ORTHORF: EXPLORING ORTHOGONALITY IN OBJECT-CENTRIC REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural synchrony is hypothesized to help the brain organize visual scenes into structured multi-object representations. In machine learning, synchrony-based models analogously learn object-centric representations by storing binding in the phase of complex-valued features. Rotating Features (RF) instantiate this idea with vector-valued activations, encoding object presence in magnitudes and affiliation in orientations. We propose Orthogonal Rotating Features (OrthoRF), which enforces orthogonality in RF's orientation space via an inner-product loss and architectural modifications. This yields sharper phase alignment and more reliable grouping. In evaluations of unsupervised object discovery, including settings with overlapping objects, noise, and out-of-distribution tests, OrthoRF matches or outperforms current models while producing more interpretable representations, and it eliminates the post-hoc clustering required by many synchrony-based approaches. Unlike current models, OrthoRF also recovers occluded object parts, indicating stronger grouping under occlusion. Overall, orthogonality emerges as a simple, effective inductive bias for synchrony-based object-centric learning.

1 INTRODUCTION

Decomposing scenes into constituent parts is a long-standing strategy in computer vision. Classical approaches factorized images into surfaces and objects with properties like reflectance, albedo, and geometry (Coakley, 2003). Deep learning has renewed this effort by learning structured representations (Zhang et al., 2013; Dittadi, 2023). In practice, this is often operationalized as Object-Centric Learning (OCL) (Greff et al., 2020), where models discover modular, compositional object representations that support generalization and relational reasoning on many downstream visual tasks (Ding et al., 2021; Bapst et al., 2019; Mandikal & Grauman, 2021). At its core, much like human perception (Spelke, 1990), OCL addresses the binding problem (Roskies, 1999): flexibly integrating features (such as color, shape, texture) into a unified perception. This view aligns with cognitive and neuroscientific accounts that posit neural synchrony (Singer, 2007) as a key mechanism, whereby temporally synchronized oscillations bind distributed information into coherent objects (see Fig 1).

A dominant OCL design to binding uses a collection of discrete latent vectors, named "slots" (Locatello et al., 2020), each dedicated to the features of a single object. As slot-based methods evolve, radically new ideas are emerging in OCL, some of which are inspired by synchrony (Mozer et al., 1991; Reichert & Serre, 2013). In this less-explored paradigm, binding is expressed via the relative phases of complex-valued neural activations, with phase-space distances serving as an implicit relational metric between object instances. Recently, synchrony-based models (Löwe et al., 2022; Gopalakrishnan et al., 2024; Miyato et al., 2024) have achieved unsupervised object discovery on synthetic and more naturalistic scenes, surpassing prior supervised approaches (Mozer et al., 1991).

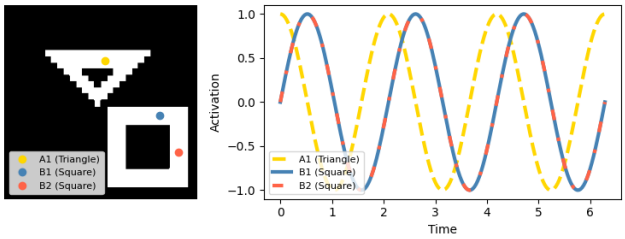

Figure 1: Binding-by-synchrony. Local excitation/inhibition partitions neurons into phase-based groups. Same-phase firing (e.g., B1–B2) encodes the same object; out-of-phase firing (e.g., A1 vs. B1/B2) encodes different objects.

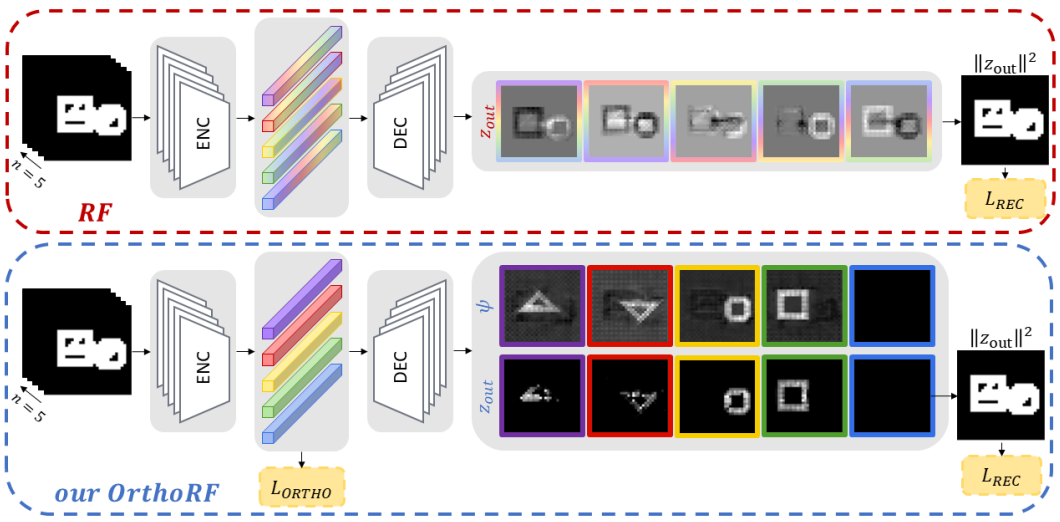

Figure 2: RF vs. our OrthoRF. In RF, object content is distributed across orientation space (seen as multicolored bars in encoder activations and mixed final outputs). OrthoRF enforces orthogonal orientation axes, routing each object to a distinct latent, yielding cleaner separation and improved handling of overlap regions.

Current state-of-the-art synchrony-based models, such as Complex-valued Autoencoders (CAE) (Löwe et al., 2022), use real-valued weights to process complex-valued activations, by sharing weights across the real and imaginary parts. In these models, magnitudes encode feature presence, whereas phases encode object affiliation. During training, they leverage the natural constructive/destructive interference through the addition of complex activations in every layer, promoting phase alignment for features of the same object and phase separation for different objects. A related work, named Rotating Features (RF) (Löwe et al., 2023), replaces complex-valued with vector-valued (i.e., n-dimensional) activations that rotate on a hypersphere, extending representational capacity beyond 2D complex planes. To (de)synchronize phases for object representation, these models use an additional inductive bias via a gating mechanism (Reichert & Serre, 2013) that strengthens interactions among similarly oriented features and weakens those among dissimilarly oriented ones. However, this mechanism can be hard to interpret. To improve interpretability, Löwe et al. (2024b) introduced cosine binding, a more transparent alternative that is based on cosine similarity between activations. This, however, entails substantial memory overhead, since it requires computing and storing many similarities between inputs and intermediate outputs.

