Angular Constraint Embedding via SpherePair Loss for Constrained Clustering

Shaojie Zhang, Ke Chen

Department of Computer Science, The University of Manchester

Constraint Clustering

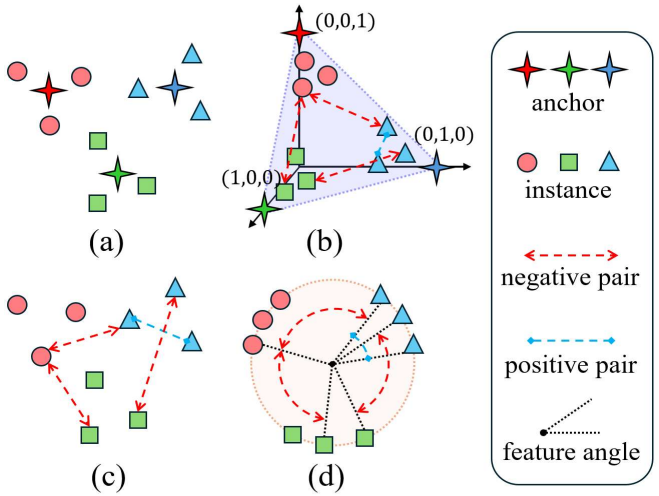
- Constrained clustering (CC) integrates weak supervision via pairwise constraints, significantly boosting clustering accuracy.
- Deep constrained clustering (DCC) surpasses traditional CC on complex, high-dimensional data and generalizes well to unseen instances.

Existing DCC

- **End-to-end DCC**
  Predefined anchors propagate errors;
  Local pairwise relations mismatch global assignments;
  Require the exact number of clusters.
- **Euclidean constraint embedding:**
  Anchor-free but relies on unbounded Euclidean distances, making pairwise separation unstable, leads to poorly distinguishable clusters.

Our method

- **Angular constraint embedding:**
  prevent conflicts, ensures equal and definitive inter-cluster distances without anchors, satisfying both positive and negative constraints



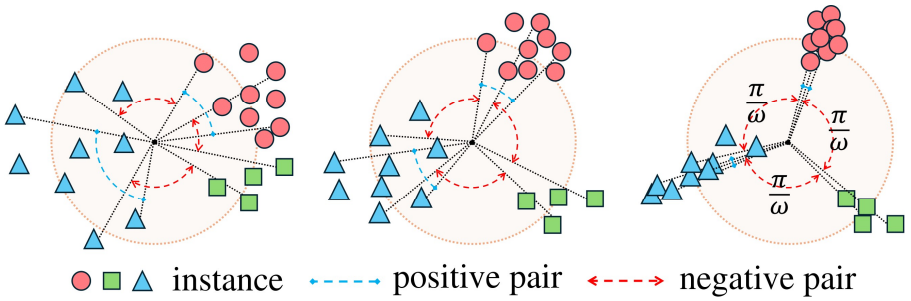
End-to-end DCC introduces anchors to transform features in (a) into soft cluster assignments in (b) for pairwise losses.

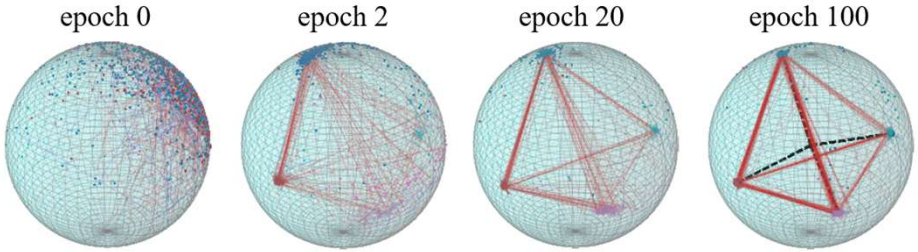
Deep constraint embedding in (c) focuses on the Euclidean distances.

Our **angular constraint embedding** in (d) operates in angular space.

SpherePair CC

- Given pairwise constraints $\mathcal{C} = \{(a_i, b_i, y_i)\}_{i=1}^{|\mathcal{C}|}$
- Embedding constrained pairs in **angular space**:
  Positive pairs ($y_i = 1$) → zero angles;
  Negative pairs ($y_i = 0$) → separated by *negative-zone* of size π/ω
- With reconstruction regularization, yields **clustering-friendly** SpherePair embedding $\mathcal{Z}_{\text{sphere}} \subset \mathbb{R}^D$
- Perform clustering (e.g., K-means) on $\mathcal{Z}_{\text{sphere}}$





Evolution of SpherePair embedding (D=3) on Reuters (K=4).

