Multi-Object Representation Learning via Feature Connectivity and Object-Centric Regularization

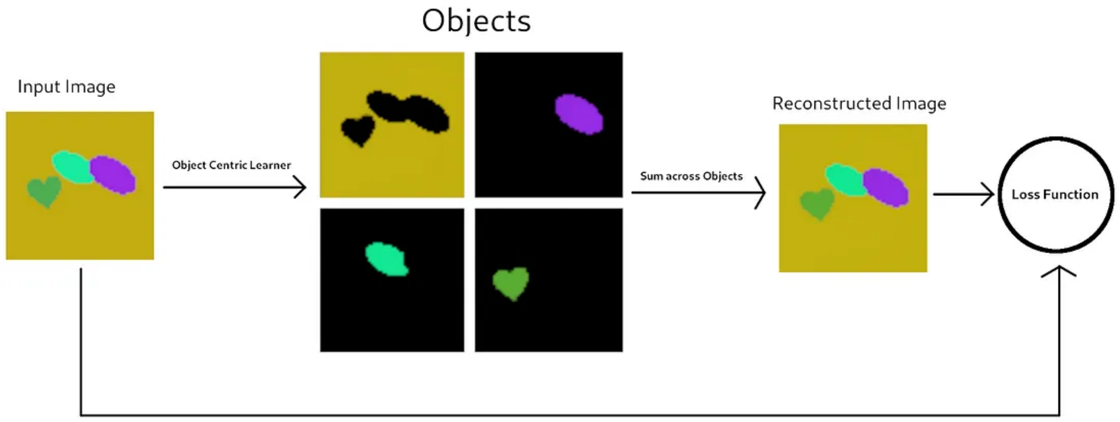
Alex Foo, Wynne Hsu, Mong Li Lee
School of Computing, National University of Singapore

Why Multi-Object Representation Learning?

- Human understanding of the world relies on objects as compositional building blocks [Kahneman et al., 1992]
- Emulating this in machine learning algorithms through object-centric representations can improve [Greff et al., 2020]:
  - Robustness
  - Sample Efficiency
  - Generalization to out-of-domain distributions
  - Interpretability

Limitations of Current Approaches

- Recent work utilizes a generative approach which optimizes pixel-based reconstruction to learn object-centric representations [Yuan et al., 2022]
- Pixel-based reconstruction prioritizes pixel accuracy over object discovery and functional feature extraction, which may lead to failure in:
  - Discovering objects [Locatello et al., 2020]
  - Obtaining useful object features such as position and shape [Kabra et al., 2019]

