CoCAE: Contrastive Training of Complex-valued Autoencoders for Object Discovery in High-resolution RGB Images

Sam Titarsolej
University of Amsterdam & Osaka University
s.titarsolej@gmail.com

Abstract

Complex-Valued Autoencoders (CAEs) have shown fascinating results in the task of unsupervised object discovery. By binding learned features that have similar activations, and unbinding learned features that have dissimilar activations using complex-valued arithmetic, CAEs manage to cluster pixels in an image that belongs to the same object in an unsupervised manner. However, CAEs are limited to single-channel (grayscale) images. Moreover, CAEs only perform well on simplistic synthesized data, such as simple 2D shapes and MNIST digits. We extend CAEs to overcome both limitations by three novel layers to handle complex values, namely, complex-value variants of max-pooling, up-sampling, and channel pooling. We also introduce a contrastive training scheme to further improve pixel separability for object discovery. We empirically show that our method outperforms existing methods in object discovery on the CLEVR and Tetrominoes datasets.

1. Introduction

Object discovery in computer vision involves recognition and conceptual separation of individual objects or concepts in a scene without any prior knowledge of what these objects or concepts are. The resulting object-enteric scene decomposition is proven to be helpful in downstream tasks such as image classification, object detection, and semantic segmentation, providing representations that generalize into unseen environments [30, 41]. Furthermore, such scene decomposition may be compatible with human perception as humans are shown to perform such scene decomposition in order to reason about and interact with the environments [25, 31, 39]. Existing research, however, requires supervision by human annotations for visual reasoning tasks such as object detection [32] and semantic segmentation [19].

Among various approaches for object discovery, such as [5, 30], **C**omplex-Valued **A**uto**E**ncoders (CAEs) [31] is particularly interesting as they make use of complex values as a basis for representing images and latent states. The autoencoder output is an image, each of whose pixels is a complex value. The magnitude of the complex value represents the intensity (*i.e.*, a pixel value in a real-valued image), whereas the phase encodes the identity of each object. Intriguingly, this capability of object discovery is learned through the reconstruction loss without any dedicated annotation.

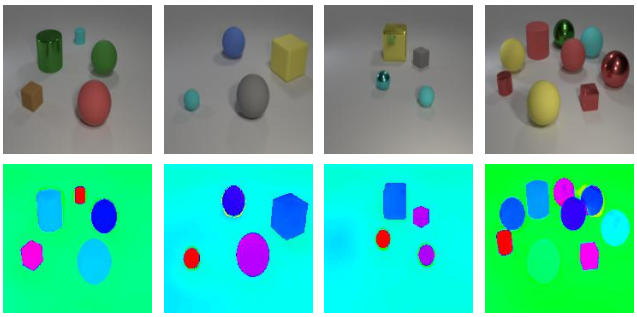

Figure 1. Qualitative results of CoCAE on the CLEVR [21] dataset with various numbers of objects. The top row represents the input image, and the bottom row represents the phase output (soft mask) of our method. Even with large numbers of objects, our method is able to capture the majority of the objects.

The major limitations of CAEs are that they only take single-channel (grayscale) and low-resolution images. These limitations may hinder CAEs from applications to real-world tasks; however, addressing these limitations, especially regarding the number of channels, is not trivial due to the use of complex values. Namely, the output for an input image with three (RGB) channels is again a three-channel image with complex-valued pixels. The channels of each pixel are supposed to encode object identity in their phases individually, which are not necessarily consistent with each other.

In an effort towards unsupervised visual reasoning, where the need for human annotations is reduced, we propose CoCAEs, **Co**ntrastive training for **C**omplex-Valued **A**uto**E**ncoders. We built it upon CAEs, addressing the

aforementioned limitations to make them work on high-resolution and multi-channel images. Our idea toward handling multi-channel images is straightforward: We newly introduce a complex channel-pooling layer to reduce phases in multiple channels. We also introduce complex max-pooling and up-sampling for high-resolution images to reduce or extend the spatial dimensionality.

We exhibit that our contrastive learning approach improves the separability of the output of CAE, for multiple objects in a single image. Where existing unsupervised representation learning methods [6, 7, 16, 18] only learn image-level representations, CoCAE enables representation extraction for specific image regions. Such learned representations could prove helpful in downstream tasks for such as image segmentation and object detection. Additionally, CoCAE could serve as a building block for explainable artificial intelligence (XAI) in computer vision applications. Previously, the object discovery capabilities of Slot Attention [30] have been applied for explainable image classification in SCOUTER [27] with minimal architecural modifiations.

The contributions proposed in this work are summarized as follows: (1) We introduce several complex-valued neural network layers to allow CAE [31] to produce meaningful phase values for multi-channel images. Moreover, through these introduced layers, our method is able to perform object discovery on a larger resolution compared to the original CAE architecture. (2) We introduce a contrastive learning scheme for training CAE on high-resolution RGB images, and demonstrate the object discovery performance of this method through experiments with the CLEVR [21] and the Tetrominoes [5] datasets. (3) We compare the object discovery performance of CoCAE with state-of-the-art object discovery techniques, and find that CoCAE outperforms Slot Attention [30] in terms of object discovery performance on these datasets due to the high resolution output masks of CoCAE.

