UNSUPERVISED OBJECT-CENTRIC LEARNING WITH BI-LEVEL OPTIMIZED QUERY SLOT ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability to decompose complex natural scenes into meaningful object-centric abstractions lies at the core of human perception and reasoning. In the recent culmination of unsupervised object-centric learning, the Slot-Attention module has played an important role with its simple yet effective design and fostered many powerful variants. These methods, however, have been exceedingly difficult to train without supervision and are ambiguous in the notion of object, especially for complex natural scenes. In this paper, we propose to address these issues by (1) initializing Slot-Attention modules with learnable queries and (2) optimizing the model with bi-level optimization. With simple code adjustments on the vanilla Slot-Attention, our model, Bi-level Optimized Query Slot Attention, achieves state-of-the-art results on both synthetic and complex real-world datasets in unsupervised image segmentation and reconstruction, outperforming previous baselines by a large margin ($\sim$10%). We provide thorough ablative studies to validate the necessity and effectiveness of our design. Additionally, our model exhibits excellent potential for concept binding and zero-shot learning. We hope our effort could provide a single home for the design and learning of slot-based models and pave the way for more challenging tasks in object-centric learning.

1 INTRODUCTION

Objects, and their interactions, are the foundations of human cognition (Spelke & Kinzler, 2007). The endowment on making abstractions from perception and organizing them systematically empowers humans the ability to accomplish and generalize across a broad range of tasks, such as scene modeling (Bear et al., 2020), visual reasoning (Yi et al., 2020), and simulating interactions (Bear et al., 2020). The key to such success lies in the emergence of symbol-like mental representations of object concepts (Whitehead, 1928). However, important as it is, disentangling object-centric concepts from visual stimuli is an exceedingly difficult task to accomplish with limited supervision (Greff et al., 2020) and requires proper inductive biases (Schölkopf et al., 2021).

Motivated by the development of symbolic thought in human cognition, slot-based representations, instance (Greff et al., 2017; 2019; Locatello et al., 2020), sequential (Gregor et al., 2015; Burgess et al., 2019; Engelcke et al., 2021), or spatial (Crawford & Pineau, 2019; Lin et al., 2020; Jiang et al., 2019), have been the key inductive bias to recent advances in unsupervised object-centric learning. Among them, the Slot-Attention module has received tremendous focus given its simple yet effective design (Locatello et al., 2020). By leveraging the iterative attention mechanism, Slot-Attention learns to compete between slots for explaining parts of the input, exhibiting a soft-clustering effect on visual signals. It is later proven to be more memory and training efficient as a plug-and-play module for unsupervised object-centric learning Locatello et al. (2020) and fostered powerful variants in understanding images (Singh et al., 2021; Xu et al., 2022), 3D scenes (Yu et al., 2022; Sajjadi et al., 2022a) and videos (Kipf et al., 2022; Elsayed et al., 2022; Singh et al., 2022).

However, as revealed by recent studies, the Slot-Attention module comes with innate discrepancies for object-centric representation learning. First, with slots randomly initialized each time, the object-centric representations obtained by these models do not necessarily bind to object concepts (Kipf et al., 2022). Intuitively, such randomness leads to undesired scenarios where slots with similar initializations compete for objects on different images. Such randomness challenges the iterative refinement procedure as it now needs to project sets of potentially similar representations to indepen-

dent constituents of the input. As discovered by Chang et al. (2022), differentiating through such recurrences contributes to various training instabilities with growing spectral norm of Slot-Attention weights. This leads to the second and perhaps least desired property of Slot-Attention; it relies heavily on hyper-parameter tuning, including gradient clipping, learning rate warm-up, *etc*., and further hurts the flexibility of Slot-Attention in adapting to broader applications with more complex signals.

To this end, we propose an extension of the Slot-Attention module, Bi-level Optimized Query Slot Attention (BO-QSA), to tackle the aforementioned problems. First, instead of sampling from a learnable Gaussian distribution, we propose directly learning the slot initializations as queries.With these learnable representations, we eliminate the ambiguous competitions between slots and provide a better chance for them to bind to specific object concepts. More importantly, we ease the difficulty in training Slot-Attention with learnable queries by formulating Slot-Attention as a bi-level optimization problem. We improve the training of query-initialized Slot-Attention with a straight-through gradient estimator by connecting our method with first-order approaches (Finn et al., 2017; Nichol & Schulman, 2018; Geng et al., 2021) in solving bi-level optimization problems. The experimental results show that the proposed BO-QSA can achieve state-of-the-art results on both synthetic and real-world image datasets with simple code adjustments to the original Slot-Attention module.

With our model significantly outperforming previous methods in both synthetic and real domains, we provide thorough ablative studies demonstrating the effectiveness of our model design. We later show that our BO-QSA possesses the potential of binding object concepts to slots. To validate this potential, we design zero-shot transfer learning experiments to show the generalization power of our model on unsupervised object-centric learning. As the experiments suggest (see Sec. 5), our model could potentially be a principle approach for unsupervised object-centric learning and serve as a general plug-and-play module for a broader range of modalities where variants of Slot-Attention prosper. We hope these efforts can help foster new insights in the field of object-centric learning.

Contributions In summary, our main contribution of this work are three-fold:

- We propose BO-QSA that leverages learnable queries as initializations in Slot-Attention and improve its learning from a bi-level optimization perspective.
- We show that, with simple code adjustments on Slot-Attention, the proposed BO-QSA achieves state-of-the-art results on several challenging synthetic and real-world image benchmarks, outperforming previous methods by a large margin ($\sim$10% improvement).
- We show the potential of our BO-QSA being a better approach on concept binding and learning generalizable representations with qualitative results and zero-shot transfer learning experiments.

2 Preliminaries

2.1 Object-Centric Representation Learning with Slot-Attention

Slot-Attention (Locatello et al., 2020) takes a set of N input feature vectors $\mathbf{x} \in \mathbb{R}^{N \times D_{\text{input}}}$ and maps them to a set of K output vectors (*i.e*., slots) $\mathbf{s} \in \mathbb{R}^{K \times D_{\text{slots}}}$. It leverages a iterative attention mechanism to first map inputs and slots to the same dimension D with linear transformations $k(\cdot)$, $q(\cdot)$ and $v(\cdot)$ parameterized by ϕ^{attn}. At each iteration, the slots compete to explain part of the visual input by computing the attention matrix $\boldsymbol{A}$ with softmax function over slots and updating slots with the weighted average of visual values:

$$\tilde{\mathbf{s}} = f_{\phi^{\text{attn}}}(\mathbf{s}, \mathbf{x}) = \left(\frac{A_{i,j}}{\sum_{l=1}^{N} A_{l,j}} \right)^{\top} \cdot v(\mathbf{x}) \quad \text{where} \quad \boldsymbol{A} = \text{softmax}\left(\frac{k(\mathbf{x}) \cdot q(\mathbf{s})^{\top}}{\sqrt{D}} \right) \in \mathbb{R}^{N \times K}.$$

The slots are initialized from a learnable Gaussian distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$. They are refined iteratively within the Slot-Attention module by passing the updates into a Gated Recurrent Unit (GRU) (Cho et al., 2014) and MLP parameterized by ϕ^{update} for T iterations:

$$\mathbf{s}^{(t+1)} = h_{\phi^{\text{update}}}(\mathbf{s}^{(t)}, \tilde{\mathbf{s}}^{(t)}), \quad \mathbf{s}^{0} \sim \mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma})), \quad \hat{\mathbf{s}} = \mathbf{s}^{(T)}. \tag{1}$$

The final prediction $\hat{\mathbf{s}}$ can be treated as the learned object-centric representation w.r.t. to input features $\mathbf{x}$. In the image domain, we take as input a set of images $\boldsymbol{I}$ and encode them with $f_{\phi^{\text{enc}}}$ to obtain features $\mathbf{x} \in \mathbb{R}^{HW \times D_{\text{input}}}$. After obtaining $\hat{\mathbf{s}}$ through the iterative refinement procedure with $h_{\phi^{\text{update}}}$, images could be decoded from these object-centric representations with a mixture-based decoder

or autoregressive transformer-based decoder. We refer the readers to Appendix A.1 for details on different decoder designs and their ways of visualizing learned object concepts.

2.2 Bi-level Optimization and Meta-Learning

The problem of bi-level optimization embeds the optimization of an inner objective within the outer objective. Normally, a bi-level optimization problem can be formulated as:

$$\min_{\theta,\phi} f(\theta,\phi) \quad s.t. \quad \theta \in \arg\min_{\theta'} g(\theta',\phi), \tag{2}$$

where we call $f(\theta,\phi)$ the outer objective function and $g(\theta,\phi)$ the inner objective function. To jointly optimize both objectives w.r.t. parameters θ and ϕ, a straightforward approach to solving Eq. (2) is to represent the inner solution of θ as a function of ϕ, *i.e.*, $\theta^*(\phi) = \arg\min_{\theta'} g(\theta',\phi)$. Then we can optimize the outer objective with gradient descent by approximating $\nabla_\phi f(\theta^*(\phi),\phi)$ as a function of ϕ. Various methods have been proposed for estimating this gradient (Pedregosa, 2016; Lorraine et al., 2020). Most intuitively, we can use the first-order approximation for updating ϕ:

$$\phi_{k+1} = \phi_k - \eta \cdot \nabla_\phi f(\theta^*, \phi_k),$$

where θ^* is treated as an optimal solution to the inner objective with no gradient and k denotes iteration. We provide discussions on more methods that approximating $\nabla_\phi f(\theta^*(\phi),\phi)$ in Appendix A.2.