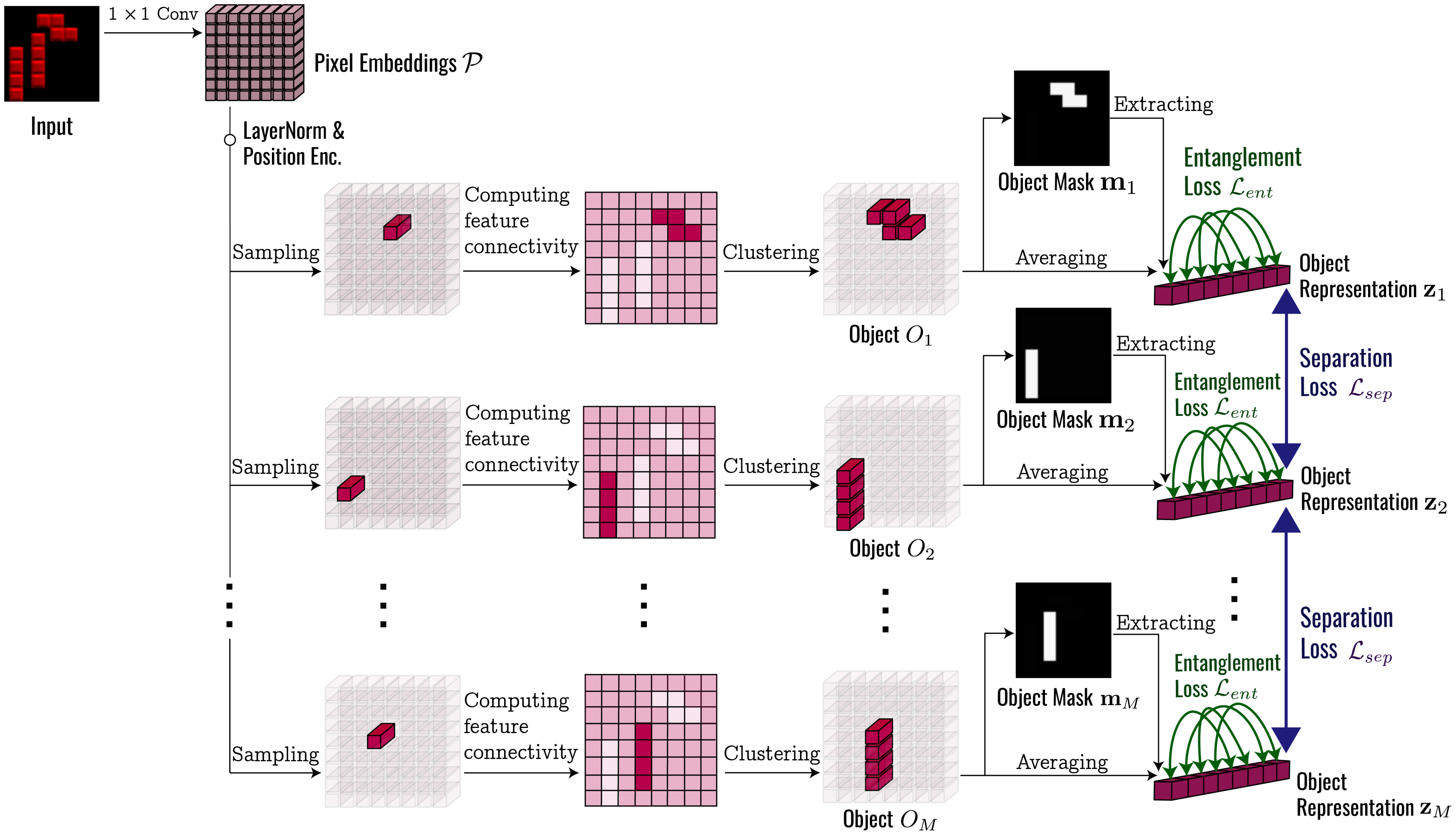


Michele De Vita: Introduction to Object Centric Learning

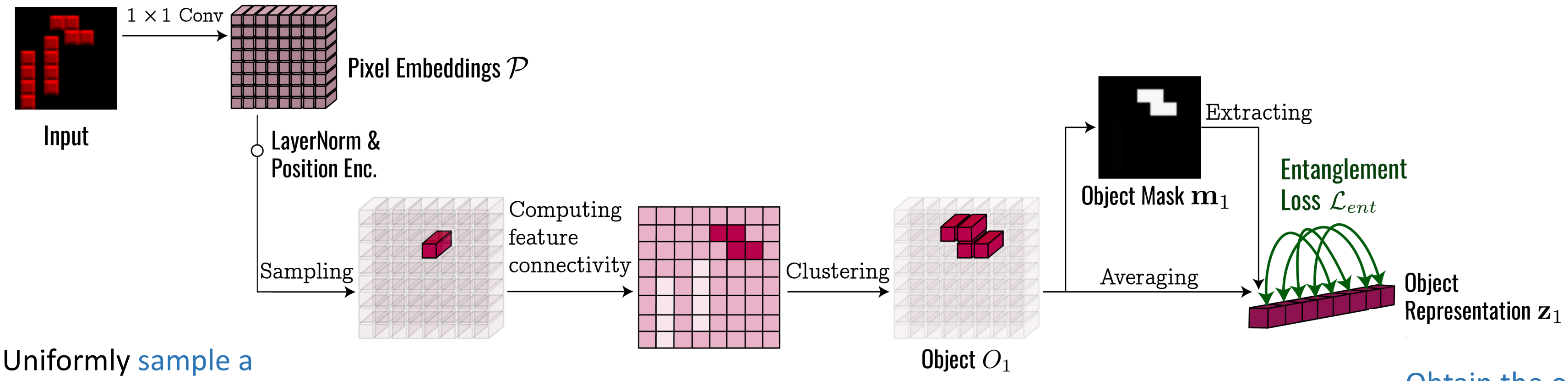
Key Contributions

- We propose a framework which leverages on feature connectivity and design two object-centric regularization terms
- We demonstrate that proposed approach:
  - Outperforms SOTA methods in discovering multiple objects from simulated, real-world, complex texture and common object images in a fine-grained manner without supervision
  - Attains sample efficiency and is generalizable to out-of-domain images
  - Learned object representations accurately predict key object properties in downstream tasks