2. Related Work

2.1. Object Discovery

The task of object discovery aims to identify and localize individual objects that compose a scene in an image, in an unsupervised manner. Many existing methods for object discovery [14, 30] involve the attention mechanism [1]. Slot attention [1] for example, performs image reconstruction through a discrete attention-based bottleneck, where the spatial information of different objects in an image is explicitly modeled in this bottleneck. These bottlenecks are then refined by iterating of the attention mechanism several times. Similarly, [5] introduce MoNET and [15] propose IODINE, which both employ a variational autoencoder [23] in for object discovery through image reconstruction. GENESIS [13] approaches object centric representation learning from a generative perspective, scenes are constructed in an iterative, component-wise manner. Several extenstions on these works have been introduced for object discovery in video [11, 22, 24, 35].

2.2. Complex-valued Neural Networks

Interest in applying complex-valued arithmetic in neural networks has grown over recent years. In [4] an overview of applications of complex-valued neural networks is provided, as well as different techniques to injecting copmlex activations into neural networks. Semantic segmentation has seen plentiful use of complex-valued neural networks in works such as [28]. More specifically, POLSAR imagery is of specific interest for many [3, 42], due to the periodic nature of POLSAR data.

2.3. Contrastive Learning

Contrastive learning is an unsupervised representation learning paradigm which aims to learn data representation where similar instances are grouped together in some representation space, and dissimilar instances are distant from each other. Recent approaches to contrastive learning in computer vision attempt to perform this separation in the feature space of image encoders [6, 7, 16, 18].

One of the main challenges of contrastive learning is the mining of positive and negative pairs. Positive pairs are used to learn which data instances to bring together, while negative pairs are used to learn which data instances to separate in the learned representation space. Many methods [7, 18] use different views and random augmentations of the same image to obtain positive pairs, and use a memory bank where previous views and augmentations are stored to obtain negative pairs. Our contrastive learning approach for CAE follows a similar positive-negative pair mining pattern using view augmentations and a memory bank.

More recent methods [6, 9, 16] overcome this pair-mining challenge by only using positive pairs in a student-teacher setup. In such setups a student model is optimized to mimic the representation of a teacher model, and the teacher model is built from previous versions of the student model based on a momentum update rule. These methods require heavy augmentations of the input images, to add noise to the optimization objective.

2.4. Autoencoders in Self Supervised Learning

Autoencoders are optimized to reconstruct their input, while being forced to do so through a information bottleneck. The purpose of the bottleneck is to reduce to dimensionality of the information contianed in the image. Traditionally however, the learned representations in the bottleneck of autoencoders often fail to capture scene level information, and instead focus on low level information such as