Synchrony-based models have important limitations. Unlike slot-based methods, which yield discrete, object-aligned slots (one slot $\approx$ one object), state-of-the-art synchrony-based models (Löwe et al., 2022; 2023; Stanić et al., 2023; Miyato et al., 2024; Gopalakrishnan et al., 2024) produce distributed representations. While this can be more flexible, it makes the output hard to use without extra machinery. In practice, they require post-hoc clustering in phase space to recover objects, grouping features by phase. This distributed coding also degrades performance in overlap regions: features from occluding objects become uncertain and drift farther from cluster centers (Löwe et al., 2023), complicating object assignment. As a result, evaluations of these models often exclude overlapping regions, potentially underrepresenting the regimes where robust binding is most needed.

In this paper, we advance synchrony-based models, specifically the RF autoencoder (Löwe et al., 2023), by improving both interpretability and representational capacity. Motivated by evidence that orthogonality enhances efficiency and encourages disentanglement (Ranasinghe et al., 2021), we propose Orthogonal Rotating Features (OrthoRF), an RF extension that enforces orthogonal object encoding in the phase space. Our approach has two components: (i) a softmax-based competitive binding that drives n-dimensional activations to specialize on distinct input components (objects), and (ii) an inner-product-based orthogonality loss that enforces a 90° separation between object representations in phase space. Together, these yield sharper phase alignment and more reliable grouping: features of the same object concentrate along a single vector dimension, producing one-

hot–like object encodings. In unsupervised object discovery evaluation, across scenarios with overlapping objects, noise, and out-of-distribution tests, OrthoRF matches or surpasses current methods, eliminates post-hoc clustering, and recovers occluded object parts in intermediate representations (a capability not shown by slot-based or prior synchrony-based models). These results underscore orthogonality as a simple yet powerful inductive bias for synchrony-based object-centric learning.

2 Background: Rotating Features

We build upon the RF autoencoder (Löwe et al., 2023), which replaces scalar features with n-dimensional vectors whose magnitudes encode feature presence and whose orientations encode object affiliation. Specifically, a standard feature vector $\mathbf{z} \in \mathbb{R}^d$ is lifted to $\mathbf{z}_{\text{rotating}} \in \mathbb{R}^{n \times d}$, whose per-feature magnitude $\mathbf{m} = \|\mathbf{z}_{\text{rotating}}\|_2 \in \mathbb{R}^d$ (the ℓ_2-norm over the n-dimension) plays the role of a standard neural activation. This lifting applies both to input images (initialized with zeros along the orientation dimension) and to activations at any layer. Given a neural layer $f_{\mathbf{w}}$ with d_{in} inputs and d_{out} outputs, an input $\mathbf{z}_{\text{in}} \in \mathbb{R}^{n \times d_{\text{in}}}$ is transformed using a weight matrix $\mathbf{w} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$, shared across the n components, and a bias $\mathbf{b} \in \mathbb{R}^{n \times d_{\text{out}}}$, as follows:

$$\boldsymbol{\psi} = f_{\mathbf{w}}(\mathbf{z}_{in}) + \mathbf{b} \ \in \mathbb{R}^{n \times d_{out}}. \tag{1}$$

To ensure that similar oriented features are processed together, RF uses a gating mechanism[1] (Reichert & Serre, 2013). Specifically, it applies the shared weights $\mathbf{w} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ to both the inputs $\mathbf{z}_{\text{in}} \in \mathbb{R}^{n \times d_{\text{in}}}$ and their per-feature magnitudes $||\mathbf{z}_{\text{in}}||_2 \in \mathbb{R}^{d_{\text{in}}}$, then combines the results:

$$\boldsymbol{\chi} = f_{\mathbf{w}}(||\mathbf{z}_{\text{in}}||_2) \in \mathbb{R}^{d_{\text{out}}}, \tag{2}$$

$$\mathbf{m}_{\text{bind}} = 0.5 \cdot ||f_{\mathbf{w}}(\mathbf{z}_{\text{in}})||_2 + 0.5 \cdot \boldsymbol{\chi} \in \mathbb{R}^{d_{\text{out}}}. \tag{3}$$

The gated magnitude $\mathbf{m}_{\text{bind}}$ is passed through ReLU to enforce non-negativity, and then used to rescale $\boldsymbol{\psi}$, ensuring the output of the layer retains $\boldsymbol{\psi}$'s orientation, as follows:

$$\mathbf{m}_{out} = \text{ReLU}(\text{BatchNorm}(\boldsymbol{m}_{\text{bind}})) \in \mathbb{R}^{d_{\text{out}}}, \tag{4}$$

$$\mathbf{z}_{\text{out}} = \mathbf{m}_{\text{out}} \odot \frac{\boldsymbol{\psi}}{||\boldsymbol{\psi}||_2} \in \mathbb{R}^{n \times d_{\text{out}}}. \tag{5}$$

The reconstructed image is obtained by computing the per-pixel magnitude of the final-layer activations, $\|\mathbf{z}_{\text{final}}\|_2 \in \mathbb{R}^{d_{\text{image}}}$ with $d_{\text{image}} = c \times h \times w$ (where c, h, w are channels, height, and width), scaling it with a learnable scalar weight $w' \in \mathbb{R}$ and bias $b' \in \mathbb{R}$, and then applying a sigmoid:

$$\hat{\mathbf{x}} = \text{Sigmoid}(w'||\mathbf{z}_{\text{final}}||_2 + b') \ \in \mathbb{R}^{d_{\text{image}}}. \tag{6}$$

During training, an MSE loss is used between the input and reconstructed images, $\mathcal{L}_{\text{REC}} = \text{MSE}(\mathbf{x}, \hat{\mathbf{x}})$. The vector-valued activations add across layers, producing constructive interference for features of the same object and destructive interference for different objects. Because regions of the same object exhibit high pointwise mutual information, destructive interference would hurt reconstruction, so training implicitly encourages within-object alignment and across-object anti-alignment in features. For object discovery, k-means is applied to the output $\mathbf{z}_{\text{final}} \in \mathbb{R}^{n \times d_{\text{image}}}$, assigning each pixel to an object-cluster. See Löwe et al. (2023) for further details on RF.