In fact, bi-level optimization is closely related to meta-learning. In meta-learning, we have meta-training tasks which comes in as N different collections of datasets $\mathcal{D} = \{\mathcal{D}_i = \mathcal{D}_i^{\text{tr}} \cup \mathcal{D}_i^{\text{val}}\}_{i=1}^N$. The inner and outer objectives in Eq. (6) are substituted by averaging training and validation errors over multiple tasks (Franceschi et al., 2018):

$$\min_{\theta,\phi} f(\theta,\phi) = \sum_{i=1}^{N} \mathcal{L}_i(\theta_i,\phi,\mathcal{D}_i^{\text{val}}) \quad s.t. \quad \theta_i = \min_{\theta_i'} \sum_{i=1}^{N} \mathcal{L}_i(\theta_i',\phi;\mathcal{D}_i^{\text{tr}}), \tag{3}$$

where $\mathcal{L}_i$ represents task-dependent error on $\mathcal{D}_i$. The final goal of meta-learning aims at seeking the meta-parameter ϕ that is shared between tasks which later enables few-shot learning and fast adaptation. With its connections with bi-level optimization, the previously mentioned optimization methods are broadly adapted for solving meta-learning problems (Finn et al., 2017; Nichol & Schulman, 2018; Rajeswaran et al., 2019).

3 Bi-level Optimized Query Slot Attention

3.1 Query Slot Attention

As mentioned in Sec. 1, the Slot-Attention module adopts a random initialization of slots and conduct iterative refinement to obtain object-centric representations $\hat{\mathbf{s}}$ as in Eq. (1). However, as argued by Kipf et al. (2022), such initializations provide no hint on the desired level of granularity, making the notion of an object ambiguous. As shown by Chang et al. (2022), this random initialization plays minimal role and could be detached from training. This indicates that the estimation of $\hat{\mathbf{s}}$ relies heavily on the task-specific iterative refining of slots over data, leaving a limited possibility for slots to bind to specific concepts and be leveraged as generalizable representations. Such a phenomenon also resembles the posterior collapse problem (Alemi et al., 2018) discovered in generative latent variable models, where the sampled latents are noisy and force decoders to ignore them.

To address this issue, we propose the Query Slot Attention (QSA) which initializes the slots in the Slot-Attention module with learnable queries ϕ^{init} as:

$$\mathbf{s}_0 = \phi^{\text{init}}.$$

Such design is motivated by how discrete VAEs leverage learnable tokens as latents for avoiding the posterior collapse (Van Den Oord et al., 2017). It facilitates an object-centric model to learn general symbolic-like representations that could be quickly adapted by refining over task-specific requirements, as discussed in Sec. 1 and Kipf et al. (2022). Similar to Jaegle et al. (2021b), the cross-attention with learnable queries in QSA can be interpreted from the data compression perspective where image features are iteratively attended to the learnable queries, forming generalizable representatives. As shown in Tab. 5, using learnable queries as slot initializations can significantly improve object-centric learning with proper optimization designs. We later show in Tab. 6 and Fig. 3 that QSA exhibits the potential for binding specific concepts to slots and generalization.

3.2 OBJECT-CENTRIC LEARNING WITH SLOT-ATTENTION AS BI-LEVEL OPTIMIZATION

Despite the good properties and potentials QSA presents, it is shown detrimental to initialize slots independently in Slot-Attention under unsupervised settings (Locatello et al., 2020). Therefore, to improve the learning of QSA model under unsupervised settings, we rewind to the optimization perspective of Slot-Attention to find better training methods.

In Slot-Attention, slots are iteratively updated with parameters $\Phi = \{\phi^{\text{init}}, \phi^{\text{attn}}, \phi^{\text{update}}\}$ for assigning input features $\mathbf{x}$ to different slots. This iterative process could be equivalently formulated as solving for the fixed points (Chang et al., 2022) of

$$\mathbf{s} = h_{\phi^{\text{update}}}(\mathbf{s}, \tilde{\mathbf{s}}) = h_{\phi^{\text{update}}}(\mathbf{s}, f_{\phi^{\text{attn}}}(\mathbf{s}, \mathbf{x})) = F_{\Phi}(\mathbf{s}, \mathbf{x}), \tag{4}$$

where $F_{\Phi}(\cdot, \cdot)$ is a certain fixed point operation. This procedure could be viewed as finding optimal solutions to an objective. For example, in K-means, a similar iterative procedure solves for cluster centroids $\mathbf{c}$ that minimize the objective $\mathcal{L}_{\text{cluster}}^{\text{K-means}} = \sum_j \sum_i ||x_i^j - c_j||_2^2$, where x_i^j denotes data samples that lies in cluster c_j. As pointed out by Greff et al. (2017); Locatello et al. (2020), Slot-Attention can also be viewed as a soft version of k-means clustering. Then, similar to K-means, we assume that the fixed point iteration in Eq. (4) accounts for the optimization of a clustering objective $\mathcal{L}_{\text{cluster}}$. Viewing from the bi-level optimization perspective, this $\mathcal{L}_{\text{cluster}}$ serves as the optimization of an inner objective, connecting the overall formulation of Slot-Attention with the bi-level optimization problem described in Sec. 2.2 by:

$$\min_{\mathbf{s}, \Phi} \sum_{i=1}^{M} \mathcal{L}(\mathbf{x}_i, \mathbf{s}_i, \Phi) \quad s.t. \quad \mathbf{s}_i^* = \arg\min_{\mathbf{s}_i} \mathcal{L}_{\text{cluster}}(\mathbf{x}_i, \mathbf{s}_i, \Phi), \tag{5}$$

where $\mathbf{x}_i$ and $\mathbf{s}_i$ denote the input feature from the i-th image and its corresponding slot features. Under this setting, the outer objective $\mathcal{L}$ could be for reconstruction or set prediction (Locatello et al., 2020) depending on the task.

3.3 BI-LEVEL OPTIMIZED QUERY SLOT ATTENTION

To solve the optimization problem in Eq. (5), Slot-Attention backpropagates gradients into the recursive updates, which leads to various training instabilities. As introduced by Chang et al. (2022) in Implicit Slot-Attention (ISA), such instabilities could be avoided by detaching gradients to the recursive updates, treating slots in the final iteration as an approximation of s_i^*, and computing first-order gradient approximations for updating Φ with $\mathbf{s}_i^*$. Such a design shows potential in stabilizing Slot-Attention training but lies heavily on the good approximation of the solution to the inner objective. Meanwhile, they detached the gradient to slot initializations, starting from random initializations each time for the inner optimization. Such a design puts challenges on the learning of Slot-Attention update F_{Φ} as it will be difficult to provide a good approximation of s_i^* starting from any random initialization with only a fixed number of iterations.

We propose BO-QSA to address the learning problem of QSA with first-order approximations in bi-level optimization. As shown in Algorithm 1, we initialize slots with learnable queries in BO-QSA. We then perform T steps of Slot-Attention update to obtain an approximation of $\mathbf{s}_i^*$. These near-optimal solutions of the inner objective are passed into one additional Slot-Attention step where gradients to all previous iterations are detached. In contrary to ISA, we use a straight-through gradient estimator (Van Den Oord et al., 2017) to backpropagate gradients to slot initialization queries as well. We believe such designs help find good starting points for the inner optimization problem on clustering. From the meta-learning perspective, such an attempt shares similar insights with first-order meta-learning methods (Finn et al., 2017; Nichol & Schulman, 2018), where we use

Algorithm 1: BO-QSA

Input: input features `input`, learnable queries `init`, number of iterations T
Output: object-centric representation `slots`
Modules : stop gradient module SG($\cdot$), slot attention module SA($\cdot, \cdot$)

```
slots = init
for t = 1, ··· , T do
  | slots = SA(slots, inputs)
slots = SG(slots) + init - SG(init)
slots = SA(slots, inputs)
return slots
```

the gradient at some task-specific optimal solution $\mathbf{s}_i^*$ of the inner optimization for optimizing slot initialization queries which are shared across datasets on the outer objective. This meta-learning perspective also indicates the potentials of our BO-QSA for fast adaptation and generalization. In our experiments, we use image reconstruction as the outer objective and show that the BO-QSA design not only enhances QSA significantly in unsupervised object-centric learning but also shows the potential for concept binding and zero-shot transfer. We leave this discussion to Sec. 5.

4 Related Work

Unsupervised Object-Centric Learning Our work falls into the recent line of research on unsupervised object-centric learning on images (Greff et al., 2016; Eslami et al., 2016; Greff et al., 2017; 2019; Burgess et al., 2019; Crawford & Pineau, 2019; Engelcke et al., 2020; Lin et al., 2020; Bear et al., 2020; Locatello et al., 2020; Zoran et al., 2021). A thorough review and discussion on this type of method could be found in Greff et al. (2020). One critical issue of these methods is on handling complex natural scenes. Singh et al. (2021); Lamb et al. (2021) leverages a transformer-based decoder with Slot-Attention for addressing this problem. Similar attempts have also been made by exploiting self-supervised contrastive learning (Choudhury et al., 2021; Caron et al., 2021; Wang et al., 2022; Hénaff et al., 2022) and energy-based models (Du et al., 2021; Yu et al., 2022). Our work builds upon Slot-Attention by extending it with learnable queries and a novel optimization method for learning. Our compelling experimental suggests our model could potentially serve as a general plug-and-play module for a wider range of modalities where variants of Slot-Attention prosper (Kipf et al., 2022; Elsayed et al., 2022; Singh et al., 2022; Yu et al., 2022; Sajjadi et al., 2022a;b).

Query Networks Sets of latent queries are commonly used in neural networks. These methods leverage permutation equivariant network modules (*e.g.* GNNs (Scarselli et al., 2008) and attention modules (Vaswani et al., 2017)) in model design for solving set-related tasks such as clustering (Lee et al., 2019), outlier detection (Zaheer et al., 2017; Zhang et al., 2019), *etc*. These learned latent queries have been shown to have good potential as features for tasks like contrastive learning (Caron et al., 2020), object detection (Carion et al., 2020) and data compression (Jaegle et al., 2021a;b). In contrast to recent success of query networks in supervised or weakly-supervised learning (Carion et al., 2020; Zhang et al., 2021; Kipf et al., 2022; Elsayed et al., 2022; Xu et al., 2022), Locatello et al. (2020) demonstrates the detrimental effect of using independently initialized slots in Slot-Attention learning. However, we show that our BO-QSA method successfully overcomes this issue and generalizes the success of query networks to the domain of unsupervised object-centric learning.

