

# Multi-view Diffusion Geometry via Intertwined Diffusion Trajectories

Argyris Kalogeratos

joint work with

Gwendal Debaussart-Joniec

ENS Paris-Saclay, Centre Borelli



Geometric Machine Learning Conference 2026 @ Paris

# Outline

---

Motivation & Background

The MDT Framework

Sampling and Learning of Trajectories

Experiments

Conclusions

Part 1

# **Motivation & Background**

# Multi-view data: same objects, multiple representations...

---

## Multi-view data

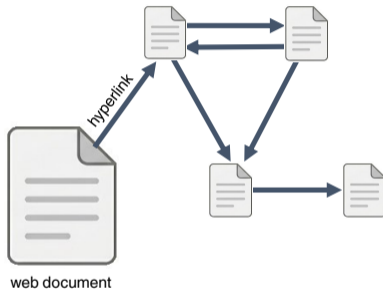
- Multiple modalities:  
text, hyperlinks, image, audio...
- Multiple measuring instruments/sensors
- Multiple feature extractions or preprocessings

## Approaches to combine views

- Representation *fusion*
- Representation *alignment*

## Tasks

- Manifold/graph learning
- Data clustering

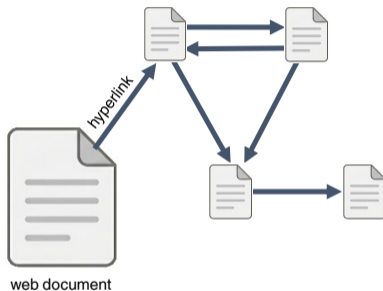


# Multi-view data: same objects, multiple representations...

---

## Challenges

- Concatenation meltdown:  
fusing inhomogeneous representations –  
different dimensions, sparsity, structure
- Superposition
- Fusion stage: early- vs. late-fusion



Views with different dimensions



Views with different sparsity



# Multi-view data: same objects, multiple representations...

---

## Challenges

- Concatenation meltdown:  
fusing inhomogenous representations –  
different dimensions, sparsity, structure
- Superposition
- Fusion stage: early- vs. late-fusion
- Generating new views
- Finding a common representation space

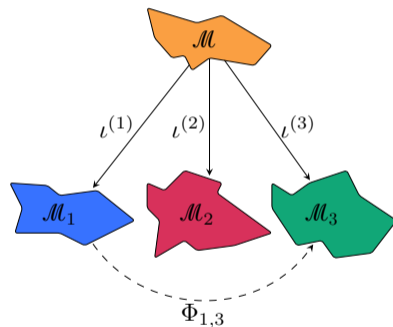


# Multi-view data: same objects, multiple representations...

---

## Challenges

- Concatenation meltdown:  
fusing inhomogenous representations –  
different dimensions, sparsity, structure
- Superposition
- Fusion stage: early- vs. late-fusion
- Generating new views
- Finding a common representation space
- Common latent manifold assumption



$\mathcal{M}$ : latent common manifold

$\mathcal{M}_i$ : observed view deriving by a smooth embedding of  $\mathcal{M}$ ;  
those manifolds are diffeomorphic to each other

# Why diffusion-based methods for multi-view data?

---

## Diffusion-based methods

- Rooted on Markov chains, specifically random walks
- Random walks on  $K$ -NN graphs reveal multi-scale structure
- Heat-kernel intuition: powers of a similarity operator *denoise* it by emphasizing low-freq. eigenvectors (implicit spectral analysis)

...

- Fuse data in an *operator space* (e.g. similarity matrix/graph) and not in the original feature space
- Can handle both point-cloud data and graphs
- Reliable, interpretable, long history in clustering, manifold learning, viz.

[Coifman and Lafon, 2006, Haghverdi et al., 2015]

*Well-defined and well-behaved in the single-view case,  
challenging generalization to multiple views.*

# Single-view Diffusion Maps

## From point-cloud data to an operator

- Kernel  $\mathbf{K}_{ij} = \kappa(x_i, x_j)$ ,  $x$ : datapoints
- Degree matrix  $\mathbf{D} = \text{diag}(\mathbf{K} \mathbf{1}_{N \times N})$
- Diffusion operator (row-stochastic):

$$\mathbf{P} = \mathbf{D}^{-1} \mathbf{K} \in \mathbb{R}^{N \times N}$$

- $\mathbf{P}^t$ : homogeneous Markov chain after  $t$  steps

## Diffusion distance

$$(\mathcal{D}^t(i, j))^2 = \sum_k \frac{1}{\pi(k)} ([\mathbf{P}^t]_{ik} - [\mathbf{P}^t]_{jk})^2$$