OC-Net: Overview



OC-Net: Object Discovery



Uniformly sample a pixel embedding yet to be assigned to an object

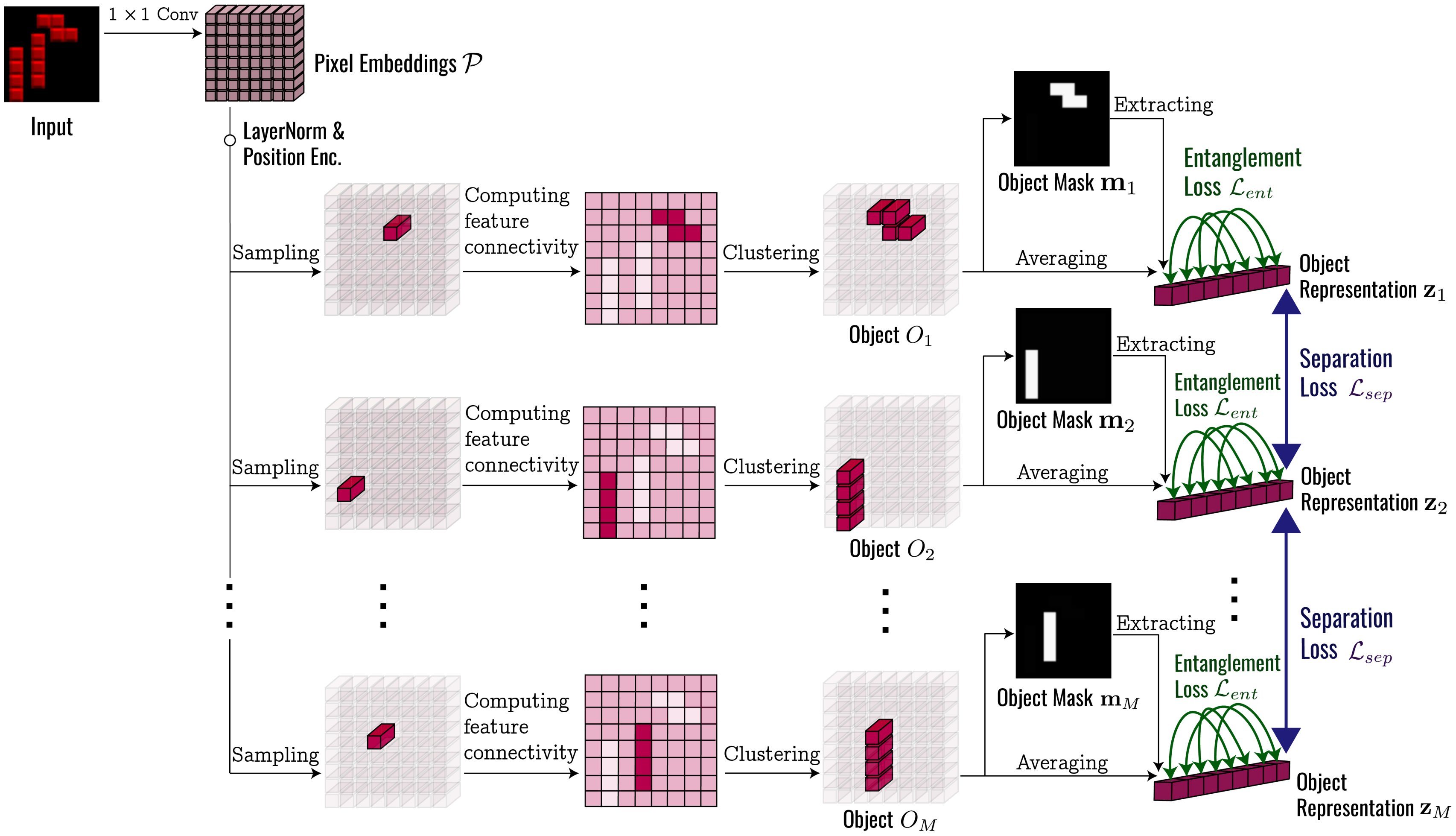
Use Dijkstra's algorithm to compute the shortest distance of the sampled pixel embedding to all other embeddings, where distance between a pair of neighbouring pixels is defined as:

$$\mathrm{sim}(\mathbf{p}_m, \mathbf{p}_k) = \sqrt{\sum_{d=1}^{D}(\mathbf{p}_m[d] - \mathbf{p}_k[d])^2}$$

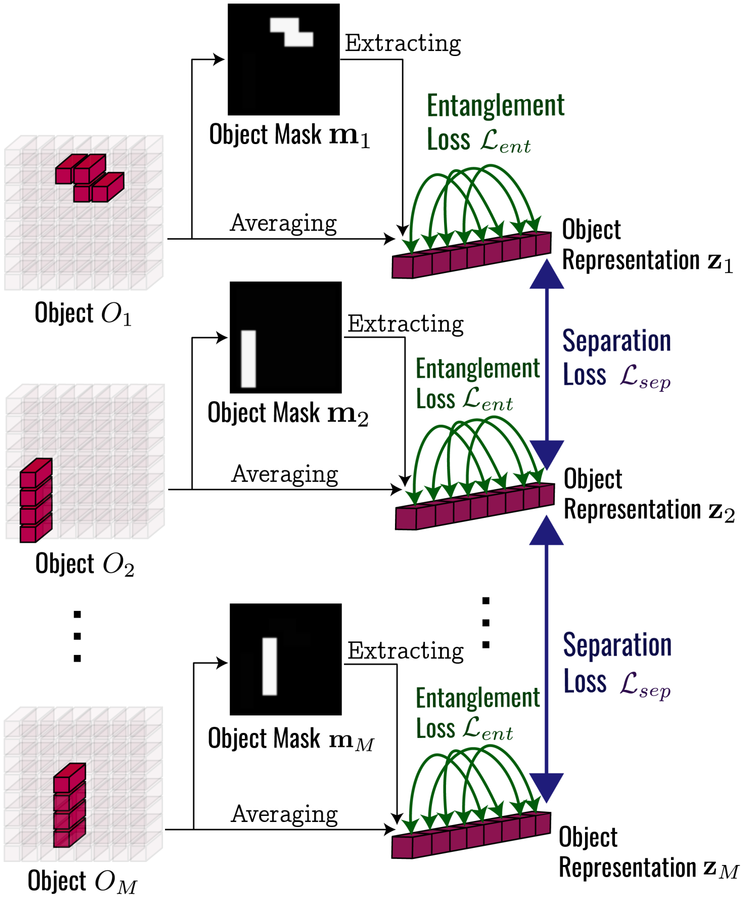
Consider pixel embeddings to be part of the same object as the sampled pixel embedding according to a threshold