Bi-level Optimization and Meta-Learning Our works is closely related to bi-level optimization methods with iterative fixed update rule for solving the inner objective. Specifically, methods are designed with implicit differentiation (Amos & Kolter, 2017; Bai et al., 2019) to stabilize the iterative update procedure. Similar formulations are also found in meta-learning literature where FOMAML (Finn et al., 2017) and Reptile (Nichol & Schulman, 2018) provide first-order approximation of the bi-level optimization problem in meta-learning, and Rajeswaran et al. (2019) provides a unified view on the optimization problem with implicit gradients. Concurrent work from Chang et al. (2022) formulate the Slot-Attention learning from an implicit gradient perspective with gradient stopping derived from first-order hyper-gradient methods (Geng et al., 2021). However, they start from the pure optimization perspective, ignoring the important role of slot initializations in generalization and concept binding. As our experiments suggest, their gradient stopping methods do not necessarily guarantee superior performance compared to the original Slot-Attention. We leave the details to Sec. 5.3 for a more in-depth discussion.

5 Experiments

In this section, we evaluate the proposed BO-QSA on both synthetic and real-world image datasets. Specifically, we aim to address the following questions with our experimental results:

- How good is our proposed BO-QSA on both synthetic and complex natural scenes?
- How important is the query and the optimization method in BO-QSA?
- Does BO-QSA possess the potential for concept binding and zero-shot transfer?

Table 1: Multi-object segmentation results on ShapeStacks and ObjectsRoom. We report ARI-FG and MSC-FG of all models with (mean $\pm$ variance) across 3 experiment trials. We visualize the best results in bold.

Model	ShapeStacks		ObjectsRoom	
	$\uparrow$ ARI-FG	$\uparrow$ MSC-FG	$\uparrow$ ARI-FG	$\uparrow$ MSC-FG
MONet-G (Burgess et al., 2019)	0.70$\pm$0.04	0.57$\pm$0.12	0.54$\pm$0.00	0.33$\pm$0.01
GENESIS (Engelcke et al., 2020)	0.70$\pm$0.05	0.67$\pm$0.02	0.63$\pm$0.03	0.53$\pm$0.07
Slot-Attention (Locatello et al., 2020)	0.76$\pm$0.01	0.70$\pm$0.05	0.79$\pm$0.02	0.64$\pm$0.13
GENSIS-V2 (Engelcke et al., 2021)	0.81$\pm$0.01	0.67$\pm$0.01	0.86$\pm$0.01	0.59$\pm$0.01
SLATE (Singh et al., 2021)	0.65$\pm$0.03	0.63$\pm$0.05	0.57$\pm$0.03	0.30$\pm$0.03
ours (BO-QSA+transformer)	0.68$\pm$0.02	0.70$\pm$0.02	0.68$\pm$0.03	0.72$\pm$0.03
ours (BO-QSA+mixture)	**0.93$\pm$0.01**	**0.89$\pm$0.00**	**0.87$\pm$0.03**	**0.80$\pm$0.02**

We provide details in the following sections with thorough comparative and ablative experiments. We clarify the datasets and metrics selected for evaluating our model on each domain:

Synthetic Domain For the synthetic domain, we select two well-established challenging multi-object datasets Shapestacks (Groth et al., 2018) and ObjectsRoom (Kabra et al., 2019) for evaluating our BO-QSA model. Specifically, we consider three metrics to evaluate the quality of object segmentation and reconstruction. Adjusted Rand Index (ARI) (Hubert & Arabie, 1985) and Mean Segmentation Covering (MSC) (Engelcke et al., 2020) for segmentation and Mean Squared Error (MSE) for reconstruction. We report the first two segmentation metrics over foreground objects (ARI-FG and MSC-FG) following the default evaluation setting in recent works. In addition to the two datasets selected, we also conduct additional experiments on CLEVRTEX (Karazija et al., 2021) and PTR (Hong et al., 2021). We leave the discussion on their results in Appendix B.2.

Real-world Images For the real image domain, we use unsupervised foreground extraction as a representative task for evaluating our method. Specifically, we select Stanford Dogs (Khosla et al., 2011), Stanford Cars (Krause et al., 2013), CUB200 Birds (Welinder et al., 2010), and Flowers (Nilsback & Zisserman, 2010) as our benchmarking datasets. We select these datasets instead of other real-world image datasets used in previous object-centric learning literature (*e.g.* APC (Zeng et al., 2017), Sketchy (Cabi et al., 2020)) as they are mainly designed for robotics research under constrained environments and thus less challenging. Our selected benchmarks contain richer visual concepts with more variations (*e.g.* texture, lighting) and, consequently, are more suitable for evaluating object-centric models. We use mean Intersection over Union (mIoU) and Dice as metrics for evaluating the quality of foreground extraction.

5.1 Object Discovery on Synthetic Datasets

Experimental Setup We explore our proposed BO-QSA with two types of decoder designs, mixture-based and transformer-based, as discussed in Sec. 2.1 and Appendix A.1. We follow the decoder architecture in Slot-Attention (Locatello et al., 2020) for mixture-based decoders and SLATE (Singh et al., 2021) for transformer-based decoders. For both types of models, we use the Slot-Attention module with a CNN image encoder and initialize slots with learnable embeddings. We leave the model implementation details to Appendix A.3.

Results We report multi-object segmentation results on the synthetic domain in Tab. 1 and visualize qualitative results in Fig. 1. As shown in Tab. 1, our BO-QSA achieves the state-of-the-art results with large improvements over previous object-centric learning methods on all metrics in ShapeStacks and ObjectsRoom. We also observe more stable model performance, *i.e.* smaller variances in results, across different trials of experiments. Our model with mixture-based decoder obtains the best overall performance on all datasets. More specifically, our mixture-based BO-QSA significantly outperforms the vanilla Slot-Attention model ($\sim$15% on average) with minimal architectural differences. This validates the importance and effectiveness of the learnable queries and our optimization method. We will continue this discussion in Sec. 5.3.

Mixture-based vs. Transformer-based Decoder Interestingly, we observe inferior segmentation performance but superior reconstruction performance of transformer-based variants of Slot-Attention (*i.e.*, SLATE) on synthetic datasets. We follow experimental settings in Singh et al. (2021) and report

Table 3: Unsupervised foreground extraction results on CUB200 Birds (Birds), Stanford Dogs (Dogs), Stanford Cars (Cars), and Caltech Flowers (Flowers). We visualize the best results in bold.

Model	Birds		Dogs		Cars		Flowers	
	↑ IoU	↑ Dice	↑ IoU	↑ Dice	↑ IoU	↑ Dice	↑ IoU	↑ Dice
ReDO (Chen et al., 2019)	46.5	60.2	55.7	70.3	52.5	68.6	76.4	-
IODINE (Greff et al., 2019)	30.9	44.6	54.4	67.0	51.7	67.3	-	-
OneGAN (Benny & Wolf, 2020)	55.5	69.2	71.0	81.7	71.2	82.6	-	-
Slot-Attention	35.6	51.5	39.6	55.3	41.3	58.3	30.8	45.9
Voynov et al. (2020)	68.3	-	-	-	-	-	54.0	-
DRC (Yu et al., 2021)	56.4	70.9	71.7	83.2	72.4	83.7	-	-
Melas-Kyriazi et al. (2021)	66.4	-	-	-	-	-	54.1	-
SLATE	36.1	51.0	62.3	76.3	75.5	85.9	68.1	79.1
ours (BO-QSA+mixture)	25.1	39.2	36.8	53.6	69.1	81.5	36.1	51.6
ours (BO-QSA+transformer)	**71.0**	**82.6**	**82.5**	**90.3**	**87.5**	**93.2**	**78.4**	**86.1**

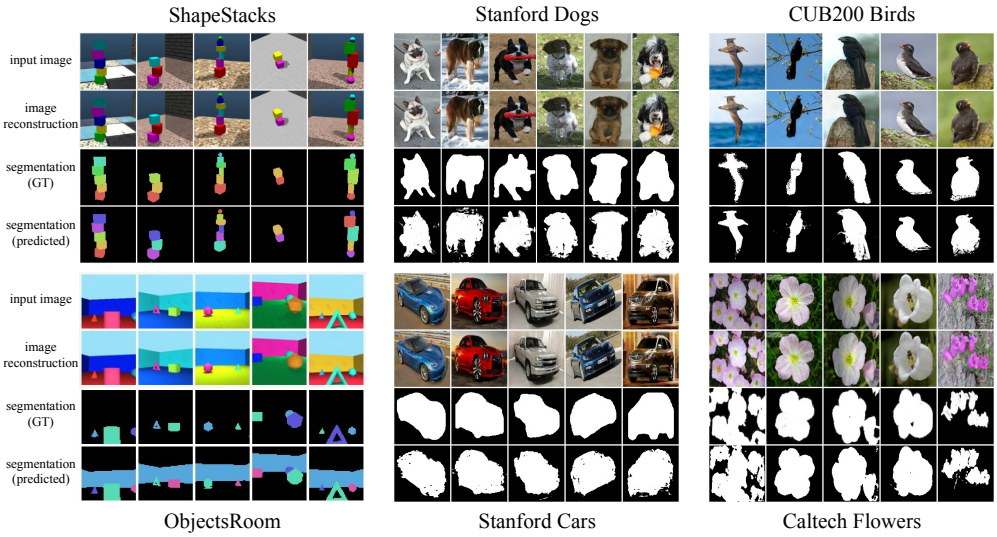

Figure 1: Visualization of our predicted segmentation and reconstruction results on synthetic and real-world images. We color the predicted mask that has maximum intersection with the ground-truth background in black.

the MSE of models with the best random seed on each dataset. As shown in Tab. 2, transformer-based methods provide better reconstruction results. We attribute their low segmentation performance to mask prediction design. Without a separate mask prediction head, such methods rely on the attention matrix computed over input features for segmentation. With image encoders aggregating local features, the attention scores computed at object boundaries will be similar as they share similar information. This leads to coarse object masks, hindering the segmentation performance, especially on synthetic datasets. Nevertheless, we observe consistent improvement by applying our slot encoder to both mixture-based and transformer-based decoders.

Table 2: Reconstruction results on ShapeStacks and ObjectsRoom (MSE↓). We compare mixture-based and transformer-based decoder designs.

