# Enhancing Interpretable Object Abstraction via Clustering-based Slot Initialization

Ning Gao[1,2] Bernard Hohmann[2] Gerhard Neumann[2]

[1]Bosch Center for Artificial Intelligence [2]Autonomous Learning Robots, KIT

BMVC 2023

## Problem

Object-centric representations using slots have shown advances towards efficient, flexible and interpretable abstraction from low-level perceptual features in a compositional scene.

However, a few limitations remain:

1. Random initialization requires more interactive refinement and leads to a slow inference.
2. Fixed number of slots need to be predefined.

## Contributions

**Intuiition.** Visual input includes a strong inductive bias about the represented scene.

1. We propose the conditional slot initialization using clustering algorithms instead of random initialization.
2. We analyze the effect of permutation symmetry including invariance and equivariance on the object-centric slot representations.
3. We apply mean-shift clustering on the perceptual features which allows to generate flexible number of slots.

## Method

We initialize slot representations conditioned on the input features. **Symmetric Geometry.** The order of the predicted slots should either remain the same w.r.t. the permutation of the cluster centers (**permutation invariance**) or change correspondingly in the same order as the cluster centers (**permutation equivariance**).

- Non-permutation symmetric (*kmeans*) K-means is applied on the pixel-wise convolutional perceptual feature to get the feature-based cluster centers, which are mapped to $K$ slots.


- Permutation-invariant (*pseudoweights*) A simple mapping between M cluster centers and K slots breaks the permutation symmetry and cannot generalize well on unseen objects. We propose a permutation-invariant model named Pseudoweights, which uses positional encoding to identify different slots.

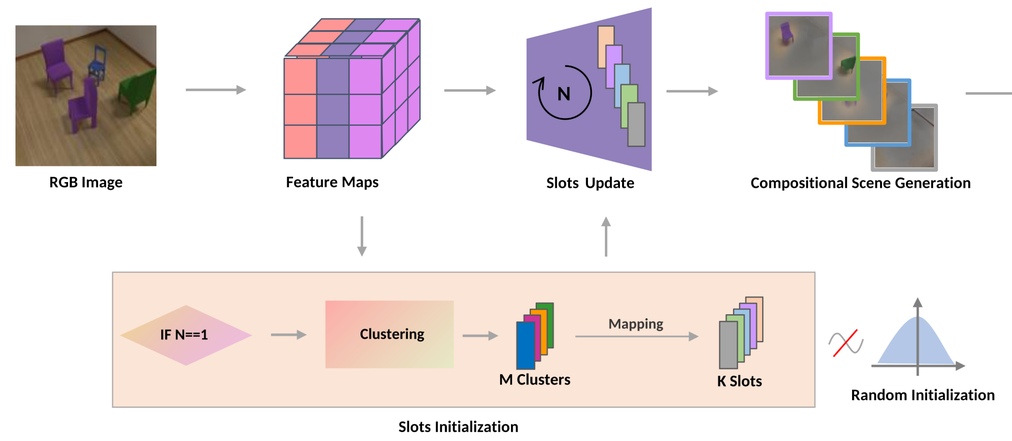

$$\mathbf{p}_k = \left(\sin\left(\frac{\pi}{D'}k\right), \cos\left(\frac{\pi}{D'}k\right), ..., \sin\left(\pi k\right), \cos(\pi k)\right) \tag{1}$$

The slots are then initialized as the weighted sum over the cluster centers by $\mathbf{w}$:

$$\mathbf{z}_k = \sum_{m=1}^{M} \mathbf{w}_{k,m} \cdot \mathbf{c}_{k,m}. \tag{2}$$