3 Method

3.1 Motivation

Synchrony-based architectures such as RF (Löwe et al., 2024a) show that these n-dimensional features can support object discovery, yet their representations are distributed (see Fig. 2) across dimensions, which in turn demands post-hoc clustering (e.g., k-means) to recover objects. This dependence makes the pipeline fragile and less practical; a single object may occupy multiple dimensions, creating redundancy and blurring boundaries, particularly in overlap regions where features drift away from cluster centers (as noted by Löwe et al. (2024a)) and assignments become uncertain. That uncertainty, however, carries informative cues: RF's behavior in overlaps reveals occlusion signals

[1] There is an inconsistency between the description in the paper and the implementation in the code regarding the calculation of $\mathbf{m}_{bind}$. While the paper states that $\mathbf{m}_{bind} = 0.5 \cdot ||\boldsymbol{\psi}||_2 + 0.5 \cdot \boldsymbol{\chi}$, the code utilizes the formula described in Equation 3 for $\mathbf{m}_{bind} \in \mathbb{R}^{d_{out}}$. We used the latter implementation as it performs better.

that slot-based OCL methods (Anciukevicius et al., 2020) seldom exploit. We investigate whether an architectural bias can preserve RF's strengths while addressing these drawbacks. Guided by evidence that orthogonality sharpens discrimination and fosters disentanglement (Lezama et al., 2018; Ranasinghe et al., 2021; Sun et al., 2017; Chen et al., 2020; Liu et al., 2018; Wang et al., 2018), we impose orthogonality constraints in the RF orientation space so that each object collapses onto a single component in this n-dimensional orientation space, reducing redundancy, removing the need for clustering, and converting overlap-driven uncertainty into a reliable cue for occlusion recovery.

3.2 Orthogonal rotating Features

In this section, we present the architectural modifications that yield competitive binding, and an orthogonality loss enforcing 90° separation among latents in the orientation space. We also highlight key properties that emerge from these modifications. Fig. 2 illustrates the overall OrthoRF architecture.

Competitive binding in orientation space We model object–component assignment as a discrete competition: each object should map to one component in the n-dimensional orientation space. Inspired by multi-class classification, where a softmax layer maps logits to a categorical distribution, and by OCL methods such as Slot Attention (Locatello et al., 2020), we use the same mechanism to induce competition that drives object-oriented specialization across components. In OrthoRF autoencoder, we apply a per-layer softmax over orientation components, yielding winner-take-most assignments and object-wise specialization. To improve stability and prevent component collapse (e.g., all features mapped to one component while others are never used), we apply centering before the softmax, but only to the encoder's output vectors. Empirically, this removes biases that would otherwise let a single component dominate (Caron et al., 2021). Specifically, after Eq. 1, we use the intermediate output $\boldsymbol{\psi} \in \mathbb{R}^{n\times d}$ (rows i: orientation components; columns j: features). For each feature index j, we apply softmax over components i after subtracting the per-feature mean logit to obtain assignment probabilities, as follows:

$$\psi'_{ij} = \frac{\exp\big(\psi_{ij} - \bar{\psi}_j\big)}{\sum_{k=1}^{n} \exp\big(\psi_{kj} - \bar{\psi}_j\big)}, \quad \text{where} \quad \bar{\psi}_j = \frac{1}{n}\sum_{k=1}^{n} \psi_{kj}. \tag{7}$$

The remaining steps follow Section 2.

Orthogonality regularization We enforce orthogonality among latent orientation components at the encoder output, since this stage aggregates global features and offers a lower-dimensional representation, reducing computational cost. To implement this, we use the encoder's output $\mathbf{z} \in \mathbb{R}^{\mathrm{bs}\times n\times z_{\mathrm{dim}}}$, (bs: batch size, n: orientation components, z_{dim}: feature dimension). For each sample $i \in \{1,\dots,\mathrm{bs}\}$ and feature $j \in \{1,\dots,z_{\mathrm{dim}}\}$, we center across orientation components:

$$\tilde{z}_{ikj} \;=\; z_{ikj} - \bar{z}_{ij}, \qquad \text{where} \qquad \bar{z}_{ij} \;=\; \frac{1}{n}\sum_{m=1}^{n} z_{imj}. \tag{8}$$

After centering, we stack the n latent vectors for a sample i as the rows of $\tilde{Z}_i \in \mathbb{R}^{n\times z_{\mathrm{dim}}}$ and define the Gram matrix, as follows:

$$G_i \;=\; \tilde{Z}_i \tilde{Z}_i^{\top} \;\in\; \mathbb{R}^{n\times n}. \tag{9}$$

Here, $(G_i)_{k\ell} = \langle \tilde{Z}_{i,k,:},\, \tilde{Z}_{i,\ell,:} \rangle$ is the (unnormalized) inner product between the component vectors k and ℓ. If different components encode distinct information, these inner products should be small (ideally zero) off the diagonal. Then, we penalize the square off-diagonal mass, averaged over samples and unique component pairs:

$$\mathcal{L}_{\mathrm{ortho}} \;=\; \frac{1}{\mathrm{bs}\, n(n-1)} \sum_{i=1}^{\mathrm{bs}} \Big(\mathrm{offdiag}(G_i) \Big)^2 \;=\; \frac{1}{\mathrm{bs}\, n(n-1)} \sum_{i=1}^{\mathrm{bs}} \sum_{k\neq\ell}^{n} (G_i)_{k\ell}^2. \tag{10}$$

By squaring and averaging the $(G_i)_{k\ell}$ terms, we drive cross-component similarities toward zero, thereby decorrelating the embeddings and promoting orthogonality. The factor $n(n-1)$ normalizes by the number of ordered pairs (or twice the number of unique pairs). Finally, the full objective includes the orthogonality term weighted by λ:

$$\mathcal{L}_{\mathrm{total}} \;=\; \mathcal{L}_{\mathrm{REC}} \;+\; \lambda\, \mathcal{L}_{\mathrm{ortho}}, \qquad \lambda > 0. \tag{11}$$

Table 1: Object discovery on 4Shapes dataset (mean ± standard deviation over 4 seeds). OrthoRF matches RF under identical output post-processing, but surpasses it on the shape-completion task, measured by $\mathrm{MBO_i^{OV}}$ (OV: overlapping regions), when evaluated on ψ. *MSE, ARI-BG, and $\mathrm{MBO_i}$ are taken from Löwe et al. (2023); RF's ARI-BG and $\mathrm{MBO_i}$ were recomputed (see Table 5 Appendix). Best results are highlighted as **first**, second, and third.