Model	ShapeStacks	ObjectsRoom
Slot-Attention (mixture)	80.8	20.4
ours (mixture)	**72.0**	**8.1**
SLATE (transformer)	52.3	16.3
ours (tranformer)	**49.3**	**14.7**

5.2 Unsupervised Foreground Extraction on Real Datasets

Experimental Setup For real-world experiments, we use the same slot encoder design used in Sec. 5.1 with a 4-layer CNN image encoder and initialize slots with learnable queries. Following Yu et al. (2021), we report the best model performance on all datasets. During the eval-

Table 5: Ablative experiments on different slot initialization and optimization methods. We visualize the best results in bold and underline the second-best results. (*Note that SA represents Slot-Attention with our slot-encoder design and different from the original one reported in Tab. 3.)

Method	Dogs		ShapeStacks	
	↑ IoU	↑ Dice	↑ ARI-FG(%)	↑ MSC-FG(%)
SA*	71.0	81.9	86.7	<u>84.8</u>
I-SA	80.8	89.2	<u>88.3</u>	76.8
BO-SA	<u>80.9</u>	<u>89.3</u>	87.7	66.6
QSA	64.5	72.9	88.1	76.1
I-QSA	59.3	77.6	84.6	81.8
BO-QSA (ours)	**82.5**	**90.3**	**92.9**	**89.2**

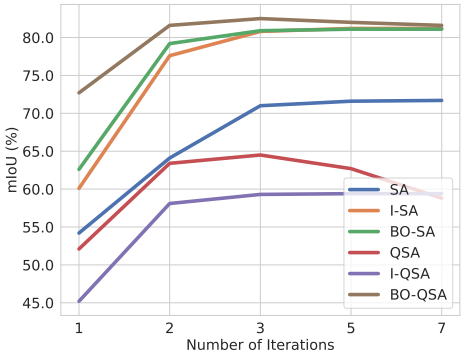

Figure 2: Effects of iterative updates in testing.

uation, we select the slot's mask prediction that has a maximum intersection with the ground-truth foreground mask as our predicted foreground. We leave model implementation and training details to Appendix A.3.

Results We show quantitative experimental results in Tab. 3 and visualize qualitative examples in Fig. 1. As shown in Tab. 3, our method significantly outperforms all existing baselines on the task of foreground extraction, achieving new state-of-the-art on all datasets. We recognize the discrepancy of mixture-based decoders in both Slot-Attention and our mixture-based design in modeling real-world images, reflecting similar discoveries from recent works (Singh et al., 2021) that mixture-based decoder struggles in modeling real-world images. On the other hand, our transformer-based model shows significant improvement over the vanilla version. Notably, our method outperforms a broad range of models, including GAN-based generative models (*i.e.* OneGAN, Voynov et al. (2020)), and large-scale pre-trained contrastive methods (*i.e.* MoCo-v2, BYOL, R2O). As shown in Tab. 4, our method achieves comparable results with state-of-the-art self-supervised contrastive learning methods without large-scale pre-training and data augmentation. This result sheds light on the potential of object-centric learning as a pre-training task for learning general visual representations.

Table 4: Unsupervised segmentation results on CUB200 Birds (mIoU↓). *Contrastive learning methods are pre-trained on ImageNet and generate object predictions with K-means feature clustering.

Model	Birds
MoCo v2 (Chen et al., 2020)	63.5
BYOL (Grill et al., 2020)	56.1
R2O (Gokul et al., 2022)	**71.2**
ours (BO-QSA+transformer)	**71.0**

5.3 Ablative Studies

Experimental Setup We perform ablative studies over our designs by comparing them with different model variants. For slot initialization, we consider two types of initialization methods: (1) sampling from a learnable Gaussian distribution as proposed in the original Slot-Attention module (SA), and (2) our Query Slot-Attention design (QSA). For optimization, we consider three types of methods: (1) the original optimization method in Slot-Attention, (2) the ISA method introduced by Chang et al. (2022) where gradients are not backpropagated through the iterative refinement but to parameters for attention ϕ^{attn} and update ϕ^{update}, and (3) our method where gradients are also not backpropagated through the iterative refinement, but to both parameters for attention ϕ^{attn}, update ϕ^{update} and initialization queries ϕ^{init} with a straight-through estimator. For simplicity, we term these optimization methods with prefixes (I-) for ISA and (BO-) for our method. We run all ablations over our 4-layer CNN slot-encoder design. We choose ShapeStacks and Stanford Dogs as representative datasets for ablative studies and provide training details in Appendix A.3.

Results We show experimental results in Tab. 5 and Fig. 2. First, from Tab. 5, we observe that BO-QSA significantly outperforms other variants. Next, for sample-based slot initializations, our method shows a similar effect on improving Slot-Attention learning. For query-based slot initializations, we validate the difficulty in training query-based Slot-Attention by showing its inferior performance compared with sample-based Slot-Attention. We further show the ineffectiveness of ISA for query-based Slot-Attention as it detaches gradients to slot initializations. The experiments on query-based Slot-Attention prove that both of our design choices are necessary and effective for

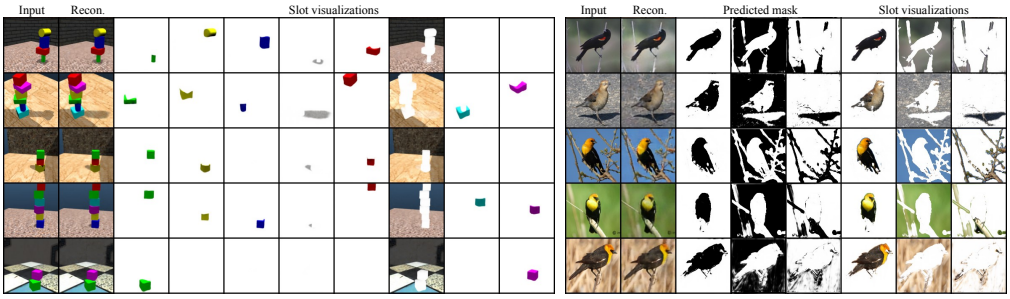

Figure 3: Visualization of learned contents for each slot. On ShapeStacks, we observe strong correlation between slots and color concepts (Left). On CUB200 Birds, the first two slots attribute for most of the foreground and background contents, while the third slot covers additional textures like branches and shadows (Right).

the superior performance of BO-QSA. To further study the effect of learned queries, we visualize in Fig. 2 where we set different numbers of iterative updates of Slot-Attention during inference on the Stanford Dogs dataset. We can see that our proposed BO-QSA significantly outperforms other variants with only one iteration step. This indicates that our query-based design can help ease the training difficulty in iterative refinement and shows the potential of these learned queries as generalizable representations.

5.4 Additional Analyses

In this section, we provide additional analyses on the potential of our BO-QSA as a concept binder for generalizing to new examples. First, we qualitatively visualize our learned content for each slot (without additional clustering) in ShapeStacks and Birds in Fig. 3. We observe high similarity within the learned content of each slot, indicating the similar concepts learned by specific slots. This shows the potential of the slots in our BO-QSA for binding specific concepts (*e.g.* color in this case). To verify this argument, we further perform a simple transfer experiment where we use models trained on dataset X for zero-shot inference on dataset Y. We term this transfer as (X→Y). As shown in Tab. 6, when adapting models trained on Stanford Dogs to zero-shot inference on Stanford Cars and Caltech Flowers, our method outperform ISA and also a majority of fine-tuned methods shown in Tab. 3. We later show in Appendix B.1 that this transfer capability is not limited to specific datasets. This further proves the potential of our learned object-centric representations for generalization purposes.

Table 6: Zero-shot transfer results (mIoU ↑).

Model	Dogs → Cars	Dogs → Flowers
SA	58.79	57.96
I-SA	69.90	58.05
BO-QSA(ours)	**75.5**	**63.4**

6 Conclusions

In this work, we introduce BO-QSA for unsupervised object-centric representation learning. We initialize Slot-Attention with learnable queries and further introduce a novel bi-level optimization method to ease the difficulty in query-based Slot-Attention learning. With simple code adjustments, we transform the Slot-Attention module into a powerful object-centric encoder that achieves state-of-the-art results on unsupervised object segmentation in both synthetic and natural image domains, outperforming previous baselines by a large margin. More importantly, our learned model exhibits concept binding effects where visual concepts are attached to specific slot queries. By further verifying our findings under zero-shot transfer learning settings, our model shows great potential as a principle approach for learning generalizable object-centric representations.

As discussed in Sec. 5.2, a promising future direction is to connect our method with self-supervised contrastive learning methods for learning general visual representations. Another direction is to study the finer level organization of object concepts under more complex scenarios (*e.g.* hierarchical grouping, language grounding, *etc.*) with weak supervision. We hope our efforts could pave the way to more exciting and challenging applications of object-centric representations.

