SHADE: Deep Density-based Clustering

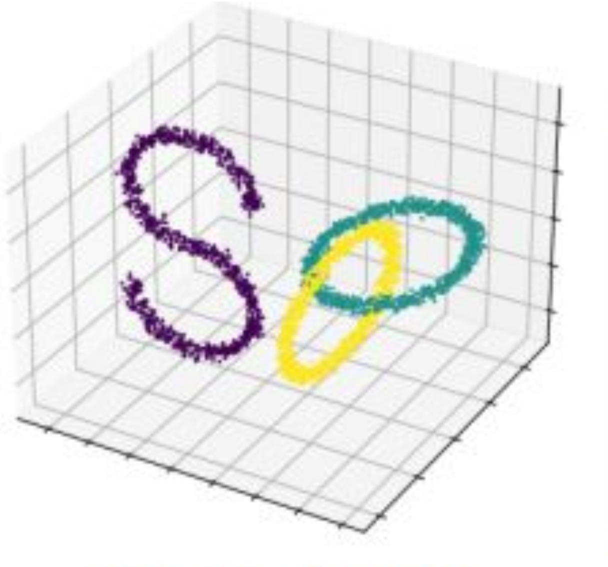
December 11th, 2024

Anna Beer*, Pascal Weber*,
Lukas Miklautz, Collin Leiber, Walid Durani, Christian Böhm, Claudia Plant

ICDM 2024, 9-12 December 2024, Abu Dhabi, UAE

Motivation

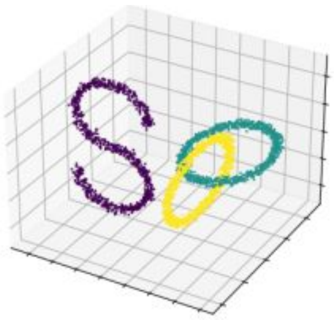
- **Density-based Clustering** is one of the main concepts of clustering
- Clusters of **arbitrary shape** are common in real world data
- Data can contain **noise points**



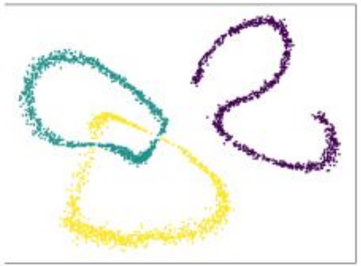
- Traditional methods are not optimal for **high-dimensional data**

Motivation

- Autoencoders (AE) optimizing the reconstruction loss are not optimal for **intertwined** or non-contractible clusters



3d data



2d embedding by an AE

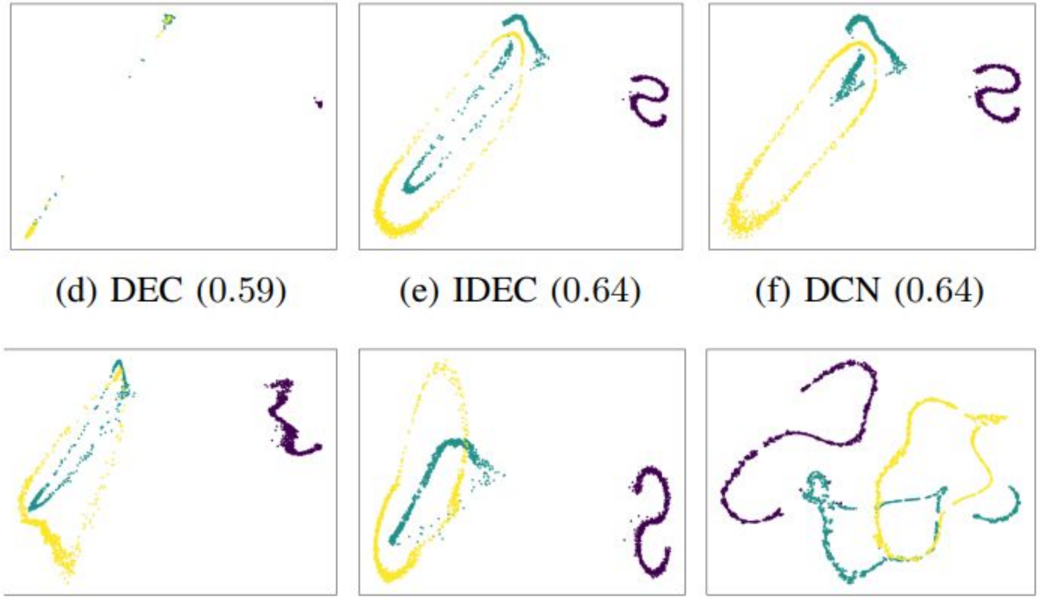
Desired embedding for clustering (SHADE)