Model	n	λ	MSE ↓	ARI-BG ↑	$\mathrm{MBO_i}$ ↑	$\mathrm{MBO_i^{OV}}$ ↑
*AE	-	-	5.492e-03 ± 9.393e-04	-	-	-
*CAE	-	-	3.435e-03 ± 2.899e-04	0.694 ± 0.041	0.628 ± 0.039	-
*$\mathrm{RF}^{\mathrm{kmeans}}_{z_{\mathrm{final}}}$	5	-	5.439e-04 ± 6.984e-05	0.975 ± 0.003	0.934 ± 0.006	0.8049 ± 0.013
	6	-	2.526e-04 ± 1.416e-05	0.991 ± 0.002	0.970 ± 0.003	0.8111 ± 0.009
	7	-	**1.642e-04 ± 1.810e-05**	0.992 ± 0.002	0.974 ± 0.003	0.8172 ± 0.001
	8	-	**1.360e-04 ± 7.644e-06**	0.987 ± 0.003	**0.968 ± 0.008**	0.8196 ± 0.001
	9	-	1.119e-03 ± 5.715e-04	0.9949 ± 0.002	**0.989 ± 0.002**	0.8170 ± 0.001
$\mathrm{OrthoRF}^{\mathrm{kmeans}}_{z_{\mathrm{final}}}$	5	0.8	**2.330e-04 ± 0.942e-03**	**0.9995 ± 0.001**	**0.9887 ± 0.003**	0.8204 ± 0.002
	6	0.5	**1.963e-04 ± 0.855e-03**	**0.9941 ± 0.009**	**0.9856 ± 0.006**	0.8151 ± 0.006
	7	0.1	1.660e-04 ± 0.942e-03	**0.9947 ± 0.006**	**0.9856 ± 0.008**	0.8161 ± 0.004
	8	0.09	3.714e-04 ± 0.768e-03	**0.9924 ± 0.009**	0.9527 ± 0.018	0.8022 ± 0.003
	9	0.08	**2.133e-04 ± 0.991e-03**	**0.9955 ± 0.002**	0.9849 ± 0.005	0.8183 ± 0.001
$\mathrm{OrthoRF}^{\mathrm{thresh.}}_{\psi_{\mathrm{final}}}$	5	0.8	2.330e-04 ± 0.942e-03	0.9934 ± 0.001	0.9843 ± 0.009	**0.9832 ± 0.006**
	6	0.5	1.963e-04 ± 0.855e-03	0.9869 ± 0.004	0.9845 ± 0.004	**0.9853 ± 0.003**
	7	0.1	1.660e-04 ± 0.942e-03	0.9763 ± 0.001	0.9730 ± 0.002	**0.9794 ± 0.005**
	8	0.09	3.714e-04 ± 0.768e-03	0.9682 ± 0.005	0.9604 ± 0.002	**0.9680 ± 0.002**
	9	0.08	2.133e-04 ± 0.991e-03	0.9631 ± 0.003	0.9678 ± 0.002	**0.9875 ± 0.009**

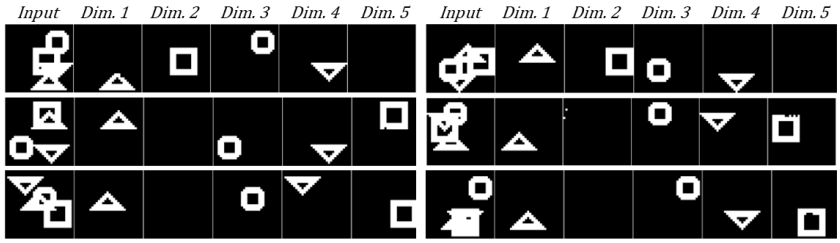

Figure 3: Qualitative OrthoRF results on 4Shapes, after thresholding ψ_{final}. Objects occupy distinct dimensions, and occluded parts are recovered.

Equivariance The OrthoRF autoencoder exhibits a key property, analogous to Slot Attention (Locatello et al., 2020), namely permutation equivariance over orientation components. For a representation $\mathbf{x} \in \mathbb{R}^{\mathrm{bs}\times n\times d}$ (e.g. output of encoder) with n the orientation axis, any permutation Π acting on this axis satisfies $f(\Pi\,\mathbf{x}) = \Pi\, f(\mathbf{x})$. This property arises from weight sharing across orientation components at every layer, which guarantees identical processing for each component.

Magnitude gating and occlusion completion In the final binding step (Eq. 5), the output is $\mathbf{z}_{\mathrm{out}} = m_{\mathrm{out}} \odot \frac{\boldsymbol{\psi}}{||\boldsymbol{\psi}||_2}$, where the magnitude $\mathbf{m}_{\mathrm{out}}$ serves as a visibility gate: visible regions pass, occluded regions are suppressed. The pre-gated content ψ shows occlusion-complete shapes (Fig. 2). A plausible explanation is that ψ is predicted from learned shape priors under the reconstruction objective, yielding completion behind occluders, while $\mathbf{m}_{\mathrm{out}}$ encodes visibility. This selective behavior depends on the softmax over orientation channels (competitive binding), which enables clean gating at the final layer.

4 Experiments

4.1 Experimental settings

Datasets We adopt the evaluation protocol of Löwe et al. (2024a; 2022); Stanić et al. (2023), which benchmarks object discovery on datasets with varying numbers of geometric shapes. Because RF emphasizes the 4Shapes dataset (see Fig. 3), we likewise evaluate OrthoRF on it. We use the binary

Table 2: Object discovery on the SEM dataset. OrthoRF outperforms RF under severe occlusions and noise, and shows stronger out-of-distribution generalization (noisy testing after clean training, and vice versa). Best results are highlighted as **first**, second, and third.