Obtain the object representation by taking the sum of the extracted mask information and the average of pixel embeddings

OC-Net: Object Discovery

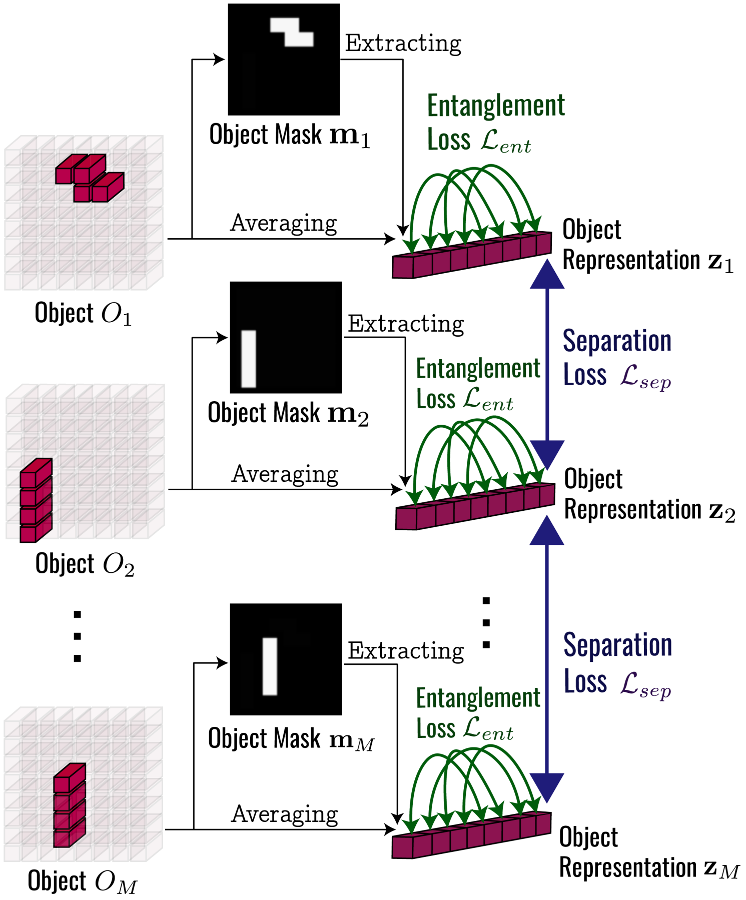


OC-Net: Object-Centric Regularization



We design two object-centric regularization terms to improve the quality of the learned object representations for downstream generalization and object discovery

OC-Net: Object-Centric Regularization



Theorem 3.1. *Let* $\mathbf{Y}$ *be the matrix of labels of the training samples and* $\mathbf{P_Z}$ *be the projection matrix of* $\mathbf{Z}$*:*

$$\mathbf{P_Z} = \mathbf{I} - \mathbf{Z}^\top(\Sigma_\mathbf{Z})^\dagger \mathbf{Z}, \tag{5}$$

where $\mathbf{I}$ *is the identity matrix,* $(.)^\dagger$ *is the pseudoinverse and* $\Sigma_\mathbf{Z} = \mathbf{Z}\mathbf{Z}^\top$ *is the unnormalized covariance matrix of* $\mathbf{Z}$*. Let* $||.||_F$ *be the Frobenius norm. Then, the following relation holds:*

$$\delta_\mathbf{Z} \leq ||\mathbf{P_Z}||_F ||\mathbf{Y}||_F. \tag{6}$$

- L_{sep} maximizes the distance between object representations in the latent space, encouraging the model to learn distinct and non-overlapping object representations:

$$\mathcal{L}_{sep} = \frac{1}{D}\sum_{d=1}^{D} \max(0, 1 - \sqrt{\sigma_d + \tau})$$

- L_{ent} minimizes the correlation between dimensions in the latent space **Z**, achieving more disentangled object representations which are easier to manipulate and analyze:

$$\mathcal{L}_{ent} = \frac{1}{D \times (M-1)} \sum_{i \neq j} \Sigma_\mathbf{Z}[i,j]$$

Performance Study

Experiment Setup: Datasets

Dataset	Type	Ground Truth	Image Size	# Samples
Multi-dSprites	Simulated	Pixel Mask	64×64	1M
Tetrominoes-NM	Simulated	Pixel Mask	35×35	1M
SVHN	Real-World	Bounding Box	Varied	530K
IDRiD	Real-World	Pixel Mask	4288×2848	81
CLEVRTEX	Complex Texture	Pixel Mask	128×128	50K
CLEVRTEX-OOD	Complex Texture	Pixel Mask	128×128	10K
Flowers	Common Object	Pixel Mask	128×128	7K
Birds	Common Object	Pixel Mask	128×128	11K
COCO	Common Object	Pixel Mask	128×128	12K

Experiments on Quality of Discovered Objects

Table 2: Evaluation scores for the discovered foreground objects.