Deep Clustering

- Deep Clustering methods usually combine an AE with some cluster loss

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{cluster}$$

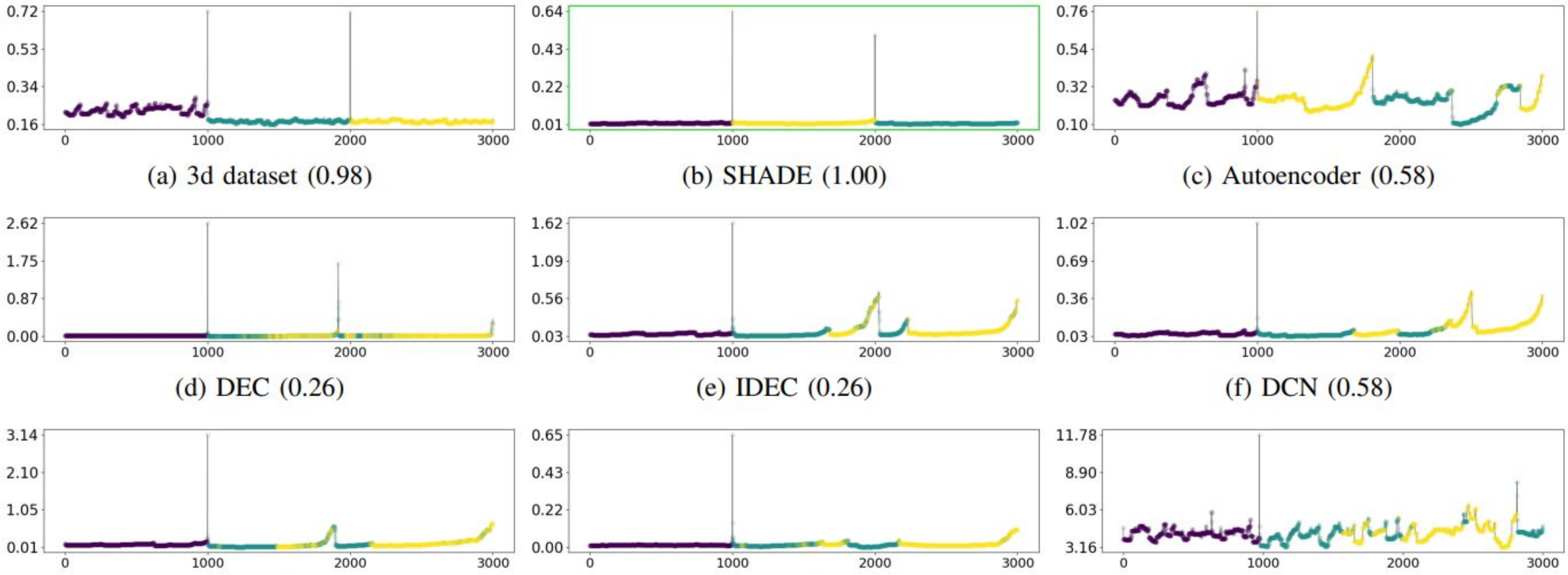
- It can be relevant to **preserve the shape** in the embedding



(d) DEC (0.59) (e) IDEC (0.64) (f) DCN (0.64)

(g) DipEncoder (0.60) (h) DipDECK (0.57) (i) DDC (0.25)