REFERENCES

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *Proceedings of International Conference on Machine Learning (ICML)*, 2018. 3

Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 136–145, 2017. 5

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. Deep equilibrium models. *Advances in Neural Information Processing Systems*, 32, 2019. 5

Daniel Bear, Chaofei Fan, Damian Mrowca, Yunzhu Li, Seth Alter, Aran Nayebi, Jeremy Schwartz, Li F Fei-Fei, Jiajun Wu, Josh Tenenbaum, et al. Learning physical graph representations from visual scenes. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 5

Yaniv Benny and Lior Wolf. Onegan: Simultaneous unsupervised learning of conditional image generation, foreground segmentation, and fine-grained clustering. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 7

Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019. 1, 5, 6, 18

Serkan Cabi, Sergio Gómez Colmenarejo, Alexander Novikov, Ksenia Konyushkova, Scott Reed, Rae Jeong, Konrad Zolna, Yusuf Aytar, David Budden, Mel Vecerik, et al. Scaling data-driven robotics with reward sketching and batch reinforcement learning. *Proceedings of Robotics: Science and Systems (RSS)*, 2020. 6

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020. 5

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 5

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021. 5

Michael Chang, Thomas L Griffiths, and Sergey Levine. Object representations as fixed points: Training iterative refinement algorithms with implicit differentiation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3, 4, 5, 8

Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 7

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 8

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 2

Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5

Eric Crawford and Joelle Pineau. Spatially invariant unsupervised object detection with convolutional neural networks. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 1, 5

Yilun Du, Shuang Li, Yash Sharma, Josh Tenenbaum, and Igor Mordatch. Unsupervised learning of compositional energy concepts. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 5

Gamaleldin F Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael C Mozer, and Thomas Kipf. Savi++: Towards end-to-end object-centric learning from real-world videos. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 5

Martin Engelcke, Adam R Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020. 5, 6

Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. Genesis-v2: Inferring unordered object representations without iterative refinement. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 6, 18

SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 5

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of International Conference on Machine Learning (ICML)*, 2017. 2, 3, 4, 5

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proceedings of International Conference on Machine Learning (ICML)*, 2018. 3

Zhengyang Geng, Xin-Yu Zhang, Shaojie Bai, Yisen Wang, and Zhouchen Lin. On training implicit models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 5

Akash Gokul, Konstantinos Kallidromitis, Shufan Li, Yusuke Kato, Kazuki Kozuka, Trevor Darrell, and Colorado J Reed. Refine and represent: Region-to-object representation learning. *arXiv preprint arXiv:2208.11821*, 2022. 8

Klaus Greff, Antti Rasmus, Mathias Berglund, Tele Hao, Harri Valpola, and Jürgen Schmidhuber. Tagger: Deep unsupervised perceptual grouping. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 5

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1, 4, 5

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *Proceedings of International Conference on Machine Learning (ICML)*, 2019. 1, 5, 7

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020. 1, 5

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *Proceedings of International Conference on Machine Learning (ICML)*, 2015. 1

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 8

Oliver Groth, Fabian B Fuchs, Ingmar Posner, and Andrea Vedaldi. Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018. 6

Olivier J Hénaff, Skanda Koppula, Evan Shelhamer, Daniel Zoran, Andrew Jaegle, Andrew Zisserman, João Carreira, and Relja Arandjelović. Object discovery and representation networks. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 5

Yining Hong, Li Yi, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Ptr: A benchmark for part-based conceptual, relational, and physical reasoning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 6, 18

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985. 6

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021a. 5

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of International Conference on Machine Learning (ICML)*, 2021b. 3, 5

Jindong Jiang, Sepehr Janghorbani, Gerard De Melo, and Sungjin Ahn. Scalor: Generative world models with scalable object representations. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019. 1

Rishabh Kabra, Chris Burgess, Loic Matthey, Raphael Lopez Kaufman, Klaus Greff, Malcolm Reynolds, and Alexander Lerchner. Multi-object datasets. https://github.com/deepmind/multi-object-datasets/, 2019. 6

Laurynas Karazija, Iro Laina, and Christian Rupprecht. Clevrtex: A texture-rich benchmark for unsupervised multi-object segmentation. In *Proceedings of Advances in Neural Information Processing Systems Datasets and Benchmarks (NeurIPS Datasets and Benchmarks Track)*, 2021. 6, 18

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, 2011. 6

Thomas Kipf, Gamaleldin F Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2022. 1, 3, 5

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of International Conference on Computer Vision Workshops (ICCVW)*, 2013. 6

Alex Lamb, Di He, Anirudh Goyal, Guolin Ke, Chien-Feng Liao, Mirco Ravanelli, and Yoshua Bengio. Transformers with competitive ensembles of independent mechanisms. *arXiv preprint arXiv:2103.00336*, 2021. 5

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*, 2019. 5

Zhixuan Lin, Yi-Fu Wu, Skand Vishwanath Peri, Weihao Sun, Gautam Singh, Fei Deng, Jindong Jiang, and Sungjin Ahn. Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020. 1, 5

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 4, 5, 6, 16, 18

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020. 3, 15

Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models. *arXiv preprint arXiv:2105.08127*, 2021. 7

Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2018. 2, 3, 4, 5

Maria-Elena Nilsback and Andrew Zisserman. Delving deeper into the whorl of flower segmentation. *Image and Vision Computing*, 28(6):1049–1062, 2010. 6

Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *Proceedings of International Conference on Machine Learning (ICML)*, 2016. 3, 15

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3, 5

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of International Conference on Machine Learning (ICML)*, 2021. 15

Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd van Steenkiste, Filip Pavetić, Mario Lučić, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022a. 1, 5

Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b. 5

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008. 5

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. In *Proceedings of the IEEE*, 2021. 1

Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate dall-e learns to compose. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021. 1, 5, 6, 8, 15, 16, 17

Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 5, 15

Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007. 1

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3, 4, 15

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5

Andrey Voynov, Stanislav Morozov, and Artem Babenko. Big gans are watching you: Towards unsupervised object segmentation with off-the-shelf generative models. 2020. 7, 8

Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

Nicholas Watters, Loic Matthey, Christopher P Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for learning disentangled representations in vaes. *arXiv preprint arXiv:1901.07017*, 2019. 15

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010. 6

Alfred North Whitehead. Symbolism: Its meaning and effect. *Journal of Philosophical Studies*, 3 (12), 1928. 1

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 5

Mao Ye, Bo Liu, Stephen Wright, Peter Stone, and Qiang Liu. Bome! bilevel optimization made easy: A simple first-order approach. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 15

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020. 1

Hong-Xing Yu, Leonidas J Guibas, and Jiajun Wu. Unsupervised discovery of object radiance fields. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2022. 1, 5

Peiyu Yu, Sirui Xie, Xiaojian Ma, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Unsupervised foreground extraction via deep region competition. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 7

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 5

Andy Zeng, Kuan-Ting Yu, Shuran Song, Daniel Suo, Ed Walker, Alberto Rodriguez, and Jianxiong Xiao. Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge. In *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2017. 6

Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

Yan Zhang, Jonathon Hare, and Adam Prugel-Bennett. Deep set prediction networks. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 5

Daniel Zoran, Rishabh Kabra, Alexander Lerchner, and Danilo J Rezende. Parts: Unsupervised segmentation with slots, attention and independence maximization. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021. 5

A Model Architecture and Design

A.1 Design of Decoders

In this section, we follow the notations used in Sec. 2.1 and describe two common approaches, mixture-based and transformer-based, for decoding images from the learned slot representations.

Mixture-based Decoder The mixture-based decoder (Watters et al., 2019) decodes each slot $\hat{\mathbf{s}}_i$ into an object image $\mathbf{x}_i$ and mask $\mathbf{m}_i$ with decoding functions $g^{\text{img}}_{\phi^{\text{dec}}}$ and $g^{\text{mask}}_{\phi^{\text{dec}}}$, which are implemented using CNNs. The decoded images and masks are calculated by:

$$\hat{\boldsymbol{I}}_i = g^{\text{img}}_{\phi^{\text{dec}}}(\hat{\mathbf{s}}_i), \quad \mathbf{m}_i = \frac{\exp\ g^{\text{mask}}_{\phi^{\text{dec}}}(\hat{\mathbf{s}}_i)}{\sum_{j=1}^{K} \exp\ g^{\text{mask}}_{\phi^{\text{dec}}}(\hat{\mathbf{s}}_j)}, \quad \hat{\boldsymbol{I}} = \sum_{i=1}^{K} \mathbf{m}_i \cdot \hat{\boldsymbol{I}}_i.$$

During training, a reconstruction objective is employed for supervising model learning. Despite its wide usage, mixture-based decoders showed limited capability at handling natural scenes with high visual complexity (Singh et al., 2021).

AutoregresObjectsRoomsive Transformer Decoder Recently, Singh et al. (2021; 2022) reveal the limitations of mixture decoder and leverage transformers and discrete VAEs (Van Den Oord et al., 2017; Ramesh et al., 2021) for decoding slot-based object-centric representations. To obtain decoded images $\hat{\boldsymbol{I}}$, they learn a separate discrete VAE for firsting encoding $\boldsymbol{I}$ into a sequence of L tokens $\mathbf{z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_L\}$ with dVAE encoder $f^{\text{dVAE}}_{\phi^{\text{enc}}}$. Next, they use a transformer decoder $g^{\text{transformer}}_{\phi^{\text{dec}}}$ to auto-regressively predict image tokens with learned slot representation $\hat{\mathbf{s}}$:

$$\mathbf{o}_l = g^{\text{transformer}}_{\phi^{\text{dec}}}(\hat{\mathbf{s}}; \mathbf{z}_{<l}) \quad \text{where} \quad \mathbf{z} = f^{\text{dVAE}}_{\phi^{\text{enc}}}(\boldsymbol{I}).$$

To train the entire model, we have the reconstruction objective supervising the learning of $\mathbf{z}$ with dVAE decoder $g^{\text{dVAE}}_{\phi^{\text{dec}}}$. Next, the objective for object-centric learning relies on the correct prediction from the auto-regressive transformer for predicting correct tokens:

$$\mathcal{L} = \mathcal{L}_{\text{dVAE}} + \mathcal{L}_{\text{CE}} \quad \text{where} \quad \mathcal{L}_{\text{dVAE}} = ||g^{\text{dVAE}}_{\phi^{\text{dec}}}(\mathbf{z}) - \boldsymbol{I}||_2^2, \ \mathcal{L}_{\text{CE}} = \sum_{l=1}^{L} \text{CrossEntropy}(\mathbf{z}_l, \mathbf{o}_l)$$

Under this setting, the model does not predict additional masks and relies on the attention $\boldsymbol{A}$ within the Slot-Attention module for obtaining slot-specific object masks. Although such models can achieve competitive results on real-world synthetic datasets, yet as our experiments suggest, they can be inferior to mixture-based decoders on segmentation in synthetic datasets. We suspect that this originates from the low resolution when discretizing images into tokens.