Theoretical foundation

Determine ω and D

- Conflict-free and equidistant embedding is theoretically attainable only when $D \geq K - 1$, and the lower bound of ω is $\pi / \arccos\big(-1/(K-1)\big)$.
- The optimal setting: $\omega = 2$ for sufficiently large $D \geq K$. This ensures conflict-free embedding for any $K > 1$ while providing maximal inter-cluster separation.

Geometric configuration

- Clusters on hypersphere form a K-vertex regular simplex in a $(K-1)$-dimensional subspace.
- Minor deviations from the ideal geometric configuration under near-zero residual loss

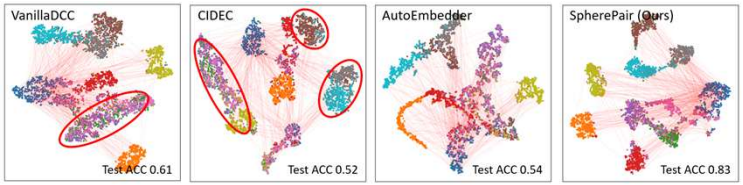
PCA-based K-inference

- The K-vertex regular simplex lies in a $(K-1)$-dimensional subspace. Thus, the intrinsic dimension $d^* = K - 1$.
- Rapidly infer d^* (and hence K) by tracking minimal inter-cluster angle invariance (in practice, we use ρ-fraction averaging for stability) across PCA dimensions.
- Avoids both retraining required by end-to-end DCC and cumbersome post-clustering validation.

Experiments

Comparative performance

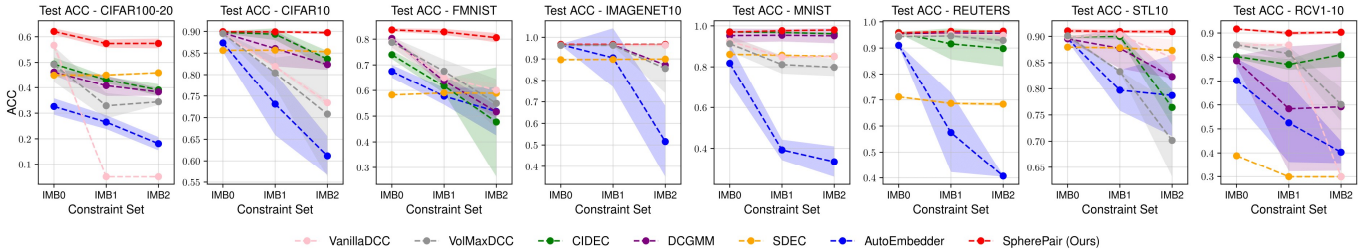
- Outperforms all 6 baselines across 8 datasets and 3 constraint levels, ranking 1st in >60/72 cases and 2nd in almost all others.
- Insensitive to the clustering methods.
- Robust even without pretraining.



t-SNE plots of FMNIST embeddings with imbalanced constraints

Imbalanced constraints

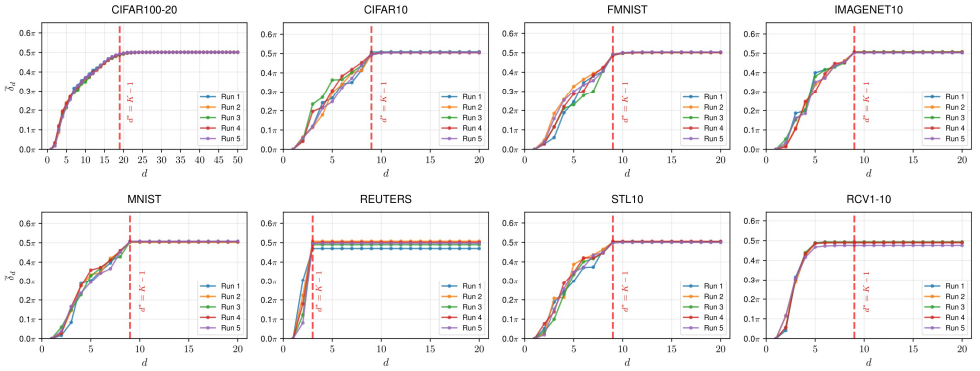
- Shows strong stability and consistently surpasses all baselines.
- Preserves coherent structures while forming discriminative clusters.



Test ACC performance of 7 models across 8 datasets under the balanced vs. imbalanced constraints

Unknown cluster number

- Rapid inference via a single closed-form PCA solution.
- Generally robust across datasets and constraint levels, with limited performance on strongly imbalanced RCV1-10.