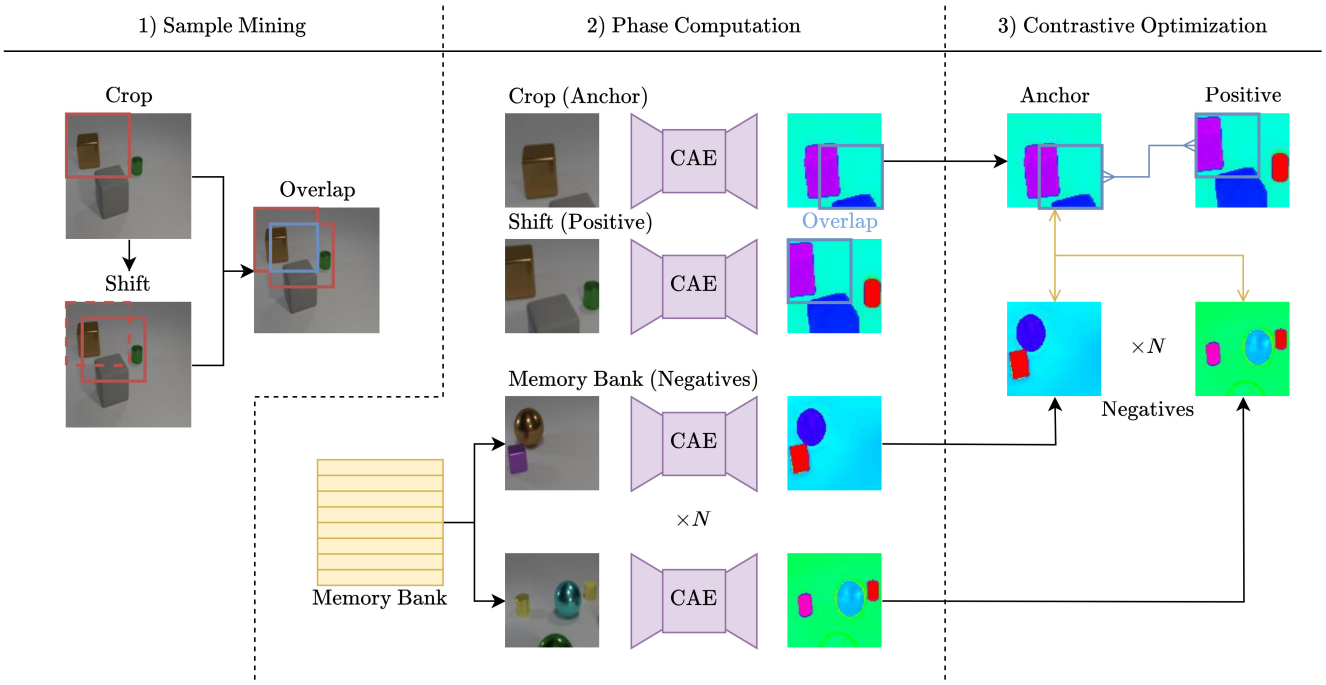

Figure 2. Contrastive learning scheme overview. In the first step, samples are mined by creating a random crop for each image in the training set, and randomly shifting these crops. The points of overlap between the original crop and the shifted crop is extracted and stored. The original crop is added to a memory bank. Secondly, the original crop, the shifted crop and random samples from the memory bank are forwarded through the CAE network. The resulting phase outputs are used in the third step, where the InfoNCE loss (equation 6) is employed to minimize the cosine distance (equation 7) between the phase outputs for the original crop and the shifted crop, while maximizing the distance between the phase outputs for the original crop and the random samples from the memory bank.

local textures. This is a result of the optimization objective of autoencoders, where small deviations in pixel distribution in the input reconstruction can cause large penalties.

In an effort to overcome the challenges autoencoders face in construction meaningful latent representations, Masked Image Modeling (MIM) [8, 17, 38, 40] focuses on reconstructing only patches in vision transformers [10, 37]. In these approaches, a portion of the input of the vision transformer is masked, while the vision transformer is optimized to reconstruct these masked patches. In some approaches an additional contrastive loss is applied to the image encoding [2], to further improve the separability of the learned representations.

3. CoCAE Method

Our approach to object discovery leverages the phase output of Complex-valued Autoencoders [31]. We introduce several architectural modifications to the complex-valued network, and a novel contrastive learning scheme to improve the object discovery capabilities of CAE in the following sections.

3.1. CAE Overview

In [31] the Complex-valued Autoencoder (CAE) is introduced for learning distributed object-centric representations. CAE uses complex-valued activations instead of real-valued activations, and leverages complex-valued arithmetic to separate pixels in an image. Complex values consist of a real part and an imaginary part, which together exist in the complex plane. A complex value can also be seperated into a magnitude, representing the norm in the complex plane, and a phase, representing the angle in the complex plane. In CAE, magnitudes represent the pressence of a learned feature and the phases should represent which learned features are bound together in a single image.

To achieve this phase encoding, [31] describe three mechanisms: synchronization, desynchronization and gating. Synchronization and desynchronization occur naturally in complex-valued neural networks due to constructive and destructive interference in complex-valued arithmetic. To provide the complex-valued neural network with precise control over phase-shifts, CAE separately applies the weights of each network layer to both the imaginary and real parts of the layer inputs:

$$\boldsymbol{\psi} = f_{\mathbf{w}}(\mathrm{Re}(\mathbf{z})) + f_{\mathbf{w}}(\mathrm{Im}(\mathbf{z})) \cdot i \in \mathbb{C}^{d_{\text{out}}} \tag{1}$$

(equation 2 in [31]). After obtaining intermediate value $\boldsymbol{\psi}$, biases are applied to the magnitudes and phases of $\boldsymbol{\psi}$ (equation 3 in [31]). Finally, a gating mechanism is applied in each layer in CAE that reduces the influence of out-of-phase inputs:

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

$$\mathbf{m}_{\mathbf{z}} = \frac{1}{2}\mathbf{m}_{\boldsymbol{\psi}} + \frac{1}{2}\boldsymbol{\chi} \in \mathbb{R}^{d_{\text{out}}} \tag{3}$$

(equation 4 in [31]), where $\mathbf{m}_{\boldsymbol{\psi}}$ is the magnitude of $\boldsymbol{\psi}$ after applying the magnitude bias. Furthermore, [31] describe that by applying the activation functions in each layer of the CAE to the magnitude of the layer output only, the model remains in full control over the phase-value outputs and thus is able to bind learned features:

$$\mathbf{z}_{\text{out}} = \mathrm{ReLU}(\mathrm{BatchNorm}(\mathbf{m}_{\mathbf{z}})) \circ \exp(i^{\boldsymbol{\varphi}_{\boldsymbol{\phi}}}) \in \mathbb{C}^{d_{\text{out}}}, \tag{4}$$

(equation 5 in [31]) in which $\boldsymbol{\varphi}_{\boldsymbol{\phi}}$ is the phase value of $\boldsymbol{\psi}$ after applying the phase bias.