A.2 Optimization

Recall the bi-level optimization problem we introduced in Sec. 2.2.

$$\min_{\theta,\phi} f(\theta, \phi) \quad s.t. \quad \theta \in \arg\min_{\theta'} g(\theta', \phi), \tag{6}$$

where we call $f(\theta, \phi)$ the outer objective function and $g(\theta, \phi)$ the inner objective function. To jointly optimize both objectives w.r.t. parameters θ and ϕ, a straightforward approach to solving Eq. (6) is to represent the inner solution of θ as a function of ϕ, *i.e.*, $\theta^*(\phi) = \arg\min_{\theta'} g(\theta', \phi)$. Then we can optimize the outer objective with gradient descent:

$$\nabla_\phi f(\theta^*(\phi), \phi) = \nabla_\phi \theta^*(\phi) \nabla_1 f(\theta^*(\phi), \phi) + \nabla_2 f(\theta^*(\phi), \phi),$$

However, the difficulty of this method lies in the calculation of $\nabla_\phi \theta^*(\phi)$ where we need to solve linear equation from implicit gradient theorem:

$$\nabla_{1,2} g(\theta^*(\phi), \phi) \nabla_\phi \theta^*(\phi) + \nabla_{2,2} g(\theta^*(\phi), \phi) = 0.$$

If $\nabla_{2,2} g(\theta^*, \phi)$ is invertible, we can solve for $\nabla_\phi \theta^*(\phi)$ and obtain the gradient update on ϕ:

$$\phi_{k+1} = \phi_k - \xi \left(\nabla_2 f_k - (\nabla_{1,2} g_k)^\top (\nabla_{2,2} g_k)^{-1} \nabla_1 f_k \right)$$

where $\nabla_1 f_k = \nabla_2 f(\theta^*(\phi_k), \phi_k)$ and $\nabla_1 f_k = \nabla_1 f(\theta^*(\phi_k), \phi_k)$. Various methods have been proposed to approximate the solution (Pedregosa, 2016; Lorraine et al., 2020) and we refer the authors to Ye et al. (2022) for a thorough review of related methods.

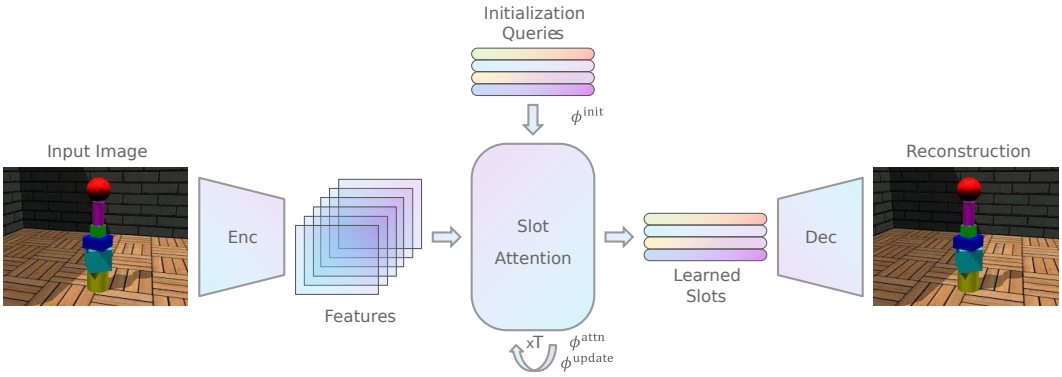

Figure 4: An illustrative visualization of our proposed BO-QSA slot-encoder. During the backward pass, BO-QSA uses straight-through gradient estimator to backpropagate gradients directly to ϕ^{init}, ϕ^{attn}, and ϕ^{update} without gradients into the iterative process.

Layer	Kernel Size	Stride	Padding	Channels	Activation
Conv	5x5	1(2)	2	64	ReLU
Conv	5x5	1	2	64	ReLU
Conv	5x5	1	2	64	ReLU
Conv	5x5	1	2	64	ReLU

Table 7: Configuration of CNN encoder used in our model. The values in parentheses are adopted for CLEVRTex and ShapeStacks

A.3 Implementation Details

We provide a visualization of our designed slot-encoder in Fig. 4 and discuss the implementation details for different experimentObjectsRoomal settings in the following sections.

A.3.1 Slot Initialization

We initialize all models with the number of slots shown in Tab. 11. During training, we add a small perturbation to the queries by sampling from a zero-mean distribution with variance σ as we found it empirically helpful for better performance. We perform annealing over σ to gradually eliminate the effect of this random perturbation during training. We adopt the cosine annealing strategy such that σ starts from 1 and gradually anneals to 0 after N_σ training steps, where N_σ is a hyperparameter that controls the annealing rate of σ. In our experiments, we use $N_\sigma = 0$ on Cars and Flowers and $N_\sigma = 30000$ on the rest of the datasets.

A.3.2 BO-QSA with Mixture-Based Decoders

For mixture-based decoders, we use the same Slot-Attention architecture as in Locatello et al. (2020) with slots initialized by learnable queries. Given an input image, Slot-Attention uses a CNN encoder to extract image features. After adding positional embedding, these features are input into the Slot-Attention module slot updates. Finally, these slots are decoded by the mixture decoder to reconstruct the input image. We provide the details of our image encoder in Tab. 7. For the mixture-based decoder, we use six transposed convolutional layers with ReLU activations following Locatello et al. (2020). We visualize the details of our mixture-based decoder design in Tab. 8. We train our model for 250k steps with a batch size of 128 and describe all training configurations and hyperparameter selection Tab. 9.

A.3.3 BO-QSA with Transformer-Based Decoder

For transformer-based decoders, we adopt the transformer architecture proposed by SLATE (Singh et al., 2021). For the transformer-based BO-QSA, unlike SLATE, we use the same CNN as in

Layer	Kernel Size	Stride	Padding	Channels	Activation
TransConv	5x5	2	2	64	ReLU
TransConv	5x5	2	2	64	ReLU
TransConv	5x5	2	2	64	ReLU
TransConv	5x5	2(1)	2	64	ReLU
TransConv	5x5	1	2	64	ReLU
TransConv	3x3	1	1	4	None

Table 8: Configuration of mixture decoder used in our model. The values in parentheses are adopted for ObjectsRoom

Batch Size	LR	Slot Dim	MLP Hidden Dim
128	4e-4	64	128
Warmup Steps	**Decay Steps**	**Max Steps**	**Sigma Down Steps**
5k	50k	250k	30k

Table 9: Training configuration for mixture-based model

mixture-based BO-QSA (instead of the dVAE encoder) to extract features from the image as input to the Slot-Attention module as we find such changes help solve the problem on coarse object boundary prediction mentioned in Sec. 5.1. Next, we use the same overall architecture of discrete VAE as mentioned in SLATE Singh et al. (2021). However, we change the kernel size of the DVAE encoder from 1 to 3 since we find that such changes can help increase model performance when decomposing scenes. We train our model for 250k steps with a batch size of 128, and all the training configuration in our experiments is described in Tab. 10.

Training	batch size	128
	warmup steps	10000
	learning rate	1e-4
	max steps	250k
DVAE	vocabulary size	1024
	Gumbel-Sofmax anealing range	1.0 to 0.1
	Gumbel-Sofmax anealing steps	30000
too	lr-dVAE(no warmup)	3e-4
Transformer Decoder	layers	4
	heads	4
	dropout	0.1
	hidden dimension	256
Slot Attention Module	slot dimension	256
	iterations	3
	σ annealing steps	30000(0)

Table 10: Training configuration for transformer-based model. The values in parentheses are adopted for Cars and Flowers dataset

A.3.4 BASELINES

The reproduction of Slot-Attention and SLATE follows the architecture and hyperparameter selection mentioned in their paper. Similar to our models, we train all baseline models with 250K steps on all datasets. For SLATE, we use the input image size of 96 on the ShapeStacks dataset as we find that the image size of 128 will cause all objects to be divided into the same slot, resulting in low ARI and MSC. For a fair comparison with numbers reported in SLATE's paper, we report the MSE of models by first computing per-pixel errors and then multiplying it by the total number of pixels.

	Model	Shapestacks	ObjectsRoom	Birds	Dogs	Flowers	Cars
# of slots	Slot-Attention	8	5	3	2	2	2
	SLATE	12	6	3	2	2	2
	BO-QSA+Mixture	8	5	3	2	2	2
	BO-QSA+Transformer	12	6	3	2	2	2
	Image Size	128	64	128	128	128	128

Table 11: The number of slots and image size used for each dataset

B Additional Experiments and Visualizations

B.1 Zero-shot Transfer

In this section, we continue on the discussion in Sec. 5.4 and provide additional zero-shot transfer results. Similarly, we use the notation ($X \rightarrow Y$) to denote the zero-shot adaptation of models trained unsupervisedly on dataset X to new datasets Y. We report transfer results from Stanford Dogs and CUB200 Birds to all other real image datasets. As we can see from Tab. 12, our model achieves the overall best results compared with other powerful Slot-Attention variants (models that achieve best or second-best results in our ablation studies as in Tab. 5) except for (Birds→Cars). However, our optimization method still helps improve zero-shot transfer for randomly initialized Slot-Attention.

Table 12: Zero-shot transfer results (mIoU ↑).

Model	Dogs → Cars	Dogs → Flowers	Dogs → Birds	Birds → Dogs	Birds → Cars	Birds → Flowers
SA	57.96	57.96	45.06	74.68	58.79	62.02
I-SA	58.05	58.06	48.88	71.16	69.90	68.67
BO-SA	58.10	58.10	47.96	71.81	**70.75**	67.95
BO-QSA(ours)	**75.50**	**63.43**	**52.49**	**76.66**	66.74	**70.74**

B.2 Experiments on Additional Datasets

In addition to datasets considered in Sec. 5, we conduct experiments on other synthetic datasets and visualize qualitative results. More specifically, we test our model on CLEVRTEX (Karazija et al., 2021) and PTR (Hong et al., 2021). CLEVRTEX is a synthetic dataset extending the original CLEVR dataset with complex textures, and PTR is a synthetic dataset of 3D objects from PartNet with rendering variations. We run our BO-QSA with the same configuration mentioned in Appendix A.3 previously and report the quantitative segmentation results on CLEVRTEX and visualize qualitative segmentation results on PTR.

As we can see in Tab. 13, our model also achieves state-of-the-art results on the unsupervised object segmentation task in CLEVRTEX. Interestingly, as CLEVRTEX lies in the middle of synthetic and real images, our design is limited by the selection of decoders where mixture-based decoders show low reconstruction accuracies and transformer-based decoders show lower segmentation accuracies because of the issues discussed in Sec. 5.1.