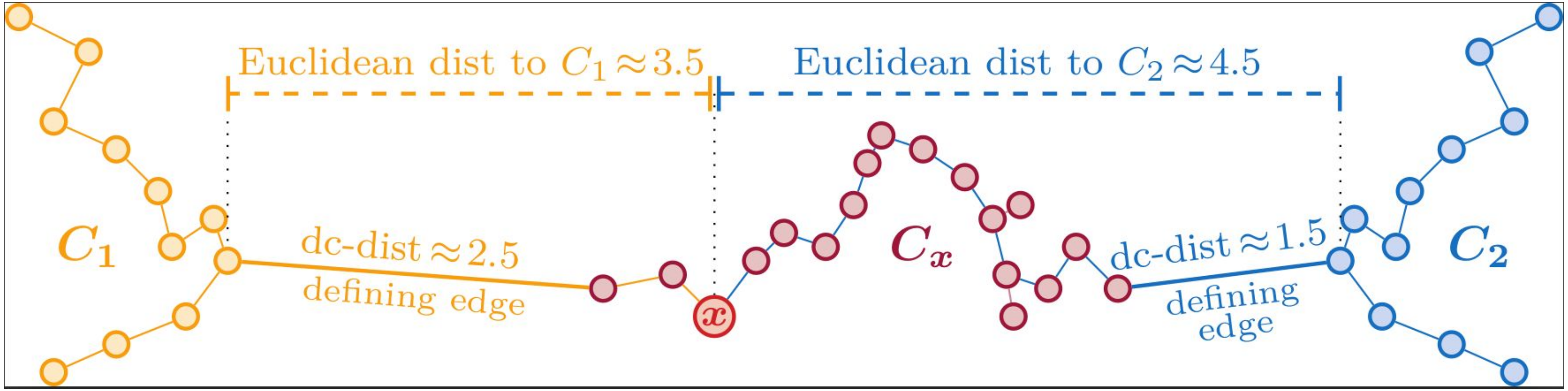
OPTICS Plots



(a) 3d dataset (0.98)
(b) SHADE (1.00)
(c) Autoencoder (0.58)
(d) DEC (0.26)
(e) IDEC (0.26)
(f) DCN (0.58)
(g) DipEncoder (0.41)
(h) DipDECK (0.67)
(i) DDC (0.44)

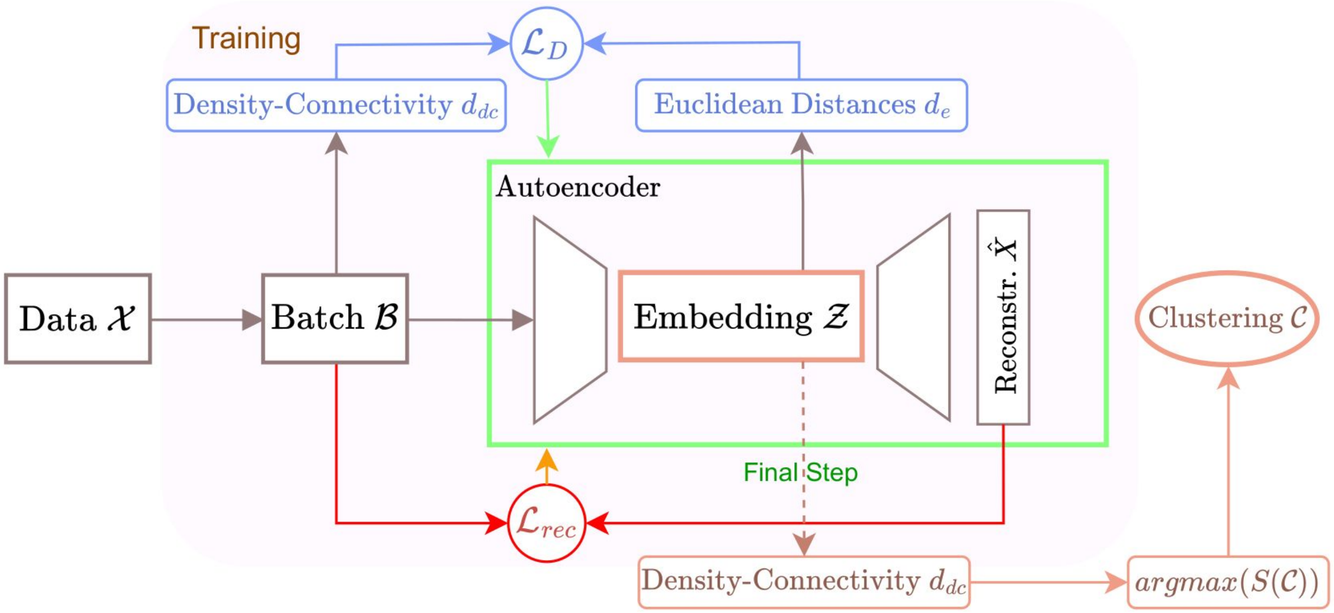
Capturing Density

- Capture the density-connectivity with **dc-distance** d_{dc}
- Similar to **minimax path distance**, enriched with concept of density by using **core distances**
- Density-connectivity loss $\mathcal{L}_d$: Similarity between Euclidean distance in low-dimensional embedding and dc-distance in high-dimensional space



Overview of SHADE

$$\mathcal{L}_d = \frac{1}{|\mathcal{B}|^2} \sum_{x_i, x_j \in \mathcal{B}} \left(d_{dc}(x_i, x_j) - d_{eucl}(z_i, z_j) \right)^2$$