3.2. CAE Architectural Modifications

To extend the ideas introduced in [31], we propose three novel complex-valued neural network layers. **Complex MaxPooling** is implemented to reduce the spatial dimensionality of high-resolution input images in the encoder network of CAE, and performs a similar operation to plain MaxPooling. In the case of Complex MaxPooling however, the max-indices are computed on the magnitude input of the module, and the pooling operation is applied to the magnitude and phase inputs seperately. Consequently, **Complex UpSampling** is applied to increase the spatial dimensionality in the decoder network. Bilinear upsampling is applied to the magnitude and phase inputs of the module seperately. Finally, **Complex ChannelPooling** reduces the channel dimension of a complex valued input from n to 1 by applying a magnitude-based weighted average over the phase-values in each of its input channels:

$$\phi'_{ij} = \sum_{c=1}^{C} \left[m_{ijc}\phi_{ijc}\right] \sum_{c'=1}^{C} \left[\frac{1}{m_{ijc'}}\right], \tag{5}$$

where i, j represent the spatial domain of an image, C is the number of input channels, and m, ϕ represent the input magnitude and phase values respectively. Through this setup, the network is forced to align phase values between channels, but still has the magnitude-channels required for RGB image encoding and decoding. Furthermore, the 1×1 convolutional layer, denoted as f_{out} in [31] is removed.

3.3. Contrastive Learning for CAE

Our contrastive learning approach for training CAE uses a view-shift approach to obtain anchors and positive samples, and a memory bank is employed to obtain negative samples. The CAE output phase-values are optimized directly. Each image in the training dataset is randomly cropped to obtain an anchor crop. This crop is then randomly shifted, and the overlapping area between the anchor crop and the resulting crop-shift is used as a positive sample. All random crops obtained during training are stored in a memory bank. This memory bank is then used to obtain random negative samples. A global overview of our training method is provided in figure 2.

Using these anchors, positive samples and negative samples, an adaptation of the InfoNCE [36] is optmized:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \frac{\exp(f(\mathbf{p},\mathbf{c})/\tau)}{\sum_{\mathbf{p}'\in\mathbf{P}} \exp(f(\mathbf{p}',\mathbf{c})/\tau)}\right], \tag{6}$$

where $\mathbf{c} \in [0, 2\pi]^{hw}$ is the phase output for the anchor image, $\mathbf{p} \in [0, 2\pi]^{hw}$ is the phase output for the positive sample, and $\mathbf{P}$ is the set of phase outputs for randomly sampled crops from the memory bank with the addition of $\mathbf{p}$. To add significant noise to the training process, a number of pixels is sampled for optimization according to a uniform distribution with $p < 0.005$. Furthermore, τ is the temperature hyperparameter. The score function $f(\mathbf{p}, \mathbf{c})$ is defined to be the cosine distance between $\mathbf{p}$ and $\mathbf{c}$:

$$f(\mathbf{p}, \mathbf{c}) = \frac{1}{K} \sum_{k\in\mathbf{K}} -\cos(|p_k - c_k|), \tag{7}$$

where $\mathbf{K}$ is the set of pixels sampled for contrastive optimization. This contrastive loss function exists as an auxiliary loss objective, besides to the reconstruction loss which is defined as:

$$\mathcal{L}_{\text{MSE}} = \|\hat{\mathbf{y}} - \mathbf{x}\|_2^2. \tag{8}$$

Combining both the reconstruction loss and the auxiliary contrastive loss, we arrive at our combined loss objective for CoCAE:

$$\mathcal{L}_{\text{CoCAE}} = \mathcal{L}_{\text{MSE}} + \alpha\mathcal{L}_{\text{InfoNCE}}, \tag{9}$$

where α is a scalar hyperparameter used to scale the influence of the contrastive loss function on the optimization process. Finally, further noise is injected into the contrastive optimization scheme, through random augmentations to the positive and negative image samples. These augmentations include Gaussian blur and color jitter.

Dataset	Model	MSE ↓	ARI-FG ↑	ARI-Full ↑	Inter Cluster ↑	Intra Cluster ↓
CLEVR6	Autoencoder	4.539e-5	-	-	-	-
	CAE [31]	4.302e-5	0.00	0.03	0.04	0.98
	CAE_{RGB}	4.251e-5	0.31	0.53	0.19	0.66
	Slot Attention [30]	**4.197e-5**	0.82	0.84	-	-
	CoCAE (ours)	4.674e-5	**0.84**	**0.87**	**0.83**	**0.18**
CLEVR8	Autoencoder	**1.143e-5**	-	-	-	
	CAE [31]	1.147e-5	0.00	0.03	0.02	0.99
	CAE_{RGB}	1.802e-5	0.28	0.49	0.18	0.65
	Slot Attention [30]	1.348e-5	0.74	0.76	-	-
	CoCAE (ours)	1.203e-5	**0.79**	**0.80**	**0.59**	**0.20**
Tetrominoes	Autoencoder	1.136e-5	-	-	-	-
	CAE [31]	1.457e-5	0.00	0.03	0.01	0.92
	CAE_{RGB}	1.024e-5	0.42	0.62	0.29	0.52
	Slot Attention [30]	1.096e-5	**0.98**	0.97	-	-
	CoCAE (ours)	**1.029e-5**	**0.98**	**0.99**	**0.73**	**0.02**