Train / Test	Model	n	Noise-free MSE ↓	Noise-free ARI-BG ↑	Noise-free MBO_i ↑	Noisy MSE ↓	Noisy ARI-BG ↑	Noisy MBO_i ↑
Noise-free	K-means	-	-	0.8427	0.6546	-	-	-
	Histogram	-	-	0.8011	0.6582	-	-	-
	$RF^{kmeans}_{z_{final}}$	5	**0.0001**	0.9551	0.6834	**0.0057**	0.6942	0.4146
	$OrthoRF^{thesh.}_{\psi_{final}}$	5	0.0002	**0.9908**	**0.7171**	0.0624	**0.7610**	**0.5644**
Noisy	K-means	-	-	-	-	-	0.3381	0.3442
	Histogram	-	-	-	-	-	0.3333	0.3612
	$RF^{kmeans}_{z_{final}}$	5	**0.0042**	0.8816	0.6044	**0.0007**	0.7043	0.4154
	$OrthoRF^{thesh.}_{\psi_{final}}$	5	0.0051	**0.9836**	**0.6705**	**0.0007**	**0.8356**	**0.6268**

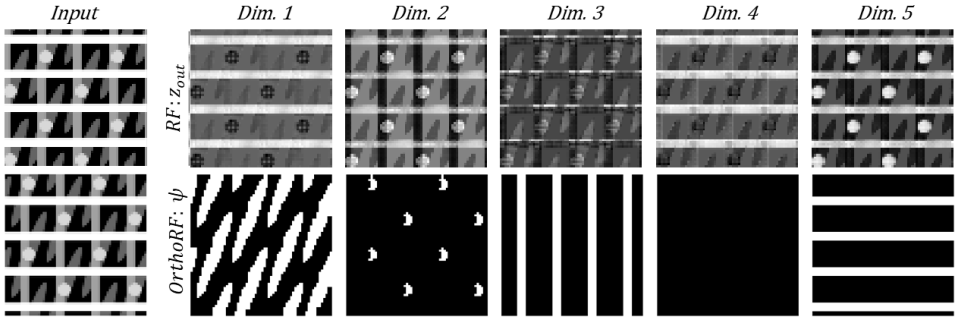

Figure 4: Qualitative results on noise-free SEM test images for $n = 5$. RF's output (top row) is spread across the orientation components, whereas the thresholded OrthoRF's output (bottom row) separates the SEM layers and reveals occluded structures.

variant rather than RGB, motivated by RF's degraded performance with color; its accuracy drops markedly on 4Shapes even with only five colors, when no pretrained features (from DINO Caron et al. (2021)) are used. Each binary image contains four objects (square, up-pointing triangle, down-pointing triangle, and circle) placed at random locations with overlap, all sharing a single color.

For a more realistic application, we introduce a synthetic Scanning Electron Microscope (SEM) dataset of semiconductor materials[2] (see Fig. 4). Each image shows a four-layer stack along the vertical z-axis, viewed from above. Each layer contains a single shape class (horizontal lines, circles, vertical lines, or ellipses) repeated across that layer. Layers are horizontally displaced relative to one another, with offsets varying per image. We provide two variants: (i) noise-free, and (ii) noisy (Gaussian blur + additive Gaussian noise) to mimic SEM acquisition artifacts. We choose this synthetic dataset because it tests three key challenges for object-centric models: (i) severe inter-layer occlusions from higher z-layers, (ii) domain relevance to semiconductor metrology, where separating stacked layers is essential, and (iii) robustness to acquisition noise. This dataset is not publicly available, but it can be easily recreated.

Evaluation metrics Following standard practices in object discovery, we evaluate performance using the Adjusted Rand Index (ARI) (Hubert & Arabie, 1985; Rand, 1971; Greff et al., 2019) and Mean Best Overlap (MBO) (Pont-Tuset et al., 2016; Seitzer et al., 2022). ARI measures clustering similarity, with 0 indicating chance-level agreement and 1 a perfect match; we compute it using decoder-generated object masks compared to ground truth, excluding the background (ARI-BG). MBO pairs predicted and ground-truth masks by overlap and averages their Intersection-over-Union (IoU) values, including background pixels, to assess alignment with object boundaries.

[2]Public source that describes the 3D-layer structure used for the SEM images of our synthetic dataset

Table 3: Quantifying similarity in phase space via mean pairwise cosine angles (degrees). On 4Shapes, OrthoRF is near-orthogonal with lower variance; on SEM, RF has a slightly higher mean but larger variance. Best results are highlighted as **first**, and second.

Data	Model	λ	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Avg.
4Shapes	RF	-	46.7 ± 0.96	80.3 ± 2.1	64.7 ± 1.5	67.7 ± 2.4	$\mathbf{87.0 \pm 1.7}$	69.28 ± 13.91
	orthoRF	0.5	$\mathbf{88.5 \pm 1.29}$	$\mathbf{88.8 \pm 1.3}$	$\mathbf{89.6 \pm 1.0}$	$\mathbf{89.3 \pm 1.0}$	78.1 ± 1.7	$\mathbf{86.86 \pm 4.39}$
SEM	RF	-	$\mathbf{82.6 \pm 2.6}$	$\mathbf{88.6 \pm 2.6}$	69.9 ± 2.1	$\mathbf{82.4 \pm 4.7}$	$\mathbf{87.1 \pm 3.3}$	$\mathbf{82.12 \pm 6.57}$
	orthoRF	0.1	80.0 ± 3.7	79.0 ± 2.2	$\mathbf{85.8 \pm 2.4}$	75.5 ± 3.1	80.7 ± 2.2	80.2 ± 3.32

Table 4: Quantifying separability in phase space via inter- and intra-cluster metrics on 4Shapes dataset. OrthoRF shows tighter clusters (lower intra) and near-orthogonal separation, whereas RF has a higher mean inter-angle but large variability. Best results are highlighted as **first**, and second.

Model	Inter-cluster ↑	Intra-cluster ↓
RF	$\mathbf{106.469 \pm 23.5752}$	17.285 ± 6.3811
OrthoRF	89.923 ± 0.0669	$\mathbf{1.085 \pm 4.3842}$

Implementation details The OrthoRF implementation follows RF (Löwe et al., 2024a) using a convolutional autoencoder. Architectural details appear in Table 6 (Appendix). Models are trained with Adam (Kingma & Ba, 2015), batch size 16, for 200k steps on all datasets. Experiments were run in PyTorch (Paszke et al., 2019) on a single NVIDIA Tesla T4 (16 GB). Additional training settings are listed in Table 7 (Appendix).