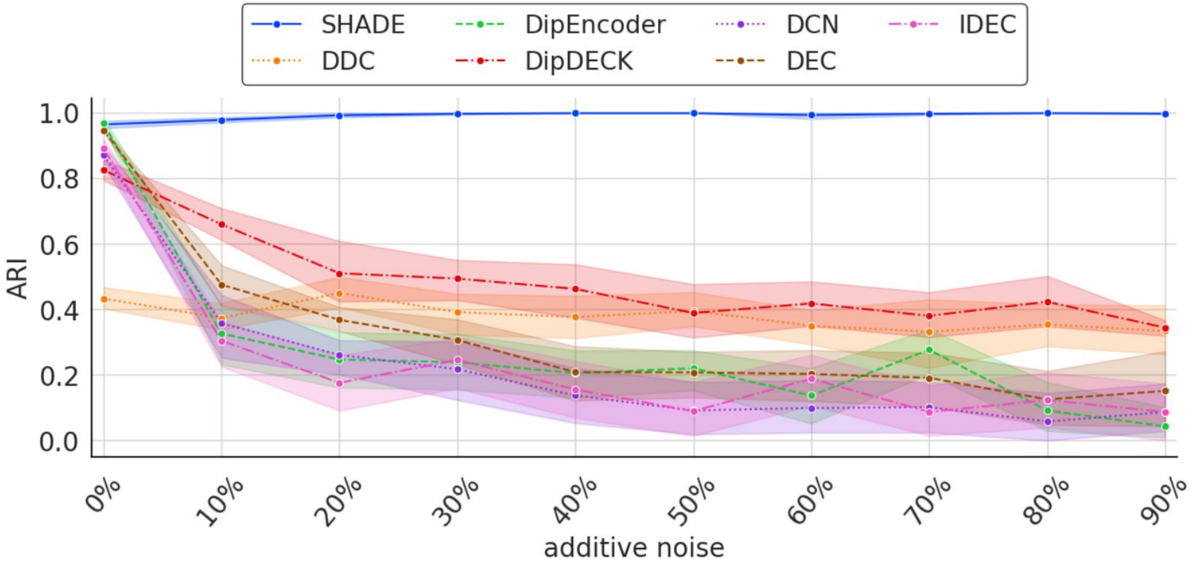


$$\mathcal{L}_{rec} = \frac{1}{|\mathcal{B}|} \sum_{x_i \in \mathcal{B}} \|x_i - \hat{x}_i\|_2^2$$

Results

	Dataset	SHADE	SHADE_1nn	DDC	DipDECK	DipEncoder	DCN	DEC	IDEC
Tabular data	Synth_low	99.4 ± 1.8	**98.9 ± 2.0**	56.9 ± 5.5	33.9 ± 6.4	10.1 ± 9.9	9.6 ± 9.7	40.2 ± 3.0	15.3 ± 8.8
	Synth_high	98.4 ± 1.2	**97.5 ± 1.4**	33.9 ± 11.1	29.9 ± 13.6	9.3 ± 10.7	8.8 ± 10.5	30.3 ± 3.5	17.9 ± 6.6
	letterrec.	43.2 ± 1.7	23.0 ± 0.9	9.9 ± 2.9	7.3 ± 3.5	24.7 ± 1.3	22.6 ± 1.0	23.9 ± 1.6	**25.2 ± 1.8**
	htru2	72.8 ± 23.4	**65.0 ± 19.5**	49.4 ± 13.0	9.7 ± 19.4	4.3 ± 0.8	49.7 ± 2.8	3.0 ± 0.5	3.2 ± 0.6
	Mice	32.5 ± 3.8	**27.7 ± 2.9**	25.2 ± 1.9	22.7 ± 4.3	21.6 ± 2.6	21.7 ± 1.4	22.0 ± 1.5	21.8 ± 1.4
	Pendigits	85.1 ± 1.4	75.1 ± 0.8	**76.9 ± 2.0**	74.3 ± 1.1	64.6 ± 3.0	61.6 ± 1.9	65.7 ± 3.3	64.9 ± 2.6
Video	Weizmann	57.1 ± 5.9	**48.2 ± 3.6**	14.7 ± 1.8	12.0 ± 1.9	23.3 ± 1.2	24.6 ± 1.1	24.9 ± 1.2	24.7 ± 1.2
	Keck	9.3 ± 0.5	**7.5 ± 0.4**	-0.2 ± 1.1	6.9 ± 0.8	7.1 ± 0.3	6.4 ± 0.5	6.1 ± 0.9	6.2 ± 0.9
Image	COIL20	82.5 ± 4.5	**68.7 ± 3.5**	62.0 ± 5.5	50.5 ± 7.8	64.0 ± 3.0	62.4 ± 2.8	63.7 ± 2.8	62.9 ± 2.9
	COIL100	78.1 ± 7.3	56.8 ± 5.0	16.4 ± 3.8	21.4 ± 3.0	54.3 ± 1.9	55.9 ± 3.0	55.8 ± 2.0	**56.9 ± 2.0**
	cmu_faces	38.9 ± 7.6	34.6 ± 6.2	35.0 ± 3.5	29.8 ± 9.8	37.9 ± 2.2	**40.3 ± 2.0**	35.8 ± 2.8	39.4 ± 3.3

Robust against Noise



Hyperparameters - Ablation Studies

- **μ / min_points**: Similar results for the tested values μ $\in$ [3, 7]
  - Default value: μ = 5
- **batchsize**: 300 is large enough to estimate the dc-distance d_{dc} good enough
  - Default value: batchsize = 500