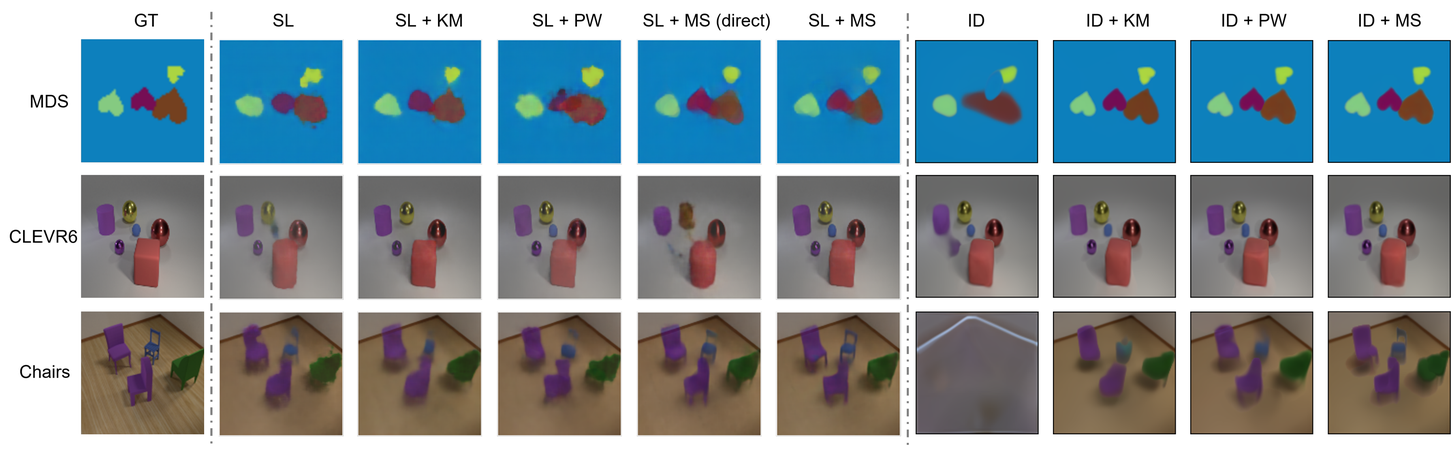
- Permutation equivariant (*mean-shift*) To further improve the generalization, we perform the mean-shift clustering algorithm over the feature space to determine the number of cluster centers without predefined. A shared mapping layer is then applied to initialize the slots based on each cluster respectively.


## Image-Conditioned Slot Initialization



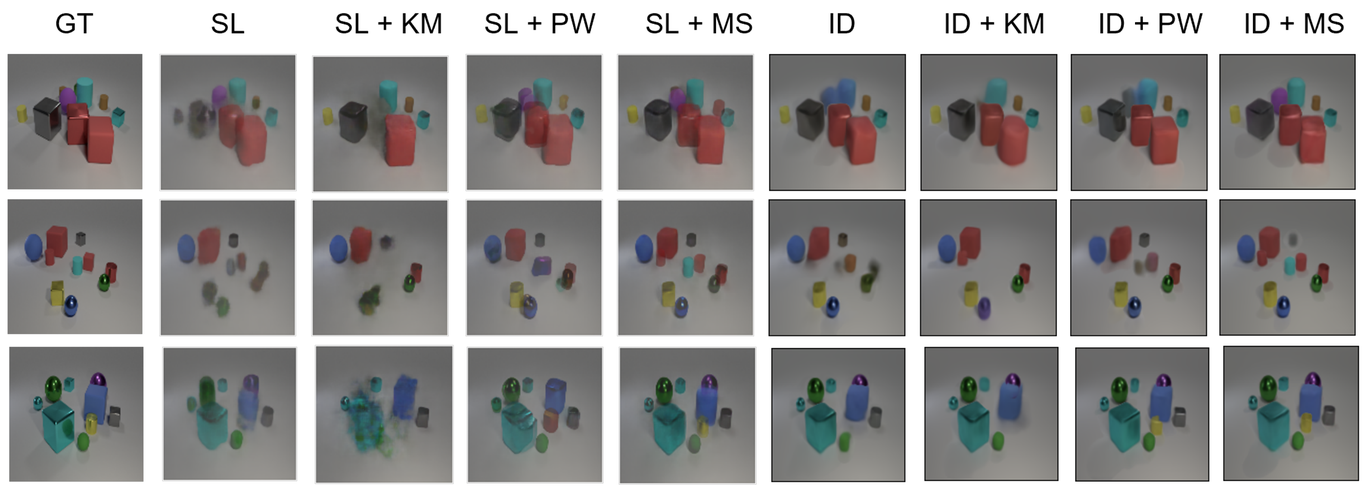

**The network architecture.** Instead of randomizing slot initialization from a common distribution widely used in prior work, we initialize slot representations conditioned on the input features. A clustering algorithm and a mapping layer are adopted.

