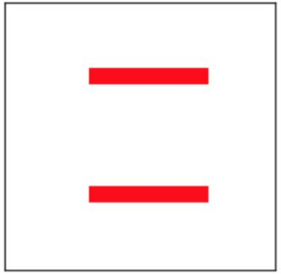
# ZeroC: Zero-shot Concept Recognition and Acquisition

ZeroC Hierarchical Composition Rule (e.g. “parallel-line”)

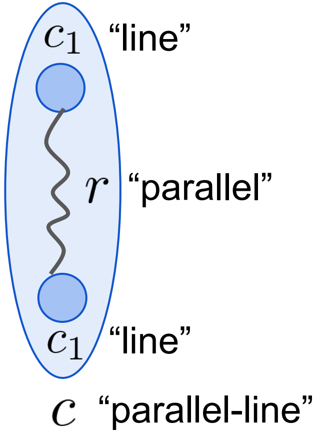
Summing the EBMs for all the nodes and edges together

$$E_{X,M,C}(x,m,c) = \underbrace{E_{X,M_1,C}(x,m_1,c_1)}_{\text{concept-EBM}} + \underbrace{E_{X,M_2,C}(x,m_2,c_1)}_{\text{concept-EBM}} + \underbrace{E_{X,M_1,M_2,R}(x,m_1,m_2,r)}_{\text{relation-EBM}}$$

Observation x:



Concept graph for “parallel-line”:



# ZeroC: Zero-shot Concept Recognition and Acquisition

x: input
m: mask
c: concept name
r: relation name

**Training:**

**Given:** data tuples of $(x, m_1, c_1)$ or $(x, m_1, m_2, r)$
**Learn:** energy-based model $E_{X,M_1,C}(x, m_1, c_1)$ or $E_{X,M_1,M_2,R}(x, m_1, m_2, r)$

We augment the state-of-the-art EBM training objective [1] with three more regularizations (from first principles) to learn:

$$L = \frac{1}{N}\sum_{n=1}^{N}\left(L_n^{(\text{Improved})} + \alpha_{\text{pos-std}}L_n^{(\text{pos-std})} + \alpha_{\text{em}}L_n^{(\text{em})} + \alpha_{\text{neg}}L_n^{(\text{neg})}\right)$$

make sure positive example have similar energy

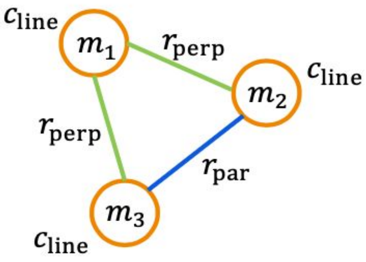
ensure consistency in concept acquisition

encourages "connected" masks

[1] Du, Yilun, et al. "Improved contrastive divergence training of energy based models." *ICML 2021*

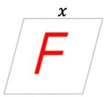
# ZeroC: Zero-shot Concept Recognition and Acquisition

**Inference:** (1) Zero-shot concept recognition




**Given:** graph structure of a hierarchical concept

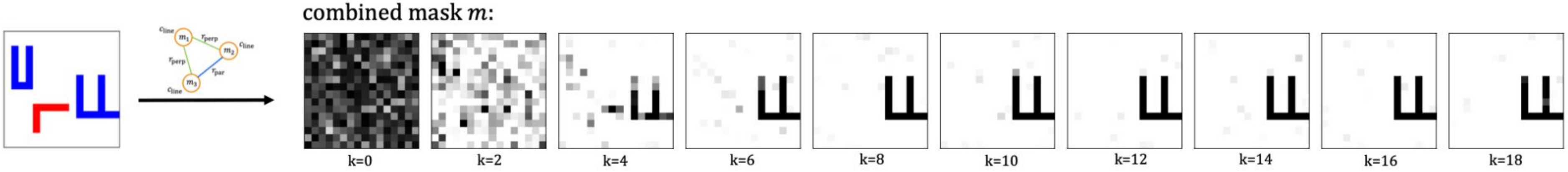
E.g. for the concept of "Fshape":