- **No other Parameters!**

Summary

- SHADE is the first deep density-based clustering method
- It preserves density-connected structures in low-dimensional embeddings
- SHADE finds noise, arbitrarily shaped clusters, and detects the number of clusters fully automatically

ArXiv: https://arxiv.org/abs/2410.06265

Code: https://github.com/pasiweber/SHADE

Density-based vocabulary

- Core Distance (where x_μ is the μ-th nearest neighbor of a point x)

$$d_{core}(x) = d_{eucl}(x, x_\mu)$$

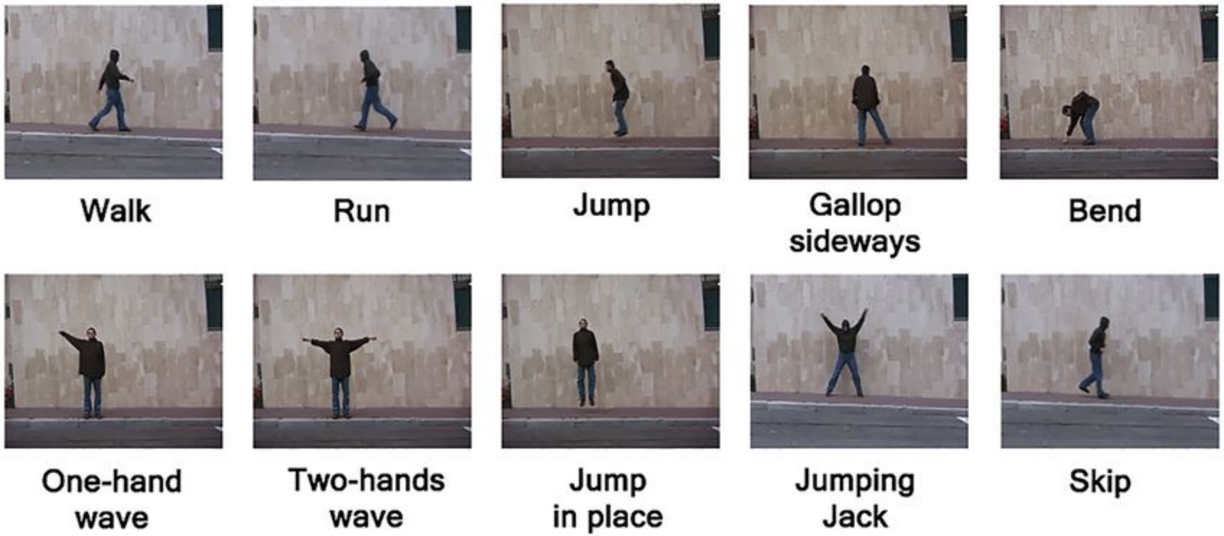
- Mutual Reachability Distance

$$d_m(x, y) = \max(d_{eucl}(x, y), core\_dist(x), core\_dist(y))$$

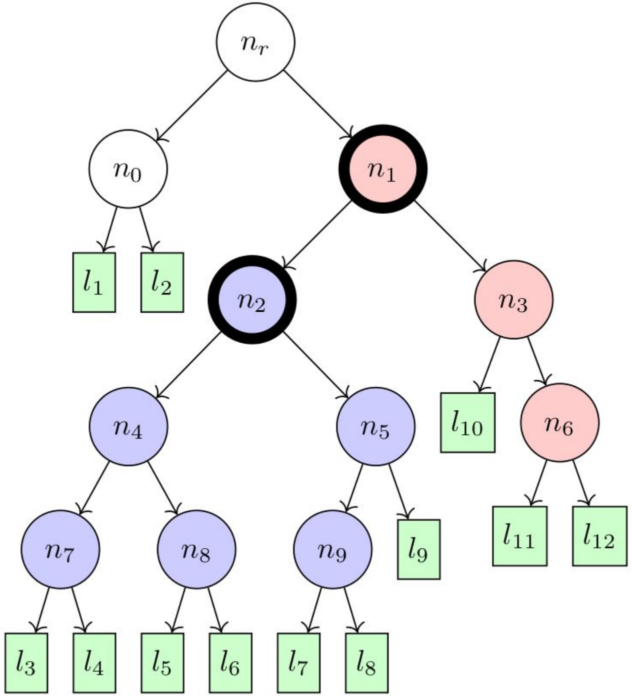
- dc-distance d_{dc}
  - Create minimum spanning tree on mutual reachability distance
  - Get minimax path distance from that tree - longest edge on the path between x and y is the dc-distance between those points

Video data often has density-based clusters in HD space

- Each scene of a movie is a density-connected cluster
  - every image/frame is similar to the next one, but beginning and end of a scene might be very different