(a) Simulated datasets

	Multi-dSprites			Tetrominoes-NM		
Method	ARI	mDice	mIoU	ARI	mDice	mIoU
SLIC	67.9±0.0	78.5±0.0	70.8±0.0	53.0±0.0	66.1±0.0	53.6±0.0
Felzenszwalb	97.4±0.0	98.6±0.0	95.0±0.0	95.0±0.0	98.0±0.0	96.9±0.0
Slot Attention	91.3±0.3	45.7±0.7	32.6±0.6	99.8±0.1	41.5±0.8	26.6±0.7
EfficientMORL	85.2±0.5	30.1±1.3	19.5±1.1	99.0±1.7	42.5±2.3	27.6±1.9
GENESIS-V2	85.0±1.3	81.5±1.9	72.2±1.4	97.6±0.5	47.1±1.1	31.0±0.8
SLATE	89.5±1.2	82.5±0.9	72.6±1.1	84.5±1.5	57.8±0.9	44.3±0.8
SysBinder	72.3±1.2	30.6±1.1	19.6±1.0	90.7±1.7	41.8±1.9	27.0±1.7
BO-QSA	90.4±1.1	91.6±1.1	88.0±1.2	99.3±0.3	40.9±1.4	25.8±1.2
OC-Net	**99.8±0.0**	**99.5±0.0**	**99.1±0.0**	**100.0±0.0**	**100.0±0.0**	**100.0±0.0**

(b) Real-world datasets

	SVHN			IDRiD		
Method	ARI	mDice	mIoU	ARI	mDice	mIoU
SLIC	5.3±0.0	50.1±0.0	34.5±0.0	32.2±0.0	12.7±0.0	8.8±0.0
Felzenszwalb	31.7±0.0	51.6±0.0	39.8±0.0	14.7±0.0	19.0±0.0	15.4±0.0
Slot Attention	38.9±1.5	51.7±1.8	36.7±1.7	28.7±1.1	8.6±1.7	5.0±1.6
EfficientMORL	32.2±1.7	49.2±2.0	34.0±1.8	16.8±1.5	11.1±2.7	7.0±1.8
GENESIS-V2	28.6±1.4	60.8±1.5	45.9±1.4	18.3±1.6	8.8±1.9	5.4±1.6
SLATE	21.2±1.2	57.0±1.3	41.7±1.5	35.6±2.1	8.1±1.2	4.7±1.8
SysBinder	15.8±1.6	49.5±1.9	34.1±1.8	25.2±1.3	16.6±1.7	11.1±1.8
BO-QSA	24.3±1.2	62.0±1.6	48.3±1.3	27.7±2.0	7.0±1.9	4.5±1.7
OC-Net	**39.7±0.1**	**64.6±0.1**	**49.9±0.1**	**39.0±0.4**	**38.1±0.2**	**31.2±0.2**

Experiments on Quality of Discovered Objects

(c) Complex textures dataset

Method	CLEVRTEX ARI	CLEVRTEX mDice	CLEVRTEX mIoU	CLEVRTEX-OOD ARI	CLEVRTEX-OOD mDice	CLEVRTEX-OOD mIoU
SLIC	27.4±0.0	20.0±0.0	13.0±0.0	25.8±0.0	21.7±0.0	14.0±0.0
Felzenszwalb	57.3±0.0	33.6±0.0	26.8±0.0	44.6±0.0	29.6±0.0	23.4±0.0
Slot Attention	58.6±1.6	35.0±1.6	26.7±1.5	51.3±1.9	34.1±1.4	25.1±1.3
EfficientMORL	59.5±1.7	37.7±1.5	31.1±1.4	53.9±2.5	32.2±2.4	25.3±2.8
GENESIS-V2	65.6±1.8	36.9±1.4	30.4±1.4	67.6±1.6	34.2±1.5	27.6±1.9
SLATE	57.5±1.8	33.3±1.6	24.4±1.5	56.6±1.3	34.7±2.1	25.3±1.8
SysBinder	61.4±1.7	31.3±1.5	23.1±1.4	61.0±2.4	32.3±2.0	23.8±1.8
BO-QSA	**70.9±1.9**	42.9±1.8	34.7±1.7	66.1±1.3	42.8±1.4	33.9±1.3
OC-Net	**70.7±0.9**	**45.1±0.9**	**37.5±0.7**	**69.8±0.8**	**43.5±0.7**	**35.0±0.6**