Table 1. Quantative results of CoCAE on different multi-object datasets. The gray rows convey results of our method. The maximum number of objects in the CLEVR4, CLEVR6 and CLEVR8 are four, six and eight respectively. Comparisons are made between a real-valued autoencoder forreconstruction quality, and a vanilla CAE [31] for phase-assignment comparisons. Between each method, the difference in reconstrution quality in terms of MSE is insifnificant. CoCAE clearly outperforms CAE [31] on all metrics and datasets. CoCAE also outperforms Slot Attention [30].

4. Experiments

Data. Our contrastive learning approach is evaluated on the CLEVR [21] dataset and on the Tetrominoes [22] dataset. Existing resarch on object discovery [12, 15, 30] evaluate their methods on the same datasets. Furthermore, these datasets provide a realistic benchmark for evaluation of CoCAE on different numbers of objects. For this purpose, five different CLEVR datasets with varying numbers of maximum objects are generated. The smallest number of maximum objects is four, the largest is eight. For each dataset the minimum number of objects is two. All images in the generated datasets are resized to be 224×224. Each generated training set consists of $50,000$ images and each generated evaluation set consists of $5,000$ images. The Tetrominoes dataset provided in the Multi-Object dataset [22] is used consisting of $1,000,000$ images. Each image is resized to be 224×224 in resolution using nearest neighbour interpolation. The first $900,000$ images are used for training, and the remaining $100,000$ images are used for evaluation.

Metrics. To evaluate the performance of CoCAE we report four quantitative results: image reconstruction quality in terms of MSE, object centric masking quality in terms of ARI-FG [20], and full masking quality in terms of ARI-Full [20]. ARI (Adjusted Random Index) quantifies the similarity between two sets of clusters - in the case of object masks these sets are the predicted mask and ground truth mask - while adjusting for random chance. ARI-FG only considers ground truth masks of foreground objects, while ARI-Full also considers the background ground truth mask.

Furthermore, we compute inter and intra cluster cosine distances for the retrieved clusters on the phase output of CoCAE, as an indication of the linear separability of these clusters. This distance is rescaled to be between 0 and 1. Both metrics are further normalized by the number of objects in an image. In the case of inter cluster distance this adjustment is to account for the maximum achievable phase distance between two objects in perfect separation. In the case of intra cluster distance, the normalization account for minimum required separation. Intuitively, these metric thus describes to which extent perfect separation is achieved in a value between 0 and 1. For inter cluster distance, cluster means are used, while for intra cluster distance, the minimum and maximum phase values within a cluster are used.

Models. The autoencoder model architectures used in our experiments follows the VGG-16 [34] architecture, similar to how the architecture is used in U-Net [33] segmentation models. We add a ComplexChannelPooling layer after the last convolutional layer in the encoder and decoder. Furthermore, we compare our training method with a real-valued autoencoder following the same architecture for image reconstruction quality. We also compare our training method with a vanilla CAE [31] for comparison of clustering performance. The ComplexMaxPooling layers are replaced by convolutional layers with stride 2, and the ComplexUpSampling layers are replaced by deconvolutional layers, such that comparisons on the same 224×224 image resolution can be made. For further comparison to object discovery methods, we compare our method with

