We propose **SORNet**: Spatial Object-centric Representation Network to learn object-centric embeddings that encode spatial relationships.

**Overview**

- **SORNet** generalizes zero-shot to scenes with unseen objects and different number of objects.
- **SORNet** is trained only on classification of logical predicates, but captures continuous spatial relationships.

**Approach**

**Embedding Network**

- Permutation-invariant object embedding

**Readout Networks**

- Number of outputs changes adaptively with number of inputs

**Results**

**Spatial Relation Prediction on CLEVR-CoGenT**

**Training** (Condition A)

- Cubes are gray, blue, brown, or yellow
- Cylinders are red, green, purple, or cyan
- Spheres can have any color

**Testing** (Condition B)

- Cubes are red, green, purple, or cyan
- Cylinders are gray, blue, brown, or yellow
- Spheres can have any color

**Readout Networks**

- Number of outputs changes adaptively with number of inputs

**Predicate Classification on Leonardo**

**Training** split

- Overall 405 colored blocks
- Randomly chosen 4 blocks in each sequence
- One task - stacking 4 blocks
- 133/96 sequences.

**Testing** split

- 7 colored blocks (unseen in train)
- Randomly chosen 4-6 blocks in each sequence
- 7 tasks different from training
- 95/26 sequences.

**Objective**

- Classifying logical predicates from RGB input

**Spatial Relation Prediction on CLEVR-CoGenT**

**Output**

Spatial Relations

**Input**

RGB frame

Canonical Object Views

**SORNet**

- left_of
- behind

**MDERK [34]**

<table>
<thead>
<tr>
<th>ValA Accuracy</th>
<th>ValB Accuracy</th>
<th>MDERK (Oracle) [34]</th>
</tr>
</thead>
<tbody>
<tr>
<td>84.950</td>
<td>59.627</td>
<td>99.006</td>
</tr>
</tbody>
</table>

**Zero-shot Accuracy**

<table>
<thead>
<tr>
<th>Scene</th>
<th>Method</th>
<th># pred</th>
<th>all</th>
<th>no surface</th>
<th>has object &amp; top</th>
<th>obj &amp; top</th>
<th>stacked</th>
<th>aligned</th>
<th>approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>1200</td>
<td>ResNet18 M-Head 100-shot</td>
<td>32</td>
<td>83.2</td>
<td>92.2</td>
<td>79.7</td>
<td>93.0</td>
<td>92.2</td>
<td>83.8</td>
<td>74.9</td>
</tr>
<tr>
<td></td>
<td>VIT-B/32 M-Head M-View 100-shot</td>
<td>32</td>
<td>83.2</td>
<td>92.2</td>
<td>79.7</td>
<td>93.0</td>
<td>92.2</td>
<td>83.8</td>
<td>74.9</td>
</tr>
<tr>
<td></td>
<td>SORNet 6-shot</td>
<td>32</td>
<td>83.2</td>
<td>92.2</td>
<td>79.7</td>
<td>93.0</td>
<td>92.2</td>
<td>83.8</td>
<td>74.9</td>
</tr>
<tr>
<td></td>
<td>SORNet M-View 1-shot</td>
<td>32</td>
<td>83.2</td>
<td>92.2</td>
<td>79.7</td>
<td>93.0</td>
<td>92.2</td>
<td>83.8</td>
<td>74.9</td>
</tr>
<tr>
<td></td>
<td>SORNet M-View 2-shot</td>
<td>32</td>
<td>83.2</td>
<td>92.2</td>
<td>79.7</td>
<td>93.0</td>
<td>92.2</td>
<td>83.8</td>
<td>74.9</td>
</tr>
<tr>
<td></td>
<td>SORNet M-View 3-shot</td>
<td>32</td>
<td>83.2</td>
<td>92.2</td>
<td>79.7</td>
<td>93.0</td>
<td>92.2</td>
<td>83.8</td>
<td>74.9</td>
</tr>
<tr>
<td></td>
<td>SORNet M-View 4-shot</td>
<td>32</td>
<td>83.2</td>
<td>92.2</td>
<td>79.7</td>
<td>93.0</td>
<td>92.2</td>
<td>83.8</td>
<td>74.9</td>
</tr>
<tr>
<td></td>
<td>SORNet M-View 5-shot</td>
<td>32</td>
<td>83.2</td>
<td>92.2</td>
<td>79.7</td>
<td>93.0</td>
<td>92.2</td>
<td>83.8</td>
<td>74.9</td>
</tr>
<tr>
<td></td>
<td>SORNet M-View 6-shot</td>
<td>32</td>
<td>83.2</td>
<td>92.2</td>
<td>79.7</td>
<td>93.0</td>
<td>92.2</td>
<td>83.8</td>
<td>74.9</td>
</tr>
</tbody>
</table>

**MDERK [34]**

<table>
<thead>
<tr>
<th>ValA Accuracy</th>
<th>ValB Accuracy</th>
<th>MDERK (Oracle) [34]</th>
</tr>
</thead>
<tbody>
<tr>
<td>84.950</td>
<td>59.627</td>
<td>99.006</td>
</tr>
</tbody>
</table>

**MDERK (Oracle) [34]**

<table>
<thead>
<tr>
<th>ValA Accuracy</th>
<th>ValB Accuracy</th>
<th>MDERK (Oracle) [34]</th>
</tr>
</thead>
<tbody>
<tr>
<td>84.950</td>
<td>59.627</td>
<td>99.006</td>
</tr>
</tbody>
</table>