(d) Common objects datasets

Method	Flowers Dice	Flowers IoU	Birds Dice	Birds IoU	COCO mDice	COCO mIoU
SLIC	30.5±0.0	18.4±0.0	33.1±0.0	20.3±0.0	36.2±0.0	24.4±0.0
Felzenszwalb	43.7±0.0	30.4±0.0	34.3±0.0	23.0±0.0	36.6±0.0	27.1±0.0
Slot Attention	43.0±1.5	28.6±1.2	42.9±2.0	27.9±1.8	24.8±2.0	15.0±1.7
EfficientMORL	59.5±2.1	45.2±2.2	44.0±1.9	30.8±1.8	28.4±2.3	18.9±2.1
GENESIS-V2	63.7±2.2	50.2±2.2	41.4±1.7	27.7±1.5	25.1±2.1	16.1±1.7
SLATE	55.6±1.2	40.8±1.8	39.5±1.5	25.9±1.8	37.0±1.9	24.4±1.8
SysBinder	45.0±1.8	30.8±1.6	33.7±1.3	21.1±2.0	18.4±1.6	10.7±1.4
BO-QSA	65.8±1.9	51.7±1.9	44.6±1.7	30.3±1.5	34.9±1.1	23.6±0.9
OC-Net	**67.2±0.2**	**54.4±0.2**	**47.8±0.2**	**33.5±0.2**	**48.2±0.2**	**35.6±0.2**

Experiments on Quality of Discovered Objects

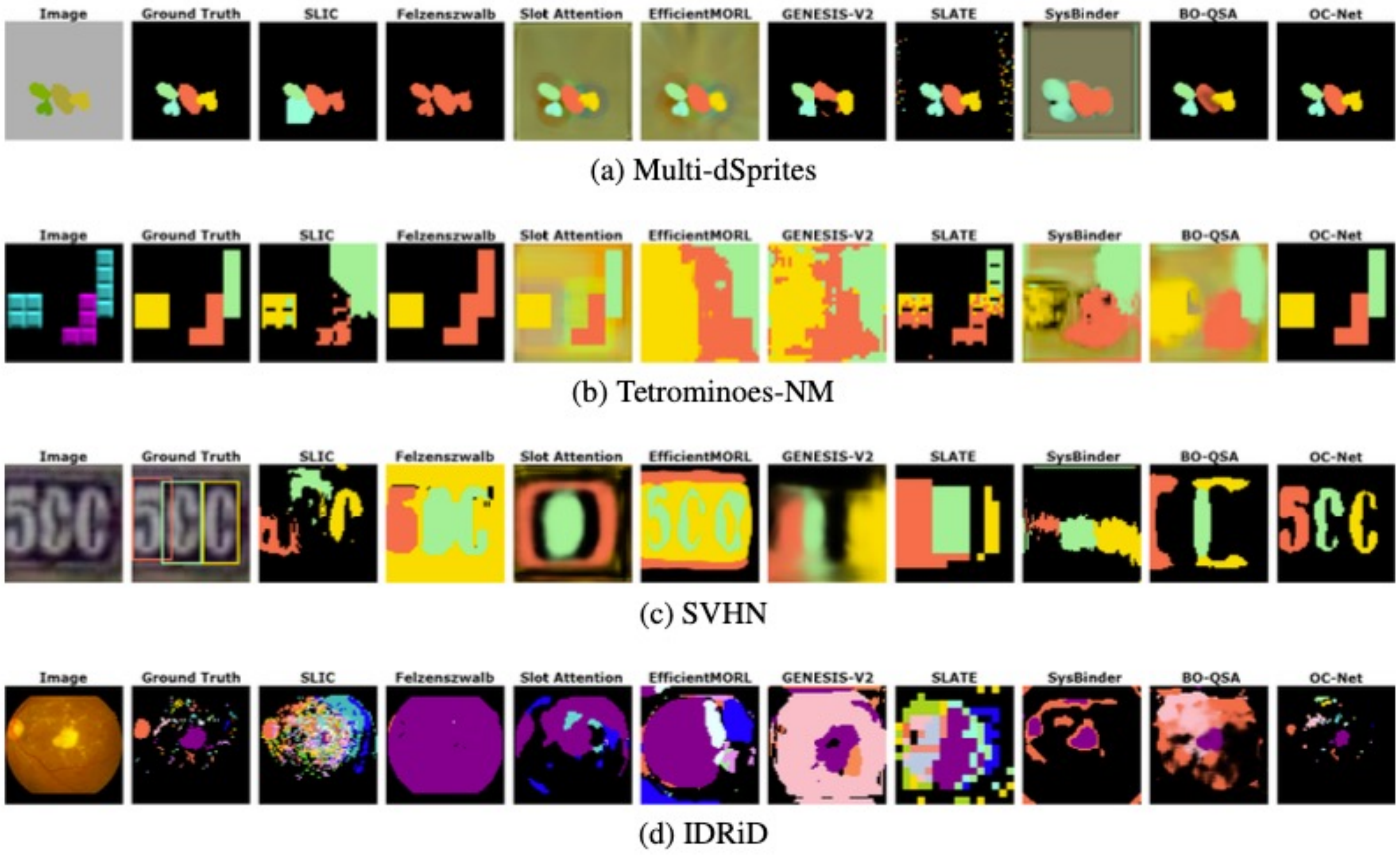
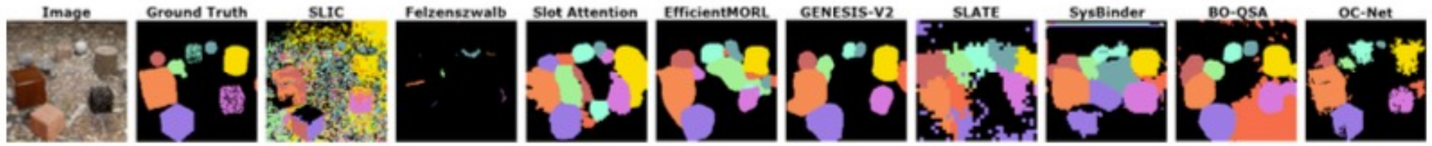
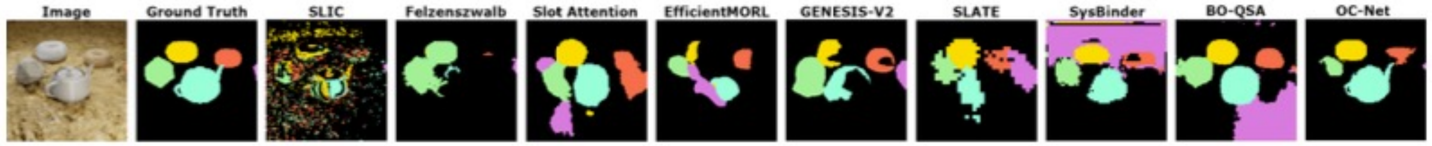