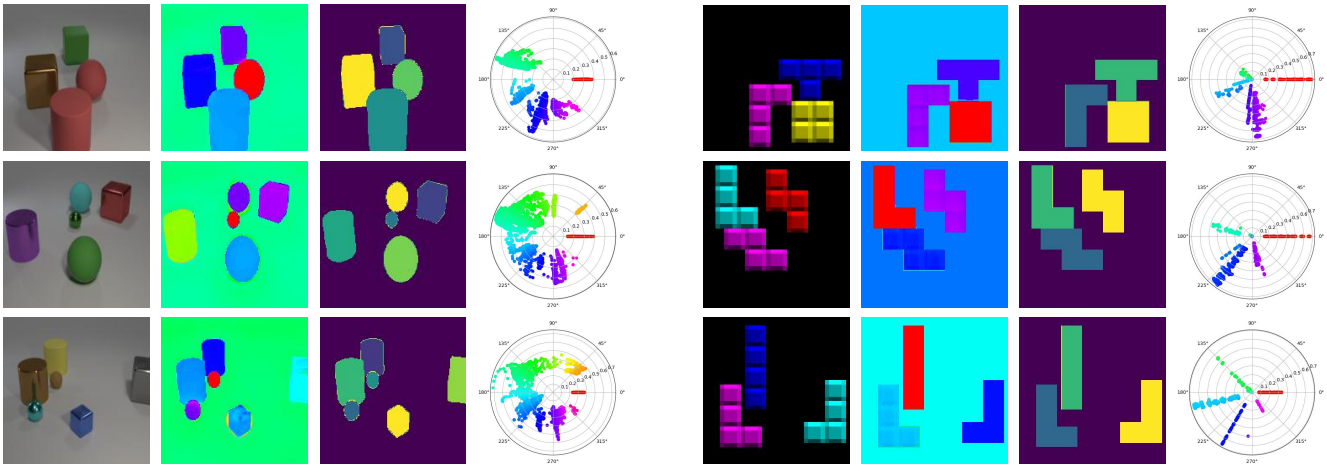

Figure 3. Qualitative results of CoCAE on the CLEVR (left) [21] and Tetrominoes (right) [5] datasets. CoCAE manages to capture object instances in the phase output, after which clustering becomes trivial. Even similarly colored objects with that border in pixel space (two red objects in top left result for example) receive significantly distant phase assignments.

Slot Attention [30]. We adapt the Slot Attention model for image reconstruction on a 224×224 resolution by adding one convolutional layer to the encoder CNN and one deconvolutional layer to the decoder CNN described in the original paper. As a result, the object attention masks become 7×7 each. The number of slots used is equal to the maximum number of objects in each dataset minus one. Further exact details of all model parameters are shared in the supplementary material.

Phases to Masks. Following the phase clustering method described in [31], the output phases of CoCAE are projected onto a unit circle in euclidean space. After this projection, K-means is applied, where the number of clusters is equal to the number of objects in the ground truth labels plus one. For the vanilla CAE, this clustering step is applied on all three image channels. Note that this step is solely in place for evaluation purposes, and could be replaced with clustering methods that do not require a fixed number of target clusters in any downstream application, as also stated in [31].

Training Settings. Training CoCAE is split in two stages. First the network is trained for 100 epochs on image reconstruction only. During this stage a batch size of 64 and a learning rate of 0.0001 is used. The network is then trained to optimize according to our contrastive objective using a batch size of 16, a learning rate of 0.0001, a single positive sample and 100 negative samples. For scaling the InfoNCE [36] part of our loss function, $\alpha = 0.00001$. Results of the optimized model after reconstruction optimization, but before contrastive optimization provided in our main results and are marked as CAE_{RGB}.

Config.	MSE	ARI-Full	Inter Cluster ↑
Dec. Only	**6.008e-5**	0.79	0.68
Enc. & Dec.	6.186e-5	**0.87**	**0.83**
Every Block	1.913e-1	0.01	0.08

Table 2. Performance of CoCAE with different placements of the ComplexChannelPooling layer on the CLEVR6 dataset. The gray row indicates configuration used in main experiments.

5. Results

Table 1 provides an overview of CoCAE performance in comparison to a real-valued autoencoder, a vanilla CAE [31], CAE_{RGB} and Slot Attention [30] on two CLEVR [21] datasets containing different numbers of objects, and the Tetrominoes dataset. The reconstruction quality between all models varies insignificantly. The masking performance of CAE_{RGB}, CoCAE and Slot Attention are significantly better compared to the vanilla CAE however. Vanilla CAE fails to produce meaningful phase outputs compared altogether, as it is not able to handle RGB images. CAE_{RGB} produces more significant separations, but only manages to capture foreground-background separation. CoCAE consistently outperforms Slot Attention [30] in terms of ARI-FG and ARI-Full. This is largely attributed to the manner in which CoCAE handles masking resolution compared to Slot Attention [30]. The phase output of CoCAE is the same resolution as the input image, while the attention mask output of Slot Attention [30] is remarkably smaller than the input image. To still achieve masks of the same input resolution, upsampling is required. During this upsampling process, details are lost. Object boundaries for example are

obscured in the spatial domain of the image. This effect is less noticeable on smaller evaluation resolutions.

In table 2, the impact of the ComplexChannelPooling layer on the performance of CoCAE is exhibited. Applying channel pooling only to the decoder - which is the minimum required for RGB images - reduces the separability of the phase output of CAE. By applying channel pooling after each convolutional block in the model, image reconstruction fails. Optimally, channel pooling is applied after the last convolutional blocks of the encoder and the decoder.

Figure 3 provides insight into the decay of the performance our method as the maximum number of objects in the CLEVR [21] dataset increases. Beyond 7 objects, the performance rapidly decreases, as the maximum achievable angle distance between phase clusters decreases. At 7 objects (plus background), the maximum achievable angle distance between object clusters equal to $\frac{2\pi}{8} = \frac{1}{4}\pi$. Compared to vanilla CAE [31] however, CoCAE provides a very significant increase in the number of objects it is able to capture. Results of vanilla CAE [31] show that performance starts to reduce crucially beyond three objects, whereas CoCAE performance remains above 0.8 ARI-Full up to seven objects.