4.2 Results

Evaluation on the 4Shapes dataset Table 1 compares OrthoRF with RF (Löwe et al., 2024a), AE (Löwe et al., 2024a), and CAE (Löwe et al., 2022) on 4Shapes under two protocols. First, visible-only object discovery evaluates model outputs ($\mathbf{z}_{\text{final}}$ for RF/OrthoRF) against ground-truth masks that exclude overlapping regions (as defined by Löwe et al. (2024a)[3]). Second, we evaluate shape completion with $\text{MBO}_\text{i}^{\text{OV}}$, scoring full-object recovery, including overlapping (OV) regions, by comparing predictions to full-shape instance groundtruth masks (the labeling scheme used is described in Appendix A.2). In the table, we first report RF's $\mathbf{z}_{\text{final}}$ with its standard post-processing—output normalization, magnitude masking, and k-means. Next, we apply the same pipeline to OrthoRF's $\mathbf{z}_{\text{final}}$ for a fair comparison. Finally, we evaluate OrthoRF's intermediate map ψ_{final} by binarizing images with a threshold of 0.1 and no further post-processing. Furthermore, we vary the orientation dimensionality n, choosing $n \geq$ (number of objects + background).

Table 1 shows that OrthoRF matches RF on visible-only discovery for ARI-BG and MBO_i metrics, but outperforms on shape completion (last column) when evaluated on ψ_{final}. With a global threshold, OrthoRF's ψ_{final} reaches approximately 0.98 $\text{MBO}_\text{i}^{\text{OV}}$ at $n = 5$, while the same value for $\mathbf{z}_{\text{out}}$ of both RF and OrthoRF is approximately 0.80. Post-processing largely explains these outcomes: because RF doesn't cleanly separate objects in $\mathbf{z}_{\text{out}}$, it uses k-means to recover memberships. However, k-means enforces one label per pixel, so overlaps get credit for only a single object. Thresholding ψ_{final} instead permits multi-label pixels in overlapping regions, improving MBO_i^{OV}. Overall, performance for both models remains stable across n (minor fluctuations). Finally, OrthoRF and RF substantially outperform AE and CAE. Fig. 3 shows qualitative OrthoRF results after thresholding ψ_{final}; objects separate into distinct dimensions, and occluded parts are recovered.

Evaluation on the SEM datasets Table 2 reports visible-only object discovery results for OrthoRF, RF, and two non-neural baselines, k-means and a histogram-based method, on clean and noisy SEM datasets. For k-means, we use scikit-learn's implementation to cluster pixel intensities. The histogram baseline uses mode/peak assignment (nearest-peak clustering); in the intensity histogram, we detect prominent peaks, and assign each pixel to the closest peak. These non-neural baselines were chosen because they are unsupervised, fast, and interpretable. They also show how much segmentation is possible from intensity alone on uniform-color SEM shapes, without heavy architectures, or additional training. Both OrthoRF and RF use $n = 5$ (4 layers + background) dimensionality in the orientation space.

[3] https://github.com/loeweX/RotatingFeatures/tree/main

Table 2 shows that OrthoRF outperforms RF on both the noise-free (ARI-BG: 0.9908 vs 0.9551) and the noisy test set (ARI-BG: 0.8356 vs 0.7043), indicating greater robustness under heavy occlusion and noise. Compared with k-means and the histogram baseline, the neural models excel especially in noisy settings, where edge blurring potentially hinders intensity-only clustering. We also evaluated out-of-distribution generalization for OrthoRF. Training on clean data and testing on noisy yields only a minor drop (ARI-BG from 0.9908 to 0.9836), indicating strong noise tolerance. In contrast, training on noisy data and testing on clean degrades more (from 0.8356 to 0.7610), likely because noise-trained models learn smoothed boundaries that underfit sharp, clean edges. Fig. 4 shows qualitative results on noise-free SEM test images. RF (top) distributes content across orientation components, whereas OrthoRF (bottom) separates SEM layers and recovers occluded structures.

Quantitative evaluation of similarity We quantify how the orthogonality constraint shapes the encoder output z_{out} by averaging pairwise cosine angles across all orientation components (Table 3). On 4Shapes, OrthoRF yields angles near 90° with far lower variability than RF (std 4.39 vs. 13.91), indicating cleaner phase separation. On SEM, OrthoRF averages 80° (softer constraint due to lower λ), while RF attains a higher mean (82.12°) but with greater variance. Across datasets, the background dimension (Dim. 5 in 4Shapes; Dim. 4 in SEM) shows the smallest angles, reflecting weaker distinctiveness but is included in all statistics.

Quantitative evaluation of separability We also quantify the orthogonality constraint in the model's output z_{final} using inter-/intra-cluster angular metrics, following Stanić et al. (2023), to gain further insight. The inter-cluster metric measures how well different objects are separated in feature space. For each image, we compute unit-normalized centroids for all objects and evaluate the pairwise angles between centroid directions. A large inter-cluster angle indicates that objects are embedded in distinct directions, while a small angle implies potential overlap. In contrast, the intra-cluster metric quantifies cluster compactness. For each object, we compute the angular deviation of each pixel feature from its centroid direction, and use the mean and standard deviation of these deviations as indicators of how tightly the object's embeddings cluster. We compute these distance metrics on a per-image basis before averaging over all samples in the 4Shapes dataset. The results in Table 4 show that OrthoRF achieves consistently lower intra-cluster dispersion compared to RF, indicating tighter object representations. While RF attains a higher mean inter-cluster angle, its variability is substantial (standard deviation ≈ 23.57), suggesting unstable separation across samples. In contrast, OrthoRF produces inter-cluster angles that are close to perpendicular with considerably lower fluctuations.

To aid interpretation, we visualize per-pixel embeddings from the models' outputs (z_{final}) by projecting them to 2D with PCA and normalizing them on the unit circle, then overlaying class centroids as arrows from the origin (see Fig. 5). As shown, OrthoRF forms tighter, better-separated clusters, while RF is noticeably more dispersed—matching the quantitative metrics in Table 4. The background cluster (label 0) is naturally more diffuse.

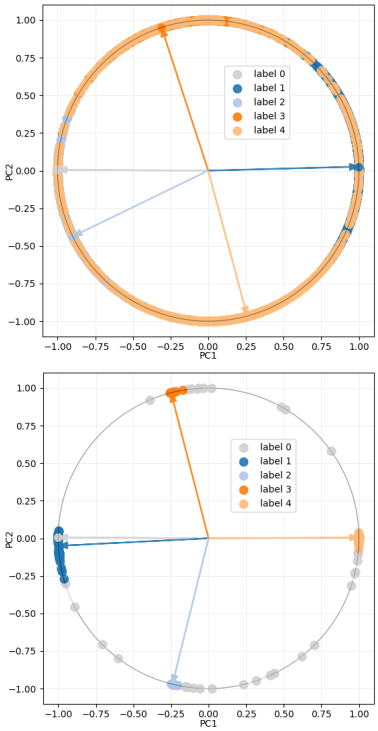

Figure 5: Principal component visualization of z_{out} on 4Shapes. OrthoRF (bottom image) embeddings form tighter clusters, whereas RF's (top image) are more dispersed. Background is always represented by label 0.