Table 13: Multi-object segmentation results on CLEVRTEX. We report ARI-FG and MSE of all models in the form of (mean ± variance) across 3 experiment trials with different random seeds. We visualize the best results in bold.

Model	CLEVRTEX	
	ARI-FG (%) ↑	MSE ↓
MONet (Burgess et al., 2019)	19.78±1.02	**146±7**
Slot-Attention (Locatello et al., 2020)	62.40±2.33	254±8
GENSIS-V2 (Engelcke et al., 2021)	31.19±12.41	315±106
DTI	79.90±1.37	438±22
ours (BO-QSA+mixture)	**80.47±2.49**	268.1±2.5

For PTR, we compare our method with the vanilla Slot-Attention module on multi-object segmentation. We report ARI-FG and MSC-FG scores of our model compared with the vanilla Slot-Attention on the PTR validation set. As we can see from Tab. 14, our model achieves similar performance compared with Slot-Attention on ARI-FG and significantly outperforms it on MSC-FG. We attribute this result to the capability of precisely segmenting objects. As ARI-FG applys masks to each slot prediction for calculating results, it does not require models to precisely segment the object from the background. However, MSC-FG uses a mIoU like measure that requires model to precisely predict the object boundaries. This indicates that our model is better at precisely segmenting objects without noises. We provide more qualitative results on CLEVRTEX and PTR in Appendix B.3. Similarly, we observe the binding of certain slots to scene backgrounds, but with more complex concepts, the binding of slots to concepts is not as straightforward as in ShapeStacks and CUB200 Birds.

Table 14: Multi-object segmentation results on PTR. We visualize the best results in bold.

Model	PTR	
	ARI-FG ↑	MSC-FG ↑
Slot-Attention	0.72	0.21
ours (BO-QSA+mixture)	**0.75**	**0.61**

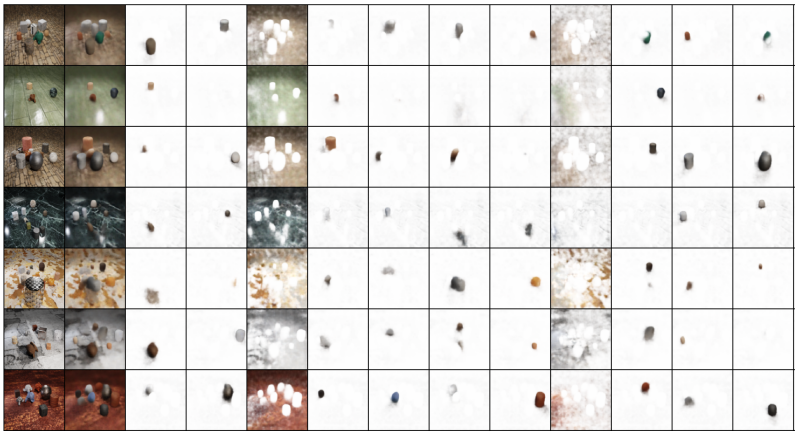

Figure 5: Unsupervised Multi-Object Segmentation on CLEVRTEX.

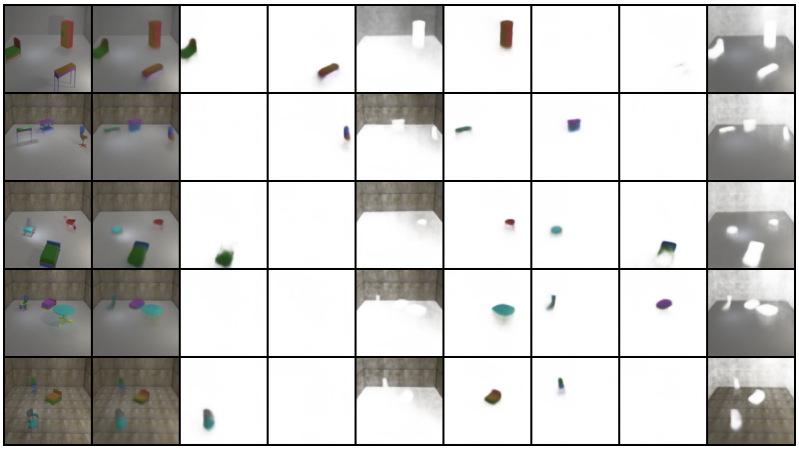

Figure 6: Unsupervised Multi-Object Segmentation on PTR.

B.3 ADDITIONAL VISUALIZATIONS

In this section, we visualize more qualitative results of our model on different datasets:

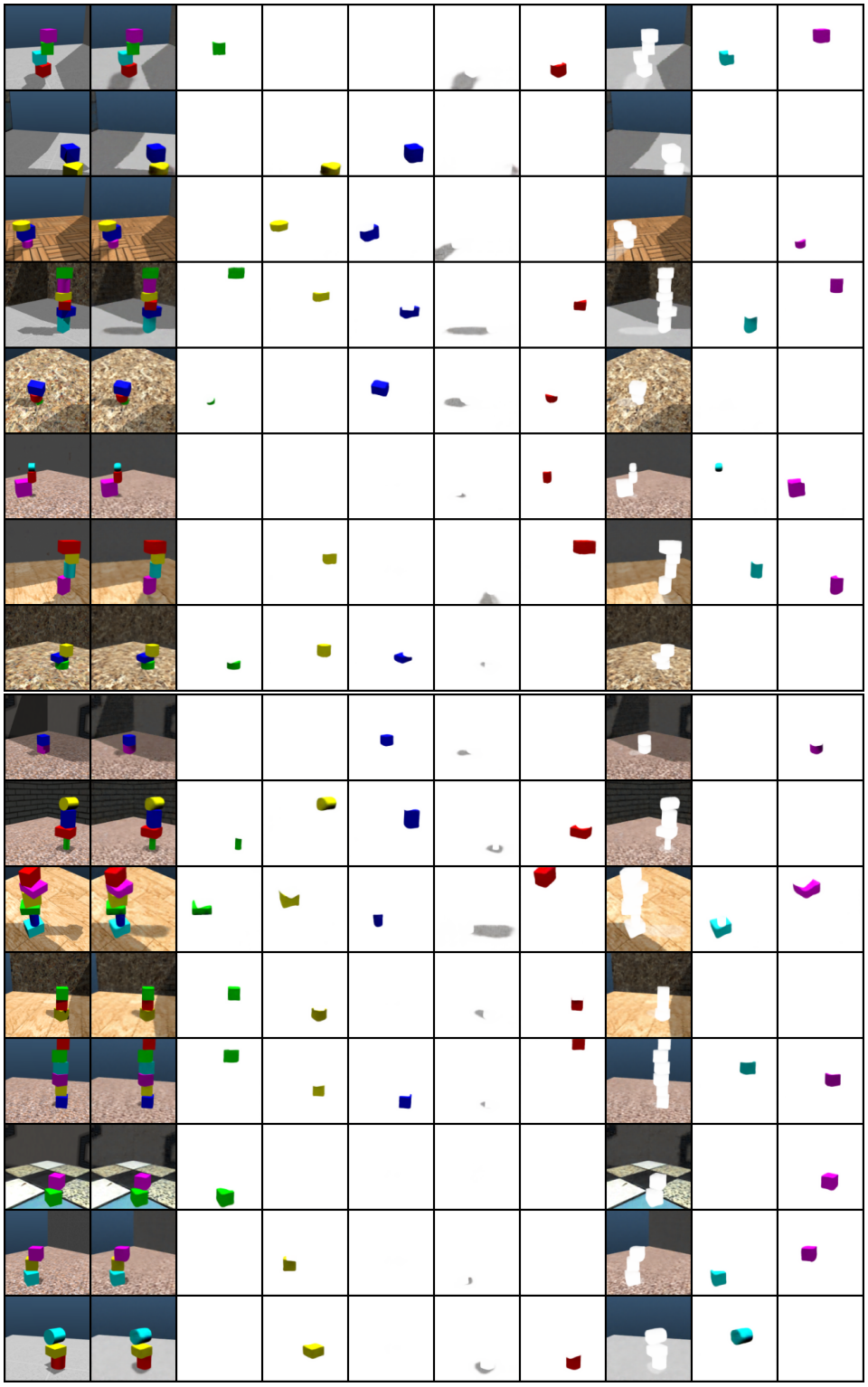

Figure 7: Unsupervised Multi-Object Segmentation on ShapeStacks.