Qualtitative results of CoCAE in figures 1, 3 and 6 demonstrate how CoCAE captures objects in the phase output. The polar projections provide insight into how well phase clusters can be separated. For most images, CoCAE is able to maximize the distance between phase clusters, as also expressed in the distance metrics in table 1. Figure 6, shows the performance of CoCAE on larger numbers of objects. Even though CoCAE is not able to capture all objects, it manages to capture the majority of objects, and separate them as the phase distance between clusters is maximized.

Figure 4. Zero-shot domain transfer of CoCAE onto the CelebAHQ-Mask [26] dataset. The model was trained only on the CLEVR10 [21] dataset, but is able to separate regions of hair, eyebrows and mouths in completely unseen images from the CelebAHQ-Mask [26] dataset.

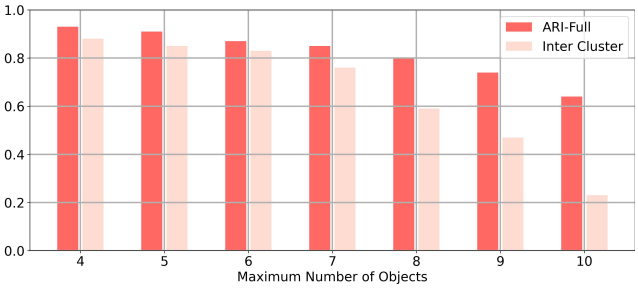

Figure 5. Decay of CoCAE performance on the CLEVR [21] in terms of ARI-Full as the maximum number of objects increases. The decay of ARI-Full is a result of less defined cluster boundaries, expressed in terms of the distance between phase clusters.

5.1. Domain Transfer

To gain insight into applications of CoCAE in natural images, we provide qualitative results of CoCAE on the CelebAHQ-Mask [26] dataset in figure 4. These results are obtained with a CoCAE model that was only trained on the CLEVR10 dataset. As the model is optimized for image reconstruction, there is room for zero-shot domain transfers. The model aims to reconstruct the input image, while matching shapes it has encountered in the training process. The results in figure 4 show that CoCAE is able to perform this domain transfer, and match several key regions in the faces depicted, such as hair, mouth and eyebrows. However, as the phase output of CoCAE on these images is very centered around 0π, clustering into meaningful masks, and thus qualitative evaluation is not possible on this dataset.

5.2. Limitations and Improvements

As the primary optimization objective of CoCAE remains image reconstruction through MSE Loss, the model looses representation capacity to unnecessary information such as precise light reflections in the CLEVR dataset [21] and shadows/highlights in the Tetrominoes dataset [5]. To avoid focus on such low-level details in natural images, [29] propose a frequency filtering technique for image reconstruction-based representation learning. These filters reduce the occurance of low-frequency image features, thus improving the focus on scene-level scene features. Such filtering could be applied to our approach to improve the object discovery quality of CoCAE.

The number of objects that CoCAE is able to capture is also limited. Performance of CoCAE significantly drops with an increased number of objects. As shown in figure 4, some objects are lost in the phase output of CoCAE, and some objects receive non-uniform phase assignments. This limitation is largely due to the shrinking of the available

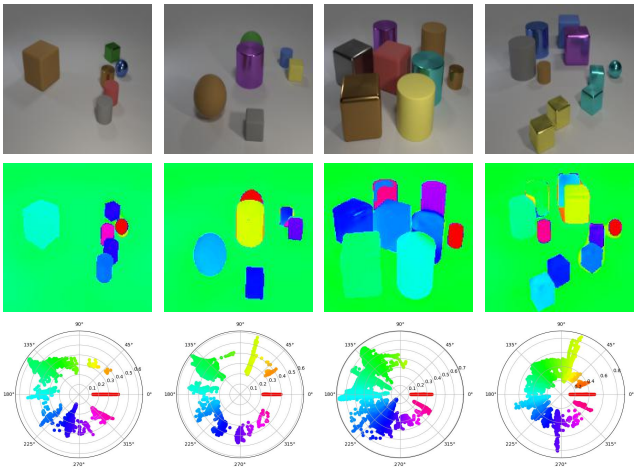

Figure 6. Qualitative results of CoCAE on the CLEVR10 [21] datasets with large number of objects. Altough CoCAE manages to capture many objects correctly, some objects are lost in the phase output of CoCAE, while others receive more than one phase assigment across the pixels than span the object.

space between pixel clusters in the phase output of CoCAE.

Finally, as illustrated by the results in figure 4, clustering phase outputs into mask is not feasible for CoCAE outputs on natural images such as the CelebAHQ-Mask dataset [26]. This currently limits the practical applicability of CoCAE. Further research into applications of contrastively trained complex-valued autoencoders in downstream tasks such as explainable computer vision, semantic segmentation and object detection is encouraged.