Extract clustering from the tree metric



$$S(c) = \left(\frac{1}{\mathcal{T}_d(a)} - \frac{1}{\mathcal{T}_d(p(a))} \right) \cdot |l(a)|$$

	Dataset	n	d	k	#noise	Source
Tabular data	Synth_low	5000	100	10	500	[18]
	Synth_high	5000	100	10	500	[18]
	HAR	10,299	561	6	0	[26]
	letterrecognition	20,000	16	26	0	[26]
	htru2	17,898	8	2	0	[26]
	Mice	1,077	68	8	0	[26]
	TCGA	801	20,264	5	0	[26]
	Pendigits	10,992	16	10	0	[26]
Video	Weizmann	5,701	77,760	90	0	[4]
	Keck	25,457	120,000	60	0	[43]
Image data	COIL20	1,440	16,384	20	0	[26]
	COIL100	7,200	49,152	100	0	[26]
	cmu_faces	624	960	20	0	[26]
	Optdigits	5,620	64	10	0	[26]
	USPS	9,298	256	10	0	[16]
	MNIST	70,000	784	10	0	[21]
	FMNIST	70,000	784	10	0	[40]
	KMNIST	70,000	784	10	0	[8]

	Dataset	Metric	SHADE	SHADE_1nn	DDC	DipDECK	DipEncoder	DCN	DEC	IDEC
Tabular data	Synth_low	ARI	99.4 ± 1.8	**98.9 ± 2.0**	56.9 ± 5.5	33.9 ± 6.4	10.1 ± 9.9	9.6 ± 9.7	40.2 ± 3.0	15.3 ± 8.8
	(*noise*: 1.1 ± 1.3)	NMI	99.7 ± 1.0	**99.2 ± 1.2**	82.9 ± 3.2	58.0 ± 5.0	31.3 ± 13.6	29.4 ± 13.4	69.4 ± 3.0	44.3 ± 9.5
	$k = 10$	k	10.0 ± 0.0	10.0 ± 0.0	16.8 ± 1.0	4.1 ± 0.5	-	-	-	-
	Synth_high	ARI	98.4 ± 1.2	**97.5 ± 1.4**	33.9 ± 11.1	29.9 ± 13.6	9.3 ± 10.7	8.8 ± 10.5	30.3 ± 3.5	17.9 ± 6.6
	(*noise*: 2.4 ± 1.2)	NMI	98.1 ± 1.2	**97.3 ± 1.3**	61.8 ± 6.1	53.2 ± 12.0	28.9 ± 13.4	26.9 ± 13.2	61.4 ± 4.8	46.7 ± 5.0
	$k = 10$	k	13.0 ± 1.5	13.0 ± 1.5	10.3 ± 2.1	4.7 ± 1.3	-	-	-	-
	HAR	ARI	36.0 ± 5.6	36.4 ± 6.4	49.4 ± 3.3	51.3 ± 4.0	60.0 ± 6.9	**66.1 ± 1.3**	63.4 ± 2.4	64.9 ± 0.9
	(*noise*: 3.2 ± 5.2)	NMI	58.5 ± 5.8	58.2 ± 5.5	68.2 ± 2.1	71.3 ± 2.0	73.6 ± 4.1	**75.4 ± 1.2**	75.0 ± 1.8	74.6 ± 0.7
	$k = 6$	k	3.3 ± 2.0	3.3 ± 2.0	4.2 ± 1.0	3.1 ± 0.3	-	-	-	-
	letterrec.	ARI	43.2 ± 1.7	23.0 ± 0.9	9.9 ± 2.9	7.3 ± 3.5	24.7 ± 1.3	22.6 ± 1.0	23.9 ± 1.6	**25.2 ± 1.8**
	(*noise*: 50.3 ± 1.8)	NMI	75.6 ± 1.0	**57.4 ± 0.5**	43.5 ± 2.5	34.4 ± 3.8	49.7 ± 0.9	46.4 ± 1.0	50.8 ± 1.0	49.7 ± 1.2
	$k = 26$	k	111.4 ± 4.9	111.4 ± 4.9	14.0 ± 1.0	13.0 ± 2.4	-	-	-	-
	htru2	ARI	72.8 ± 23.4	**65.0 ± 19.5**	49.4 ± 13.0	9.7 ± 19.4	4.3 ± 0.8	49.7 ± 2.8	3.0 ± 0.5	3.2 ± 0.6
	(*noise*: 14.0 ± 3.8)	NMI	59.9 ± 19.6	**47.4 ± 14.4**	42.1 ± 7.0	5.5 ± 11.0	10.5 ± 1.4	31.6 ± 7.0	10.8 ± 0.6	10.7 ± 0.6
	$k = 2$	k	5.9 ± 2.1	5.9 ± 2.1	3.7 ± 0.6	1.2 ± 0.4	-	-	-	-