**Compose:** ZeroC first compose an EBM based on the given graph:

$$\boldsymbol{E}(\boldsymbol{x}, \boldsymbol{m}, \boldsymbol{c}_{\textbf{Fshape}}) = E\left(x, m_1, m_2, r_{\text{perp}}\right) + E\left(x, m_1, m_3, r_{\text{perp}}\right) + E\left(x, m_2, m_3, r_{\text{par}}\right) + \sum_{i=1,2,3} E\left(x, m_i, c_{\text{line}}\right)$$

**Detection:** (infer the mask given image x and concept name c):



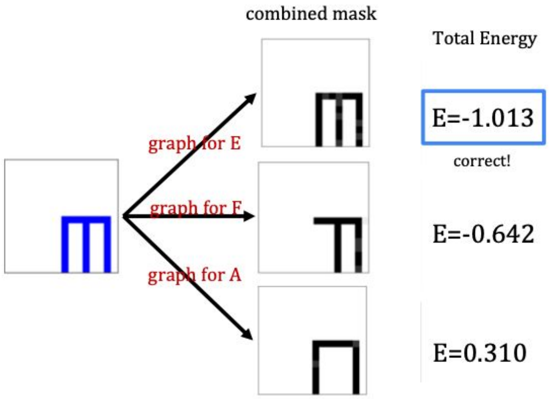
# ZeroC: Zero-shot Concept Recognition and Acquisition

**Inference:** (1) Zero-shot concept recognition

**Given:** graph structure of a hierarchical concept

E.g. for the concept of “Eshape”:

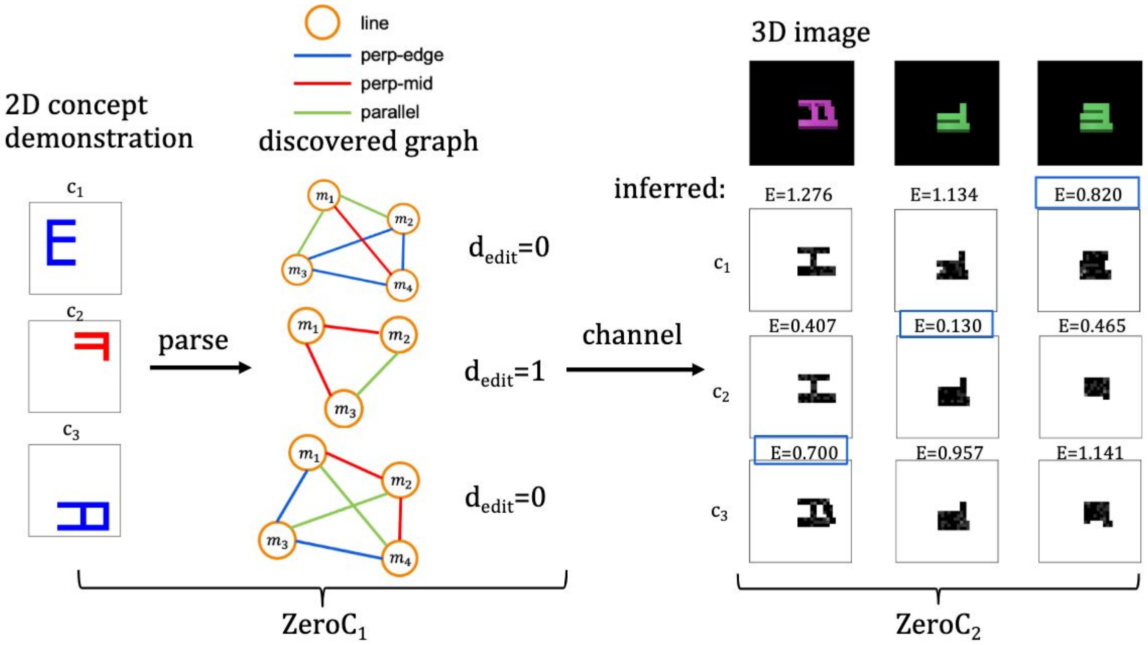
**Classification:**




correct!