## Results




**Qualitative results on the object discovery task.**

- Our method demonstrates better performance than baselines.
- Our method can reconstruct more details, even better quality than the original input on MDS dataset.



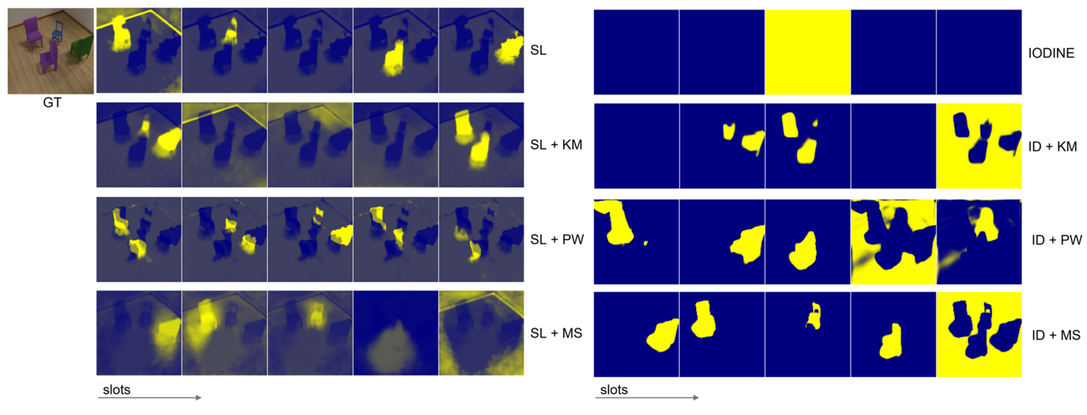

**Qualitative results on increasing objects.**

- Baselines struggle with overlapped objects.
- Our models can accurately detect overlapped objects even for difficult scenes.

| | Iter 1 | | | | Iter 3 | | | | Iter 7 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | $\text{LPIPS}_A \downarrow$ | $\text{LPIPS}_V \downarrow$ | PSNR ↑ | SSIM ↑ | $\text{LPIPS}_A \downarrow$ | $\text{LPIPS}_V \downarrow$ | PSNR ↑ | SSIM ↑ | $\text{LPIPS}_A \downarrow$ | $\text{LPIPS}_V \downarrow$ | PSNR ↑ | SSIM ↑ |
| ID | 0.4415 | 0.6071 | 12.72 | 0.3820 | 0.4477 | 0.5804 | 16.33 | 0.4908 | 0.4363 | 0.5646 | 19.53 | 0.5001 |
| ID + kmeans | 0.2108 | 0.3768 | 27.05 | 0.6202 | 0.1956 | 0.3607 | 28.75 | 0.6533 | 0.1884 | 0.3545 | 29.33 | 0.6656 |
| ID + PW | 0.2269 | 0.3734 | 27.57 | 0.6297 | 0.1973 | 0.3531 | 29.33 | 0.6642 | 0.1885 | 0.3461 | 29.92 | 0.6768 |
| ID + MS | **0.1798** | **0.3545** | **28.39** | **0.6467** | **0.1602** | **0.3343** | **30.16** | **0.6828** | **0.1528** | **0.3273** | **30.68** | **0.6951** |

**Evaluation with different number of iterations** (5 iterations are used for training).

- All models are capable of generalizing on more iterations with performance gains.
- Our method gains notable improvement at the first iteration, indicating the efficiency of the learned inductive slot initialization.




**Slot-wise mask prediction.**

- MS has better disentanglement than others and reconstructs more details.

Please refer to our paper for more details.