	Mice	ARI	32.5 ± 3.8	**27.7 ± 2.9**	25.2 ± 1.9	22.7 ± 4.3	21.6 ± 2.6	21.7 ± 1.4	22.0 ± 1.5	21.8 ± 1.4
	(*noise*: 30.0 ± 5.5)	NMI	56.2 ± 4.6	48.7 ± 2.3	**49.6 ± 2.0**	48.0 ± 5.3	38.7 ± 2.6	38.3 ± 1.5	39.4 ± 1.8	38.3 ± 1.8
	$k = 8$	k	11.9 ± 2.8	11.9 ± 2.8	15.3 ± 1.2	13.4 ± 5.3	-	-	-	-
	TCGA	ARI	86.6 ± 7.8	80.0 ± 13.7	87.5 ± 0.8	88.8 ± 4.4	**93.4 ± 6.0**	87.2 ± 5.3	85.1 ± 2.7	82.6 ± 0.9
	(*noise*: 21.8 ± 9.0)	NMI	90.8 ± 4.4	87.4 ± 7.2	91.9 ± 0.7	91.9 ± 2.1	**94.0 ± 3.8**	89.1 ± 3.1	89.5 ± 1.2	86.1 ± 1.4
	$k = 5$	k	5.7 ± 0.9	5.7 ± 0.9	6.0 ± 0.0	5.9 ± 0.5	-	-	-	-
	Pendigits	ARI	85.1 ± 1.4	75.1 ± 0.8	**76.9 ± 2.0**	74.3 ± 1.1	64.6 ± 3.0	61.6 ± 1.9	65.7 ± 3.3	64.9 ± 2.6
	(*noise*: 17.5 ± 2.2)	NMI	89.3 ± 0.9	82.7 ± 0.6	**84.1 ± 1.0**	82.0 ± 0.8	75.2 ± 1.3	72.7 ± 0.6	76.7 ± 1.4	75.9 ± 0.9
	$k = 10$	k	20.0 ± 1.8	20.0 ± 1.8	13.0 ± 0.4	14.8 ± 0.7	-	-	-	-
Video data	Weizmann	ARI	57.1 ± 5.9	**48.2 ± 3.6**	14.7 ± 1.8	12.0 ± 1.9	23.3 ± 1.2	24.6 ± 1.1	24.9 ± 1.2	24.7 ± 1.2
	(*noise*: 14.6 ± 2.6)	NMI	86.6 ± 0.6	**80.2 ± 1.1**	57.5 ± 2.0	52.8 ± 1.9	61.9 ± 0.8	62.3 ± 0.9	63.7 ± 1.0	63.3 ± 1.0
	$k = 90$	k	57.1 ± 2.1	57.1 ± 2.1	20.8 ± 2.4	24.5 ± 2.2	$88.4 \pm 1.2^{*}$	-	-	-
	Keck	ARI	9.3 ± 0.5	**7.5 ± 0.4**	-0.2 ± 1.1	6.9 ± 0.8	7.1 ± 0.3	6.4 ± 0.5	6.1 ± 0.9	6.2 ± 0.9
	(*noise*: 25.4 ± 1.1)	NMI	66.3 ± 0.3	**61.6 ± 0.3**	18.1 ± 4.3	34.9 ± 6.0	42.4 ± 0.6	39.4 ± 2.5	39.9 ± 3.5	39.2 ± 5.2
	$k = 60$	k	226.6 ± 10.4	226.6 ± 10.4	10.6 ± 3.1	32.3 ± 7.4	-	-	-	-
Image data	COIL20	ARI	82.5 ± 4.5	**68.7 ± 3.5**	62.0 ± 5.5	50.5 ± 7.8	64.0 ± 3.0	62.4 ± 2.8	63.7 ± 2.8	62.9 ± 2.9
	(*noise*: 12.5 ± 1.9)	NMI	93.6 ± 1.7	**85.6 ± 1.3**	85.5 ± 0.9	79.9 ± 2.4	80.2 ± 1.1	79.6 ± 1.3	80.6 ± 1.0	80.0 ± 1.1
	$k = 20$	k	16.5 ± 1.0	16.5 ± 1.0	14.3 ± 0.8	18.9 ± 1.0	-	-	-	-
	COIL100	ARI	78.1 ± 7.3	56.8 ± 5.0	16.4 ± 3.8	21.4 ± 3.0	54.3 ± 1.9	55.9 ± 3.0	55.8 ± 2.0	**56.9 ± 2.0**
	(*noise*: 24.9 ± 1.7)	NMI	94.4 ± 0.9	85.4 ± 0.6	69.5 ± 3.0	69.5 ± 1.7	82.6 ± 0.6	84.2 ± 0.7	**85.8 ± 0.5**	84.7 ± 0.5
	$k = 100$	k	77.8 ± 3.2	77.8 ± 3.2	18.3 ± 3.3	26.8 ± 1.9	99.8 ± 0.4	-	-	-
	cmu_faces	ARI	38.9 ± 7.6	34.6 ± 6.2	35.0 ± 3.5	29.8 ± 9.8	37.9 ± 2.2	**40.3 ± 2.0**	35.8 ± 2.8	39.4 ± 3.3
	(*noise*: 15.1 ± 4.1)	NMI	69.1 ± 5.3	65.0 ± 4.9	64.4 ± 2.3	62.3 ± 7.6	67.7 ± 1.6	**68.5 ± 1.0**	66.1 ± 2.0	68.2 ± 2.2
	$k = 20$	k	10.6 ± 1.7	10.6 ± 1.7	12.0 ± 0.9	8.9 ± 2.2	-	-	-	-