# ZeroC: Zero-shot Concept Recognition and Acquisition

**Inference:** (2) Zero-shot concept acquisition

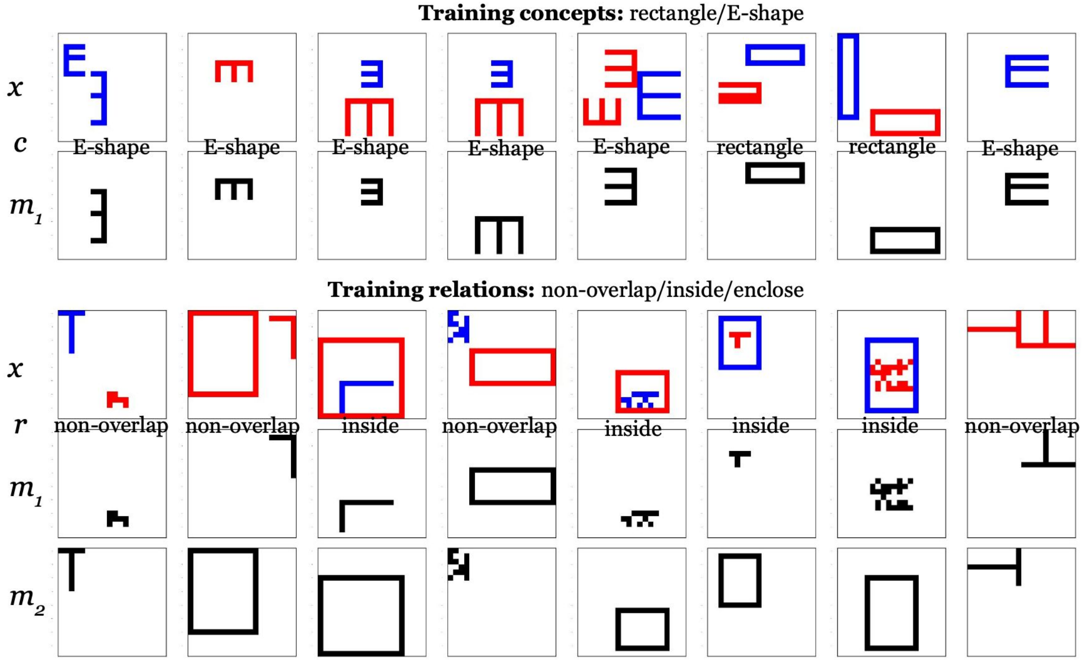



Difficult because it is a **NP-hard** subgraph isomorphism task

# Experiment 1: zero-shot recognition

Training dataset (HDConcept: elementary concepts and relations):

Training on concepts of “Eshape”, “rectangle” and relations of “inside”, “non-overlap”, “outside”:

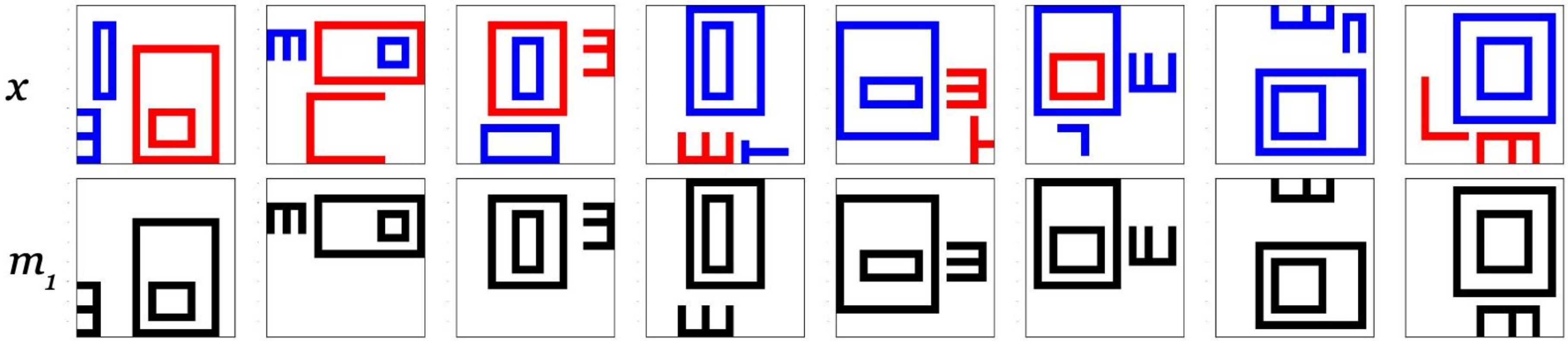


# Experiment 1: zero-shot recognition

Test dataset (HDConcept: hierarchical concepts):