6. Conclusion

In this work we presented CoCAE, a contrastive learning approach for complex-valued autoencoders for object discovery. To achieve more scalability in terms of image channels and image resolution, we introduced several complex-valued neural network layers. We proposed a contrastive training scheme for the optimization of the phase output of the network, which uses augmentations for positive-sample mining, a memory bank for negative-sample mining. An adaptation of the InfoNCE [36] loss is derived for optimization of CAE using these positive and negative samples. We empirically show that CoCAE outperforms existing object discovery methods on the CLEVR [21] and Tetrominoes [5] datasets. Finally, we discuss the limitations of CoCAE and recommendations on how to further improve the object discovery performance of CoCAE, which currently include the number of objects that CoCAE is able to process, the reconstruction loss objective for CoCAE, and the lack of application of CoCAE onto downstream tasks.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016. 2

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: BERT pre-training of image transformers. *ICLR*, abs/2106.08254, 2021. 3

[3] J. A. Barrachina, C. Ren, G. Vieillard, C. Morisseau, and J.-P. Ovarlez. Real- and complex-valued neural networks for sar image segmentation through different polarimetric representations. pages 1456–1460, 2022. 2

[4] Joshua Bassey, Lijun Qian, and Xianfang Li. A survey of complex-valued neural networks, 2021. 2

[5] Christopher P. Burgess, Loïc Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matthew M. Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. abs/1901.11390, 2019. 1, 2, 6, 7, 8

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *ICCV*, abs/2104.14294, 2021. 2

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ICML*, abs/2002.05709, 2020. 2

[8] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *ICLR*, abs/2202.03026, 2022. 3

[9] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CVPR*, abs/2003.04297, 2020. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, abs/2010.11929, 2020. 3

[11] 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, 2022. 2

[12] Patrick Emami, Pan He, Sanjay Ranka, and Anand Rangarajan. Efficient iterative amortized inference for learning symmetric and disentangled multi-object representations. *ICML*, abs/2106.03630, 2021. 5

[13] Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner. GENESIS: generative scene inference and sampling with object-centric latent representations. *ICLR*, abs/1907.13052, 2019. 2

[14] S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, Koray Kavukcuoglu, and Geoffrey E. Hinton. Attend, infer, repeat: Fast scene understanding with generative models. *NeurIPS*, abs/1603.08575, 2016. 2

[15] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Chris Burgess, Daniel Zoran, Loïc Matthey, Matthew M. Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. *ICML*, abs/1903.00450, 2019. 2, 5

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *NeurIPS*, abs/2006.07733, 2020. 2

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CVPR*, abs/2111.06377, 2021. 3

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, abs/1911.05722, 2019. 2

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. 1

[20] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, Dec. 1985. 5

[21] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 1, 2, 5, 6, 7, 8

[22] Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loic Matthey, Antonia Creswell, Matt Botvinick, Alexander Lerchner, and Chris Burgess. Simone: View-invariant, temporally-abstracted object representations via unsupervised video decomposition. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *NeurIPS*, volume 34, pages 20146–20159. Curran Associates, Inc., 2021. 2, 5

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. 2

[24] 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. *ICLR*, abs/2111.12594, 2021. 2

[25] K Koffka. *Principles Of Gestalt Psychology*. Routledge, Oct. 2013. 1

[26] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, 2020. 7, 8

[27] Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. SCOUTER: slot attention-based classifier for explainable image recognition. *ICCV*, abs/2009.06138, 2020. 2

[28] Xinzhi Liu, Jun Yu, Toru Kurihara, Liangfeng Xu, Zhao Niu, and Shu Zhan. Hyperspectral imaging for green pepper segmentation using a complex-valued neural network. *Optik*, 265:169527, 2022. 2

[29] Yuan Liu, Songyang Zhang, Jiacheng Chen, Kai Chen, and Dahua Lin. Pixmim: Rethinking pixel reconstruction in masked image modeling, 2023. 7

[30] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *NeurIPS*, abs/2006.15055, 2020. 1, 2, 5, 6

[31] Sindy Löwe, Phillip Lippe, Maja Rudolph, and Max Welling. Complex-valued autoencoders for object discovery, 2022. 1, 2, 3, 4, 5, 6, 7

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016. 1

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *MICCAI*, abs/1505.04597, 2015. 5

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 5

[35] Gautam Singh, Yi-Fu Wu, and Sungjin Ahn. Simple unsupervised object-centric learning for complex and naturalistic videos, 2022. 2

[36] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *NeurIPS*, abs/1807.03748, 2018. 4, 6, 8

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, abs/1706.03762, 2017. 3

[38] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan L. Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. *CVPR*, abs/2112.09133, 2021. 3

[39] Max Wertheimer. Untersuchungen zur lehre von der gestalt. II. *Psychologische Forschung*, 4(1):301–350, 1923. 1

[40] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *CVPR*, abs/2111.09886, 2021. 3

[41] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. *ICLR*, abs/1910.01442, 2019. 1

[42] Zhimian Zhang, Haipeng Wang, Feng Xu, and Ya-Qiu Jin. Complex-valued convolutional neural network and its application in polarimetric sar image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):7177–7188, 2017. 2