Dataset	Metric	SHADE	SHADE_1nn	DDC	DipDECK	DipEncoder	DCN	DEC	IDEC
Optdigits	ARI	93.3 ± 1.9	78.0 ± 7.4	$\mathbf{88.9 \pm 2.3}$	82.2 ± 2.3	80.2 ± 4.0	77.0 ± 3.5	80.4 ± 3.1	80.7 ± 3.6
(*noise*: 38.2 ± 5.3)	NMI	94.6 ± 1.0	83.4 ± 3.9	$\mathbf{91.7 \pm 1.3}$	86.2 ± 0.9	85.6 ± 2.0	82.9 ± 1.6	86.3 ± 1.4	86.2 ± 1.5
$k = 10$	k	12.4 ± 1.0	12.4 ± 1.0	11.1 ± 0.5	11.0 ± 1.2	-	-	-	-
USPS	ARI	88.1 ± 2.7	68.1 ± 1.5	$\mathbf{92.2 \pm 2.2}$	68.2 ± 5.1	72.5 ± 2.0	66.2 ± 1.3	73.4 ± 0.9	74.0 ± 0.9
(*noise*: 45.7 ± 2.8)	NMI	91.8 ± 1.1	75.9 ± 1.1	$\mathbf{91.0 \pm 0.8}$	78.5 ± 2.5	79.9 ± 1.6	74.5 ± 1.0	81.0 ± 0.5	81.3 ± 0.6
$k = 10$	k	9.5 ± 0.9	9.5 ± 0.9	9.8 ± 0.4	8.0 ± 1.2	-	-	-	-
MNIST	ARI	86.1 ± 8.4	54.1 ± 2.8	$\mathbf{94.5 \pm 1.2}$	80.9 ± 3.0	79.6 ± 2.8	82.0 ± 4.5	79.5 ± 2.1	81.9 ± 2.0
(*noise*: 58.9 ± 5.3)	NMI	84.5 ± 2.4	63.7 ± 1.3	$\mathbf{93.7 \pm 0.8}$	84.5 ± 1.5	84.9 ± 1.6	84.6 ± 1.9	85.1 ± 1.0	87.5 ± 0.8
$k = 10$	k	38.0 ± 9.9	38.0 ± 9.9	10.0 ± 0.0	11.3 ± 0.8	-	-	-	-
FMNIST	ARI	72.2 ± 2.7	35.7 ± 1.4	35.5 ± 7.1	46.5 ± 0.9	43.5 ± 3.6	45.8 ± 2.9	42.3 ± 4.1	$\mathbf{46.9 \pm 3.6}$
(*noise*: 66.3 ± 2.3)	NMI	73.4 ± 1.0	53.9 ± 0.9	64.1 ± 3.3	$\mathbf{65.3 \pm 0.7}$	59.4 ± 2.4	62.4 ± 2.2	59.3 ± 2.7	63.7 ± 2.5
$k = 10$	k	68.7 ± 9.2	68.7 ± 9.2	7.1 ± 0.9	9.6 ± 1.1	-	-	-	-
KMNIST	ARI	58.6 ± 9.4	27.5 ± 1.7	$\mathbf{60.3 \pm 3.5}$	33.8 ± 9.9	43.6 ± 1.1	40.3 ± 1.6	42.4 ± 0.9	42.8 ± 1.1
(*noise*: 73.4 ± 2.2)	NMI	66.2 ± 3.1	46.8 ± 0.6	$\mathbf{73.6 \pm 1.6}$	55.9 ± 5.0	56.7 ± 1.3	53.6 ± 1.4	55.7 ± 1.1	56.7 ± 1.2
$k = 10$	k	64.1 ± 10.4	64.1 ± 10.4	16.2 ± 0.7	11.7 ± 2.8	-	-	-	-