Tracking minimal inter-cluster angle invariance across PCA dimensions for K-inference

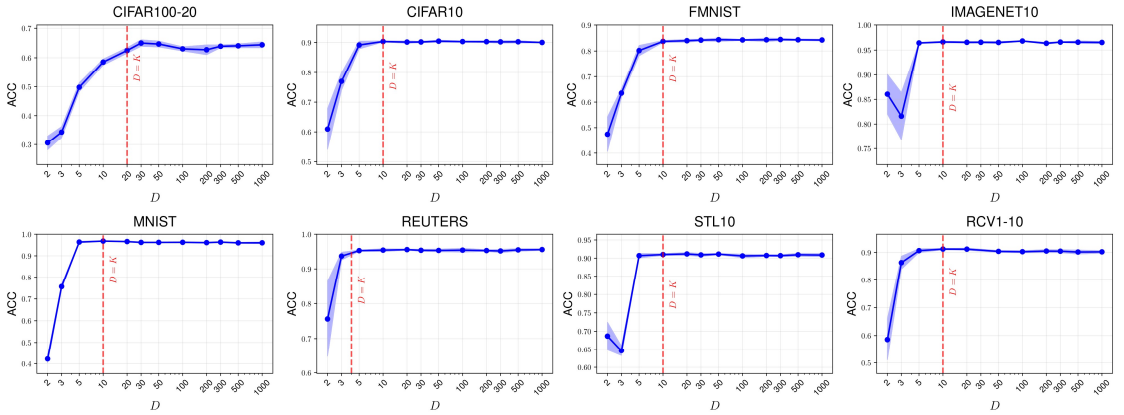
Empirical validation & hyperparameter sensitivity

Embedding dimension D:

- Stable for $D \geq K$, even up to $D = 1000$.
- Empirically validate our theoretical insights.
- Flexible under slightly sub-threshold D when K is unknown.

Regularization strength λ:

- Broadly robust in a wide range.
- Default λ= 0.02 recommended.



SpherePair test ACC vs. embedding dimension D across datasets

Tail ratio ρ in K-inference:

- Smaller ρ sharpens rises before $d^* = K - 1$, while larger ρ stabilizes the subsequent angle invariance
- $\rho \in [0.03, 0.1]$ yields clear and stable K estimation.

Other results

- Computationally efficient.
- Consistent performance across model structures.

	1k			5k			10k		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
VanillaDCC	34.3	36.3	19.3	47.4	47.1	32.2	54.5	50.3	37.6
VolMaxDCC	20.3	21.6	7.2	42.8	42.1	22.8	51.0	48.7	33.3
CIDEC (10D)	45.4	47.8	29.1	47.7	48.7	31.4	48.8	49.1	32.6
CIDEC (20D)	46.2	47.9	29.9	46.1	45.7	29.6	50.1	48.8	33.0
CIDEC (30D)	44.5	46.9	29.8	47.4	47.2	31.6	48.7	48.8	33.6
DCGMM (10D)	43.6	45.6	28.3	46.5	46.4	30.8	48.7	48.1	33.7
DCGMM (20D)	44.2	45.4	28.7	47.9	47.1	32.2	52.1	49.6	36.7
DCGMM (30D)	44.9	46.8	29.1	45.2	46.8	30.1	46.7	47.4	31.6
SDEC (10D)	44.9	47.2	28.2	45.6	48.5	29.3	45.6	48.8	30.1
SDEC (20D)	45.4	47.5	29.2	45.1	47.5	29.3	45.2	47.7	29.5
SDEC (30D)	44.3	46.3	28.5	44.7	46.8	29.2	44.7	47.2	29.2
AutoEmbedder (10D)	29.2	32.0	12.3	28.9	35.3	18.1	39.7	42.4	27.1
AutoEmbedder (20D)	21.6	23.4	7.1	14.2	13.8	4.7	31.3	36.9	20.4
AutoEmbedder (30D)	34.0	35.2	17.8	26.2	31.3	15.1	33.8	40.3	25.1
SpherePair (Ours) (10D)	46.5	46.3	30.9	54.2	49.8	37.5	57.5	51.6	41.1
SpherePair (Ours) (20D)	**48.2**	**48.0**	**32.4**	**58.8**	53.0	40.9	62.6	55.5	45.2
SpherePair (Ours) (30D)	**48.2**	47.2	31.5	58.4	**53.8**	**41.9**	**64.3**	**56.9**	**46.5**

Test performance (%) (ACC, NMI, ARI) on CIFAR-100-20 for models with varying embedding dimensions and constraint levels.

THANKS FOR WATCHING

Paper

Code