Test on hierarchical concept (e.g. Concept1) that consists of “Eshape”, “rectangle” combined in certain way. E.g.:

**Inference:** Detecting Concept1 from distractor, given x and concept1’s concept graph



# Experiment 1: zero-shot recognition

| Model | Classification (acc.) HD-Letter | Classification (acc.) HD-Concept | Detection (IoU) HD-Letter+distractor | Detection (IoU) HD-Concept+distractor |
|---|---|---|---|---|
| Statistics | 46.5 | 53.5 | 5.69 | 12.6 |
| Heuristics | (–) | (–) | 42.3 | 29.2 |
| CADA-VAE [8] | 18.0 | 66.0 | (–) | (–) |
| **ZeroC (ours)** | **84.5** | **70.5** | **72.5** | **84.7** |
| ZeroC composition without R-EBM | 62.5 | 32.5 | 45.3 | 84.3 |
| ZeroC composition without HC-EBM | 67.0 | 55.0 | 67.7 | 78.4 |
| ZeroC without $L^{(\text{pos-std})}$ | 43.6 | 65.5 | 76.1 | 81.5 |
| ZeroC without $L^{(\text{neg})}$ | 64.5 | 59.0 | 60.0 | 84.2 |
| ZeroC without $L^{(\text{em})}$ | 81.5 | 61.0 | 68.0 | 86.0 |
| ZeroC with only $L^{(\text{Improved})}$ | 27.5 | 55.5 | 49.1 | 81.7 |

- ZeroC can zero-shot recognize hierarchical concepts with reasonable accuracy
- ZeroC outperforms the strong zero-shot learning baseline of CADA-VAE
- Ablation: The different components are necessary

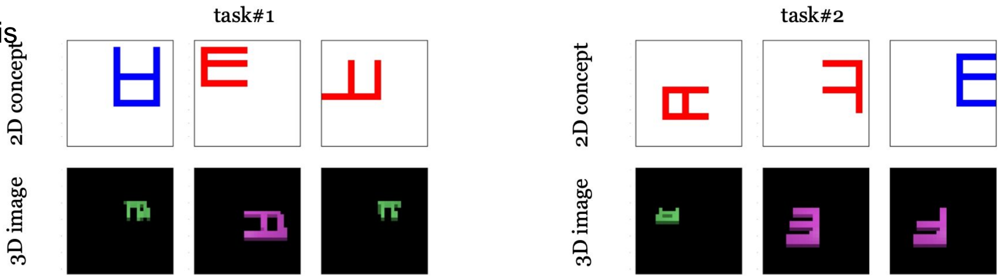
# Experiment 2: zero-shot acquisition

Table 2: Performance of models on acquiring concepts between models and domains at inference time (%).

| Model | Domain 1 (2D image) Parsing | |
|---|---|---|
| | Isomorphism (acc.) ↑ | Edit distance ↓ |
| Statistics | 2.33 | 3.14 |
| Mask R-CNN [13]+relation classification | 35.5 | 1.01 |
| **ZeroC$_1$ → ZeroC$_2$ (ours)** | **72.7** | **0.50** |
| ZeroC$_1$ without $L^{(\text{pos-std})}$ → ZeroC$_2$ | 55.2 | 1.57 |
| ZeroC$_1$ without $L^{(\text{neg})}$ → ZeroC$_2$ | 53.5 | 0.99 |
| ZeroC$_1$ without $L^{(\text{em})}$ → ZeroC$_2$ | 50.7 | 1.58 |
| ZeroC$_1$ with only $L^{(\text{Improved})}$ → ZeroC$_2$ | 11.5 | 2.00 |
| ZeroC$_2$ with ground-truth graph (upper bound) | (–) | (–) |