5 Related Work

Object-centric learning Over the years, a wide range of OCL models (Eslami et al., 2016; Greff et al., 2016; 2017; Jiang et al., 2019; Lin et al., 2020; Prabhudesai et al., 2022; Stelzner et al., 2019) have been proposed. MONet (Burgess et al., 2019),

IODINE (Greff et al., 2019), and GENESIS (Engelcke et al., 2019; 2021) propose unsupervised approaches to disentangle scenes into objects, with MONet and IODINE modeling objects independently and GENESIS capturing their interactions. Slot Attention (Locatello et al., 2020) extends this line of work by introducing an iterative attention mechanism where slots compete to bind to distinct objects. This design has inspired numerous extensions, including SLATE (Singh et al., 2021) and DINOSAUR (Seitzer et al., 2022), which integrate Transformer-based encoders and decoders (Vaswani et al., 2017) to better handle real-world images. OCL has been extended to both 3D (Chen et al., 2021; Sajjadi et al., 2022; Stelzner et al., 2021) and video data (Lai et al., 2021; Elsayed et al., 2022; Singh et al., 2022). ROOTS (Chen et al., 2021) disentangles objects via 3D-to-2D multi-view projections, while SAVi (Lai et al., 2021), SAVi++ (Elsayed et al., 2022), and STEVE (Singh et al., 2022) extend Slot Attention to videos, leveraging temporal dynamics to separate objects from each other and the background.

Synchrony-based learning Recent OCL work is shifting beyond slot-based models toward synchrony-based approaches (Löwe et al., 2022; Reichert & Serre, 2013; Löwe et al., 2024a; Stanić et al., 2024), inspired by neural binding-by-synchrony in the brain. Early studies explored supervised (Mozer et al., 1991) and weakly supervised (Ravishankar Rao & Cecchi, 2010) settings, while recent advances focus on unsupervised learning (Löwe et al., 2022; Reichert & Serre, 2013). CAE (Löwe et al., 2022) is an unsupervised model with complex-valued activations that discovers objects via phase clustering. CtCAE (Stanić et al., 2024) extends CAE with a contrastive loss to sharpen separability. RF (Löwe et al., 2024a) lifts phase representations to higher-dimensional rotations, improving object separation and expressivity beyond the 2D complex plane. In the approaches above, synchrony emerges from task objectives and network dynamics. By contrast, some works (Miyato et al., 2024; Muzellec et al., 2025) introduce an explicit synchronizer (e.g., a Kuramoto oscillator system) to create phase synchrony for object categorization.

Orthogonality Orthogonality has been widely exploited in deep learning. It has been used in network initialization (Saxe et al., 2013) and during training (Achour et al., 2022; Li et al., 2019) to improve stability and generalization. Orthogonality also has been used to create discriminative feature representations (Lezama et al., 2018; Ranasinghe et al., 2021) and disentangle features (Wang et al., 2018). In open-world object detection (Sun et al., 2024), a study utilized multiple levels of orthogonality throughout the training process to mitigate catastrophic interference and facilitate incremental learning of previously unseen objects.

6 Conclusion

Summary In this paper, we introduce the OrthoRF autoencoder to address a central RF (Löwe et al., 2024a) limitation: distributed object-centric representations break down in overlaps, where features from different objects become uncertain and undermine phase-space clustering. OrthoRF couples competitive binding with an inner-product orthogonality loss to align each object to a distinct phase axis, yielding sharper alignment and removing the clustering step. Across unsupervised object discovery OrthoRF matches or surpasses relevant models and recovers occluded parts. More broadly, orthogonality provides an effective inductive bias that regularizes distributed representations into discrete, directly resolving overlap ambiguity and improving the downstream usability.

Limitations and future work OrthoRF converges more slowly than RF ($\approx$200k vs. $\approx$100k steps), reflecting the stricter objective of enforcing orthogonal, object-aligned axes rather than distributed codes. Rarely, training can get stuck in suboptimal states (incomplete separation or multi-object collapse onto one phase axis), a phenomenon also noted in slot-based setups (Locatello et al., 2020). Increasing the orthogonality weight λ or lowering the learning rate stabilizes training. For future work, a natural extension is to integrate synchrony-based binding into attention layers. We also aim to address RF's RGB/color-channel degradation and evaluate OrthoRF on more realistic datasets (rich textures, challenging backgrounds) and in video, leveraging temporal synchrony for stability. Finally, the "lifting" from scalar to vector-valued features adds compute; more parameter-efficient encodings are a promising direction.

REFERENCES

El Mehdi Achour, François Malgouyres, and Franck Mamalet. Existence, stability and scalability of orthogonal convolutional neural networks. *Journal of Machine Learning Research*, 23(347): 1–56, 2022.

Titas Anciukevicius, Christoph H Lampert, and Paul Henderson. Object-centric image generation with factored depths, locations, and appearances. *arXiv preprint arXiv:2004.00642*, 2020.

Victor Bapst, Alvaro Sanchez-Gonzalez, Carl Doersch, Kimberly Stachenfeld, Pushmeet Kohli, Peter Battaglia, and Jessica Hamrick. Structured agents for physical construction. In *International conference on machine learning*, pp. 464–474. PMLR, 2019.

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.

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 the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chang Chen, Fei Deng, and Sungjin Ahn. Roots: Object-centric representation and rendering of 3d scenes. *Journal of Machine Learning Research*, 22(259):1–36, 2021.

Di Chen, Shanshan Zhang, Jian Yang, and Bernt Schiele. Norm-aware embedding for efficient person search. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12615–12624, 2020.

JA Coakley. Reflectance and albedo, surface. *Encyclopedia of atmospheric sciences*, 12, 2003.

David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. Attention over learned object embeddings enables complex visual reasoning. *Advances in neural information processing systems*, 34:9112–9124, 2021.

Andrea Dittadi. On the generalization of learned structured representations. *arXiv preprint arXiv:2304.13001*, 2023.