Figure 2: Visualization of discovered objects.

Experiments on Quality of Discovered Objects

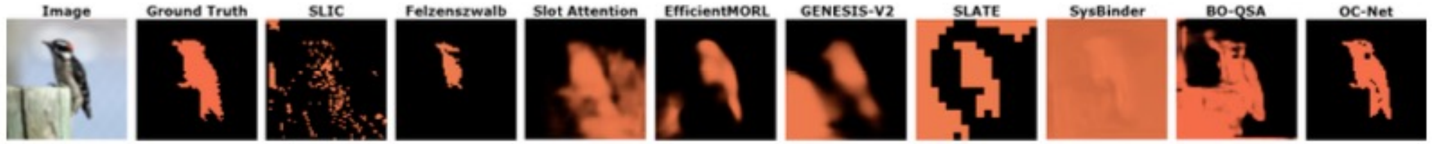


(e) CLEVRTEX



(f) CLEVRTEX-OOD

(g) Flowers

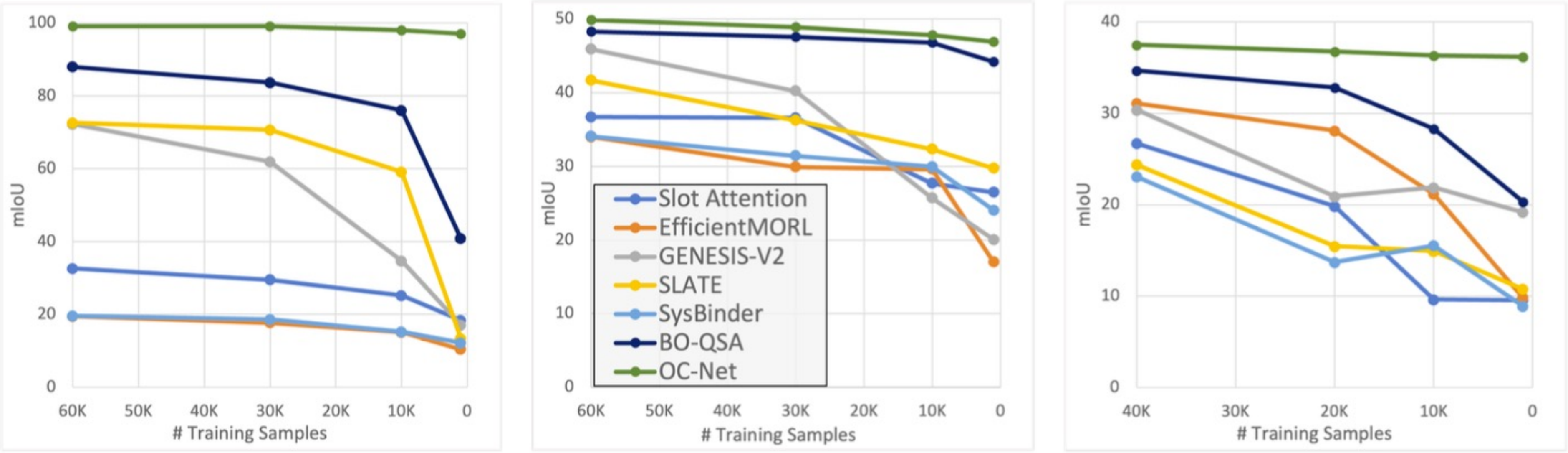


(h) Birds

(i) COCO

Experiments on Sample Efficiency

- One obstacle to unsupervised object discovery is the availability of a sufficiently large number of suitable training samples.
- Sample efficiency refers to a model's ability to learn effectively from a relatively small number of examples.