*we use a stringent **subgraph isomorphism accuracy** which is only 1 if the inferred graph is isomorphic to ground-truth.
An individual node/edge accuracy of 0.8 will result in overall accuracy of $0.8^{10} = 0.107$

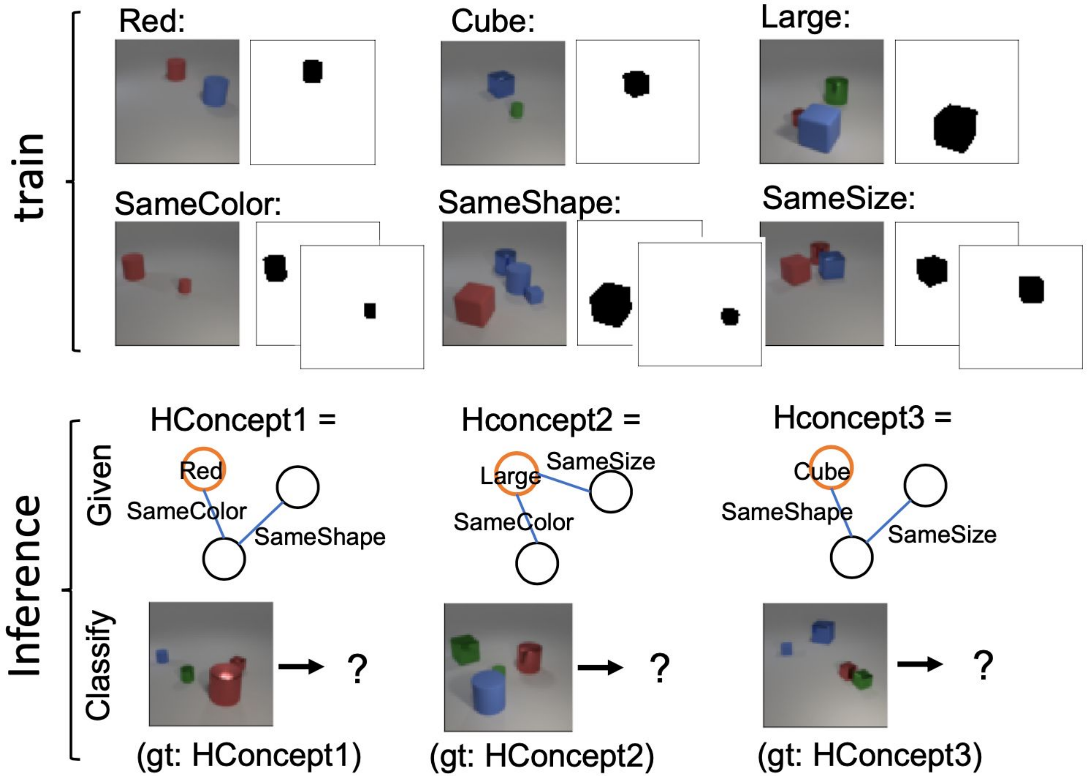
**Example task:**
2D to 3D transfer of concepts without training:



# Experiment 3: CLEVR dataset:

| Model | Classification acc (%) |
|---|---|
| Statistics | 33.4 |
| CADA-VAE | 45.3 |
| ZeroC (ours) | **56.0** |

ZeroC outperforms the strong baseline of CADA-VAE model, and able to reasonably classify the hierarchical concepts.



# Summary:

In this work, we introduce Zero-shot Concept Recognition and Acquisition (ZeroC), a neuro-symbolic architecture that can recognize and acquire novel concepts in a zero-shot way.

It is able to perform:

- **Zero-shot recognition:** recognize more complex concepts at inference, without further training
- **Zero-shot acquisition:** discover the internal structure of more complex concepts at inference, and transfer the knowledge across domains.

For more, see our paper and project page at
[http://snap.stanford.edu/zeroc/](http://snap.stanford.edu/zeroc/), or SCAN the QR code: