IMPERIAL

Cluster Contrast for Unsupervised Visual Representation Learning

Nikolaos Giakoumoglou[1] nikos@imperial.ac.uk
Tania Stathaki[1] t.stathaki@imperial.ac.uk

1. Introduction

Self-supervised learning has emerged as a powerful paradigm for learning visual representations without manual annotations. While contrastive methods excel at distinguishing between different images, clustering approaches explore similarities across images.

Our Contribution: We introduce **Cluster Contrast (CueCo)**, a framework that combines the strengths of both approaches through a "push-pull" dynamic.

2. Methodology

Our method optimizes three objectives:

1. Contrastive Loss (InfoNCE):

$$\mathcal{L}_1 = -\log \frac{\exp(\mathbf{z}^\top \cdot \mathbf{z}'/\tau)}{\exp(\mathbf{z}^\top \cdot \mathbf{z}'/\tau) + \sum_{k=1}^{K} \exp(\mathbf{z}^\top \cdot \mathbf{z}_k/\tau)}$$

2. Centroid Contrastive Loss:

$$\mathcal{L}_2 = -\log \frac{\exp(\mathbf{z}^\top \cdot \mathbf{c}_{i[\mathbf{z}]}/\tau)}{\exp(\mathbf{z}^\top \cdot \mathbf{c}_{i[\mathbf{z}]}/\tau) + \sum_{l=1}^{L} \exp(\mathbf{z}^\top \cdot \mathbf{c}_l/\tau)}$$

3. Variance Loss:

$$\mathcal{L}_3 = \frac{\left\|\mathbf{z} - \mathbf{c}_{i[\mathbf{z}]}\right\|^2}{2 \cdot \sigma_{i[\mathbf{z}]}^2 + \epsilon}$$

Final Objective:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_1 + \lambda_2 \cdot \mathcal{L}_2 + \lambda_3 \cdot \mathcal{L}_3$$

Momentum Clustering

We introduce online clustering:

$$\mathbf{c}_i \leftarrow \beta_1 \cdot \mathbf{c}_i + (1-\beta_1) \cdot \left(\frac{1}{|S_i|} \sum_{\mathbf{z} \in S_i} \mathbf{z}' \right)$$

$$\sigma_i^2 \leftarrow \beta_2 \cdot \sigma_i^2 + (1-\beta_2) \cdot \left(\frac{1}{|S_i|} \sum_{\mathbf{z} \in S_i} (\mathbf{z}' - \mu_i)^2 \right)$$

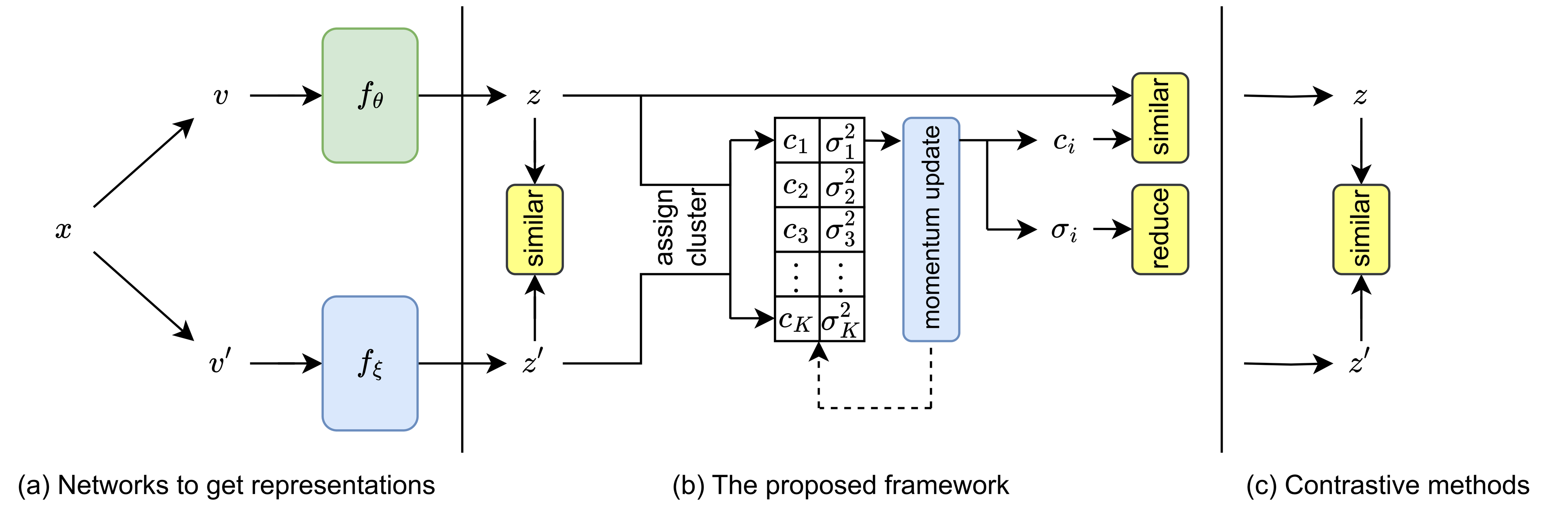

Figure 1: CueCo framework processes two augmented views through encoders, enforcing inter-class separation via contrastive loss while improving intra-class cohesion through clustering.

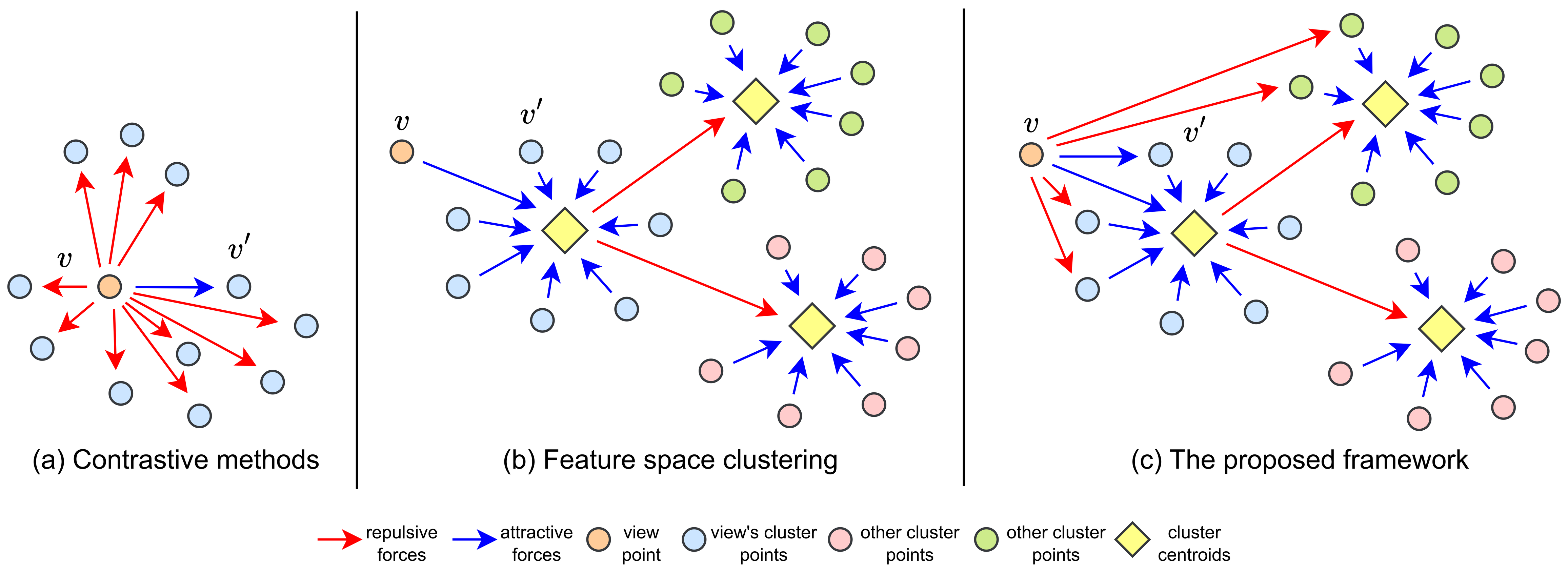

Figure 2: Visualization of push-pull dynamics: (a) Contrastive objective emphasizes class separation, (b) Clustering improves intra-class compactness, (c) CueCo combines both approaches.

3. Results

Linear Evaluation Performance

Table 1: Linear evaluation results (% accuracy) on benchmark datasets using ResNet-18.

Method	CIFAR-10	CIFAR-100	ImageNet-100
BYOL	92.61	70.18	80.09
MoCo-v2	92.94	69.54	78.20
MoCo-v3	93.10	68.83	80.86
VICReg	90.07	68.54	79.22
SwAV	89.17	64.67	74.28
CueCo (ours)	91.40	68.56	78.65

Clustering Performance

Table 2: Unsupervised clustering metrics on CIFAR datasets.

Method	CIFAR-10		CIFAR-100	
	NMI	ACC	NMI	ACC
MoCo-v2	60.96	63.51	51.77	31.72
SimCLR	69.03	74.50	50.75	32.17
CueCo	69.33	75.06	52.37	33.82

Ablation Study

Table 3: Impact of different loss components on CIFAR-100 performance.

$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	Top-1	NMI	ACC	ARI
✓			67.9	50.5	31.1	8.3
✓	✓		66.9	**54.6**	**34.5**	**14.9**
✓		✓	68.4	51.2	32.3	8.6
✓	✓	✓	**68.5**	52.3	33.8	11.3

Combining all three loss terms achieves the best balance between linear evaluation performance and clustering quality.

4. Conclusion

CueCo advances unsupervised visual representation learning by integrating contrastive learning with momentum clustering, creating a "push-pull" dynamic that simultaneously enhances inter-class separation and intra-class cohesion. The framework demonstrates competitive performance on benchmark datasets while particularly excelling in unsupervised image classification metrics. CueCo establishes a promising direction for self-supervised learning.

Affiliations

[1] Department of Electrical and Electronic Engineering, Imperial College London