(a) Multi-dSprites (b) SVHN (c) CLEVRTEX

Figure 3: mIoU scores vs decreasing number of training samples.

Experiments on Model Generalizability

- In this set of experiments, we compare the generalization ability of OC-Net with the baselines by training the models on Multi-dSprites and testing them on the other datasets.

Table 3: mIoU scores for model generalizability after training on Multi-dSprites.

Method	Tetrominoes-NM	SVHN	IDRiD	CLEVRTEX	CLEVRTEX-OOD
Slot Attention	21.8±3.5	19.5±3.8	7.5±2.5	12.2±2.2	12.3±2.2
EfficientMORL	21.2±3.8	23.4±2.8	6.5±2.6	12.7±3.2	15.2±2.0
GENESIS-V2	42.9±4.9	31.1±2.8	8.5±2.4	21.9±1.6	21.3±2.5
SLATE	51.4±1.6	21.1±2.0	10.0±1.7	12.7±2.2	12.9±1.8
SysBinder	28.5±1.8	23.8±1.1	13.9±1.8	10.6±1.5	11.6±1.9
BO-QSA	41.8±1.8	24.3±2.0	4.0±1.5	24.4±1.4	22.8±2.5
OC-Net	**100.0±0.0**	**47.5±0.5**	**29.1±0.5**	**31.7±0.6**	**31.3±0.6**

Prediction based on Learned Object Representation

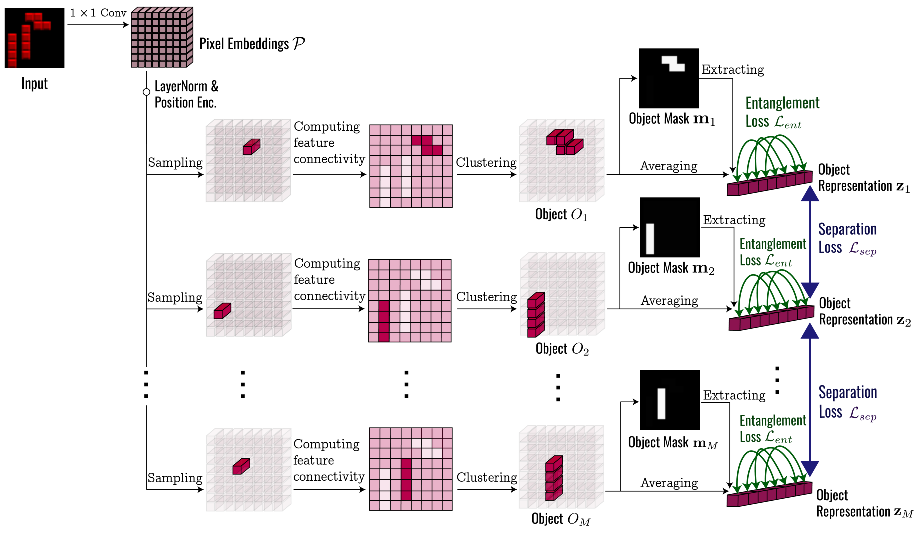
- One characteristic of an effective object-centric representation is its ability to encode object properties such as color, position and shape [Schölkopf et al., 2021].

Table 5: R^2 scores for object property prediction on simulated datasets

	Multi-dSprites			Tetrominoes-NM		
Method	Color	Position	Shape	Color	Position	Shape
Slot Attention	72.2±12	96.8±0.1	38.2±0.0	86.5±6.5	98.7±0.6	36.3±0.0
EfficientMORL	86.5±6.2	95.8±0.1	61.7±0.0	94.9±3.2	97.9±0.7	68.5±0.0
GENESIS-V2	78.1±7.5	97.1±0.7	75.8±0.0	88.1±5.8	94.6±2.6	37.9±0.0
SLATE	87.5±0.7	90.6±4.4	31.7±0.0	85.5±3.9	89.6±0.7	10.5±0.0
SysBinder	73.6±1.0	69.3±3.4	33.3±0.0	97.9±0.6	77.8±2.7	19.9±0.0
BO-QSA	96.3±1.6	97.4±0.1	75.2±0.0	98.1±0.7	98.9±0.2	52.5±0.0
OC-Net	**98.0±0.6**	**98.3±0.1**	**78.1±0.0**	**100.0±0.0**	**99.4±0.1**	**98.7±0.0**

OC-Net

- We have described a framework called OC-Net that learns object-centric representations in a fine-grained manner without supervision.
- From the results of experiments conducted on simulated, real-world, complex texture and common object images, we have demonstrated the superior quality of the object representations over current state-of-the-art. Moreover, we have highlighted the sample efficiency and generalizability of OC-Net.
- Finally, we have shown how the discovered object representations can be used to predict object properties in a downstream task, indicating its potential use for other computer vision applications where samples and ground truth labels are limited.



Thank you