Gamaleldin 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. *Advances in Neural Information Processing Systems*, 35:28940–28954, 2022.

Martin Engelcke, Adam R Kosiorek, Oiwi Parker Jones, and Ingmar Posner. Genesis: Generative scene inference and sampling with object-centric latent representations. *arXiv preprint arXiv:1907.13052*, 2019.

Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. Genesis-v2: Inferring unordered object representations without iterative refinement. *Advances in Neural Information Processing Systems*, 34:8085–8094, 2021.

SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. *Advances in neural information processing systems*, 29, 2016.

Anand Gopalakrishnan, Aleksandar Stanić, Jürgen Schmidhuber, and Michael C Mozer. Recurrent complex-weighted autoencoders for unsupervised object discovery. *Advances in Neural Information Processing Systems*, 37:140787–140811, 2024.

Klaus Greff, Antti Rasmus, Mathias Berglund, Tele Hao, Harri Valpola, and Jürgen Schmidhuber. Tagger: Deep unsupervised perceptual grouping. *Advances in Neural Information Processing Systems*, 29, 2016.

Klaus Greff, Sjoerd Van Steenkiste, and Jürgen Schmidhuber. Neural expectation maximization. *Advances in neural information processing systems*, 30, 2017.

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 *International conference on machine learning*, pp. 2424–2433. PMLR, 2019.

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

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

Jindong Jiang, Sepehr Janghorbani, Gerard De Melo, and Sungjin Ahn. Scalor: Generative world models with scalable object representations. *arXiv preprint arXiv:1910.02384*, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. international conference on learning representations (2015). *San Diego, California*, 2015.

Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1205–1214, 2021.

José Lezama, Qiang Qiu, Pablo Musé, and Guillermo Sapiro. Ole: Orthogonal low-rank embedding-a plug and play geometric loss for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8109–8118, 2018.

Shuai Li, Kui Jia, Yuxin Wen, Tongliang Liu, and Dacheng Tao. Orthogonal deep neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1352–1368, 2019.

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. *arXiv preprint arXiv:2001.02407*, 2020.

Weiyang Liu, Zhen Liu, Zhiding Yu, Bo Dai, Rongmei Lin, Yisen Wang, James M Rehg, and Le Song. Decoupled networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2771–2779, 2018.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

Sindy Löwe, Phillip Lippe, Maja Rudolph, and Max Welling. Complex-valued autoencoders for object discovery. *arXiv preprint arXiv:2204.02075*, 2022.

Sindy Löwe, Phillip Lippe, Francesco Locatello, and Max Welling. Rotating features for object discovery. *Advances in Neural Information Processing Systems*, 36:59606–59635, 2023.

Sindy Löwe, Phillip Lippe, Francesco Locatello, and Max Welling. Rotating features for object discovery. *Advances in Neural Information Processing Systems*, 36, 2024a.

Sindy Löwe, Francesco Locatello, and Max Welling. Binding dynamics in rotating features. *arXiv preprint arXiv:2402.05627*, 2024b.

Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 6169–6176. IEEE, 2021.

Takeru Miyato, Sindy Löwe, Andreas Geiger, and Max Welling. Artificial kuramoto oscillatory neurons. *arXiv preprint arXiv:2410.13821*, 2024.

Michael C Mozer, Richard Zemel, and Marlene Behrmann. Learning to segment images using dynamic feature binding. *Advances in Neural Information Processing Systems*, 4, 1991.

Sabine Muzellec, Andrea Alamia, Thomas Serre, and Rufin VanRullen. Enhancing deep neural networks through complex-valued representations and kuramoto synchronization dynamics. *arXiv preprint arXiv:2502.21077*, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):128–140, 2016.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Generating fast and slow: Scene decomposition via reconstruction. *arXiv preprint*, 2022.

Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, and Fahad Shahbaz Khan. Orthogonal projection loss. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12333–12343, 2021.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

A Ravishankar Rao and Guillermo A Cecchi. An objective function utilizing complex sparsity for efficient segmentation in multi-layer oscillatory networks. *International Journal of Intelligent Computing and Cybernetics*, 3(2):173–206, 2010.

David P Reichert and Thomas Serre. Neuronal synchrony in complex-valued deep networks. *arXiv preprint arXiv:1312.6115*, 2013.

Adina L Roskies. The binding problem. *Neuron*, 24(1):7–9, 1999.

Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd Van Steenkiste, Filip Pavetic, Mario Lucic, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. *Advances in neural information processing systems*, 35:9512–9524, 2022.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*, 2022.

Wolf Singer. Binding by synchrony. *Scholarpedia*, 2(12):1657, 2007.

Gautam Singh, Fei Deng, and Sungjin Ahn. Illiterate dall-e learns to compose. *arXiv preprint arXiv:2110.11405*, 2021.

Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos. *Advances in Neural Information Processing Systems*, 35:18181–18196, 2022.

Elizabeth S Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990.

Aleksandar Stanić, Anand Gopalakrishnan, Kazuki Irie, and Jürgen Schmidhuber. Contrastive training of complex-valued autoencoders for object discovery. *Advances in Neural Information Processing Systems*, 36:11075–11101, 2023.

Aleksandar Stanić, Anand Gopalakrishnan, Kazuki Irie, and Jürgen Schmidhuber. stanic2024contrastive training of complex-valued autoencoders for object discovery. *Advances in Neural Information Processing Systems*, 36, 2024.

Karl Stelzner, Robert Peharz, and Kristian Kersting. Faster attend-infer-repeat with tractable probabilistic models. In *International Conference on Machine Learning*, pp. 5966–5975. PMLR, 2019.

Karl Stelzner, Kristian Kersting, and Adam R Kosiorek. Decomposing 3d scenes into objects via unsupervised volume segmentation. *arXiv preprint arXiv:2104.01148*, 2021.

Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE international conference on computer vision*, pp. 3800–3808, 2017.

Zhicheng Sun, Jinghan Li, and Yadong Mu. Exploring orthogonality in open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17302–17312, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Yitong Wang, Dihong Gong, Zheng Zhou, Xing Ji, Hao Wang, Zhifeng Li, Wei Liu, and Tong Zhang. Orthogonal deep features decomposition for age-invariant face recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 738–753, 2018.

Yangmuzi Zhang, Zhuolin Jiang, and Larry S Davis. Learning structured low-rank representations for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 676–683, 2013.