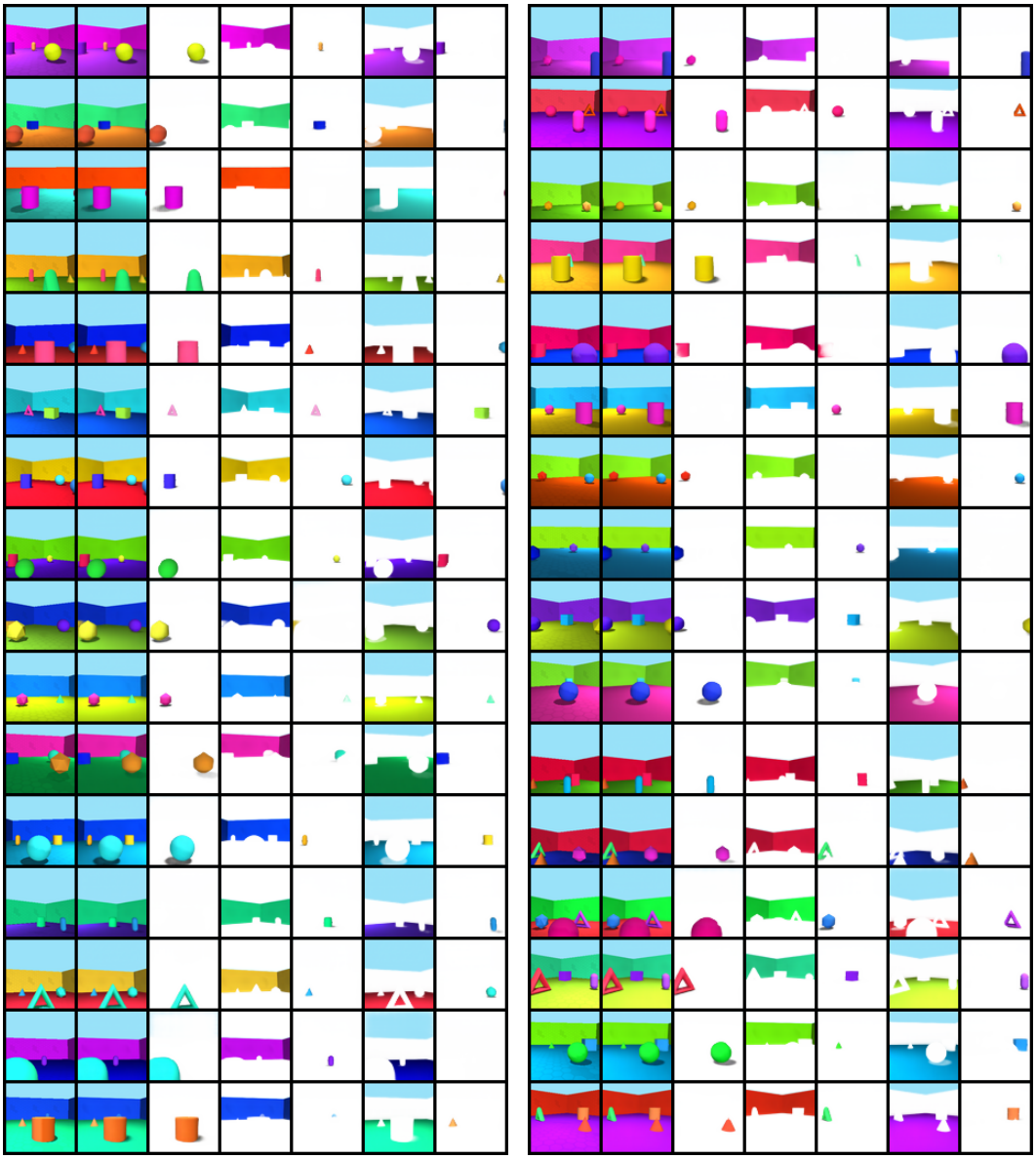

Figure 8: Unsupervised Multi-Object Segmentation on ObjectsRoom. In contrast to ShapeStacks, we observe consistent binding of slots to ground, wall, sky and also objects in the front.

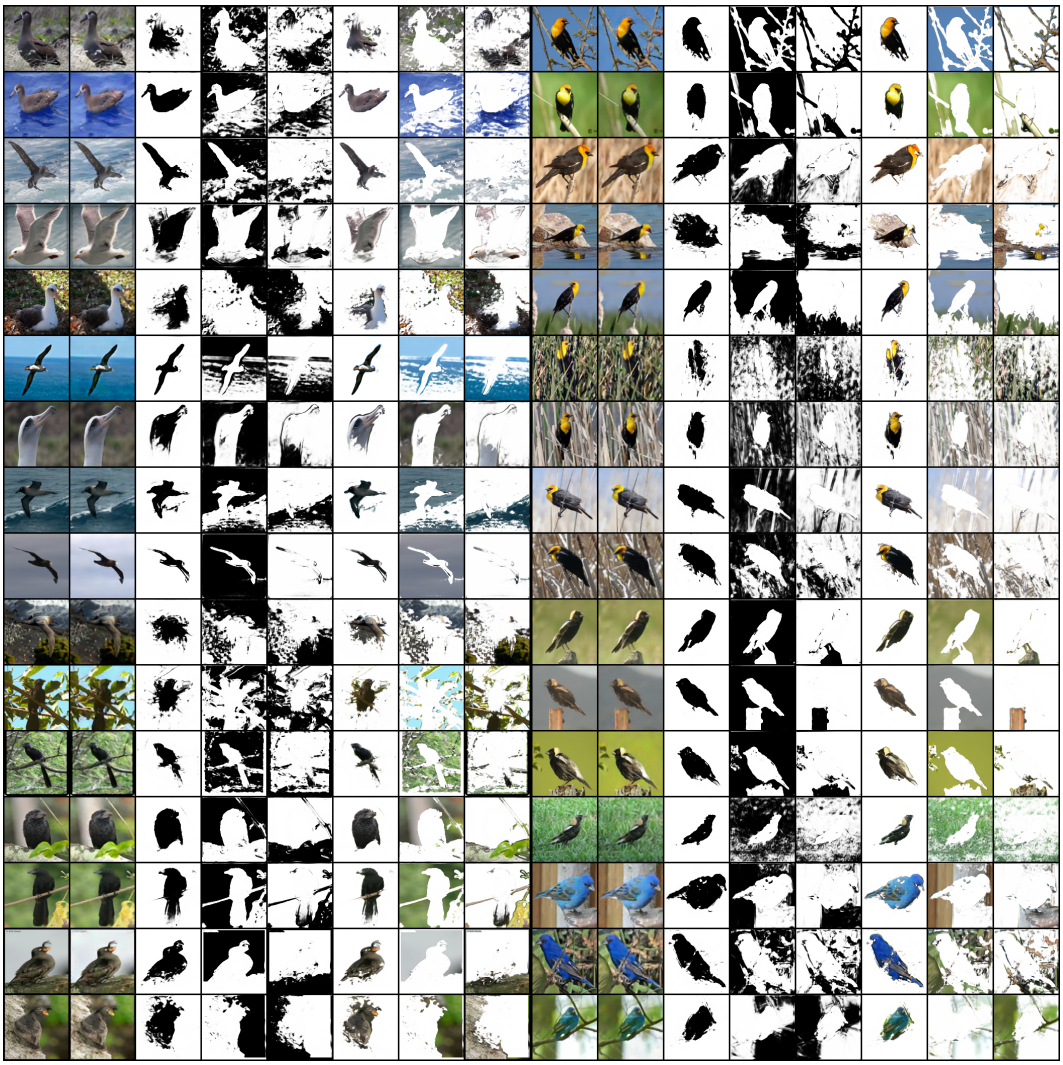

Figure 9: Unsupervised Foreground Extraction on CUB200 Birds.

Figure 10: Unsupervised Foreground Extraction on Stanford Dogs.

Figure 11: Unsupervised Foreground Extraction on Stanford Cars.

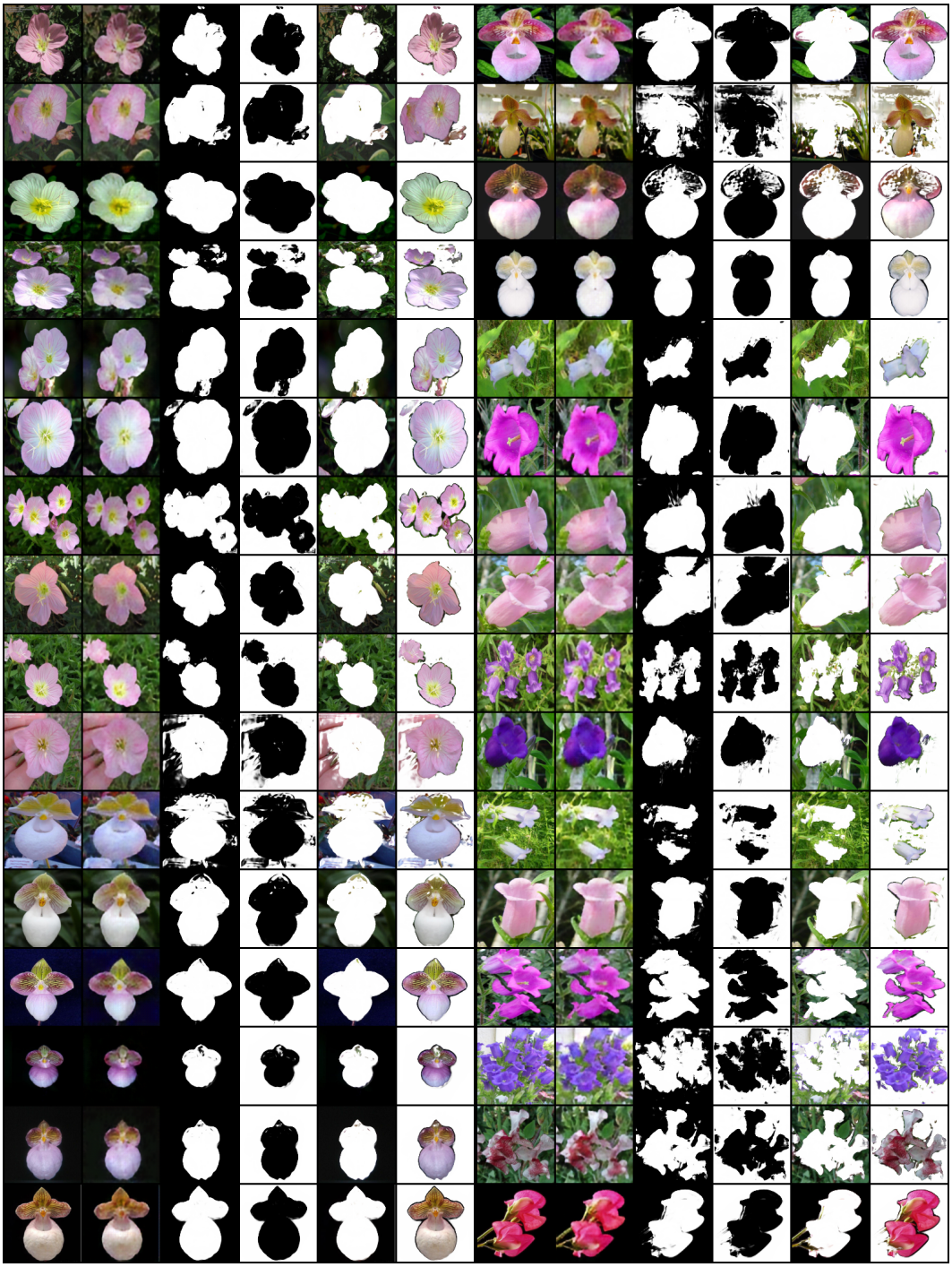

Figure 12: Unsupervised Foreground Extraction on Caltech Flowers.