$\pi$ : the stationary distribution of the Markov chain

## Diffusion Map embedding

Eigendecomposition  $\mathbf{P} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top$ :

$$\Psi^t(x_i) = e_i^\top \mathbf{U} \mathbf{\Lambda}^t$$

Isometry property:

$$\mathcal{D}^t(i, j) = \|\Psi^t(x_i) - \Psi^t(x_j)\|$$

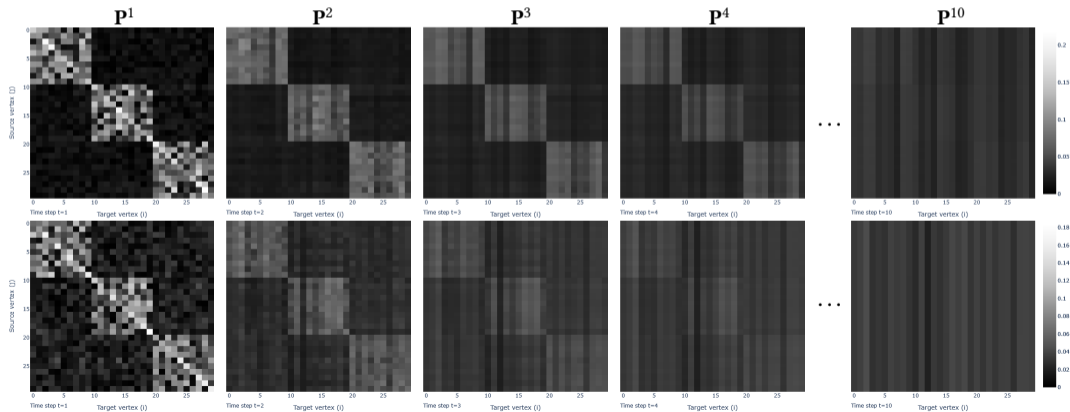
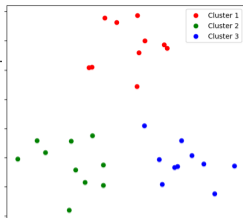
## In practice

Truncation ( $l$ ) and time selection ( $t$ )

$$\Psi^{l,t}(x_i) = [\lambda_1^t u_1(i), \dots, \lambda_l^t u_l(i)]^\top$$

- short  $t \rightarrow$  local structure
- long  $t \rightarrow$  global structure
- $t \rightarrow \infty$ : rank-1 limit is  $\mathbf{P}^t = \mathbf{1} \pi^\top$

# Speaking visual



# Existing multi-view diffusion methods: *fixed* designs!

---

**Setting:** Aligned views  $\{\mathbf{X}_v\}_{v=1}^V$  with operators  $\mathcal{P}_c = \{\mathbf{P}_v\}_{v=1}^V$  (canonical set)

**Existing operator-based methods** build a single *composite* operator  $\mathbf{Q}$

■ **Alternating Diffusion (AD):**  $\mathbf{Q}_{AD} = \mathbf{P}_1 \mathbf{P}_2$  [Katz et al., 2019]

■ **Integrated Diffusion (ID):**  $\mathbf{Q}_{ID} = \mathbf{P}_1^{t_1} \mathbf{P}_2^{t_2}$  [Kuchroo et al., 2022]

■ **Multi-View Diffusion (MVD):**  $\mathbf{K}_{MVD} = \begin{bmatrix} 0 & \mathbf{K}_1 \mathbf{K}_2 \\ \mathbf{K}_2 \mathbf{K}_1 & 0 \end{bmatrix}$  [Lindenbaum et al., 2020]  
 $\mathbf{Q}_{MVD} = \mathbf{D}_{MVD}^{-1} \mathbf{K}_{MVD}$

■ **Cross-Diffusion (CR-DIFF):** they use  $\mathbf{P}_v^\top$ , hence depart from row-stochasticity  
**Composite Diffusion (COM-DIFF)** [Wang et al., 2012, Shnitzer et al., 2018]

*Each is a single fixed rule for fusing views*

# Conceptual limitation

---

## Why *this* fixed rule?

- No theoretical or empirical reason *this* rule generalizes to arbitrary data
- Limited understanding of the associated *operator spaces* for comparing, learning, or sampling strategies
- No neutral *baseline* for evaluating sophisticated designs

### The proposed framework

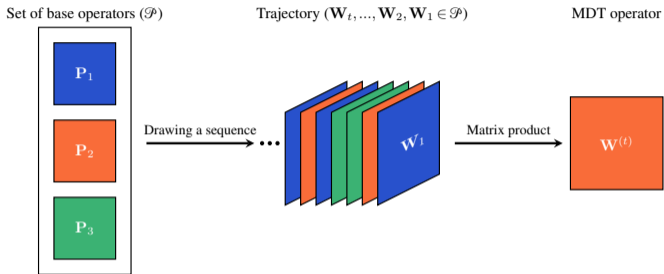
A stochastic view-interwining process that turns these fixed rules into points in a rich operator space

Method	Row-stoch.	Scales w/ $V$	Tunable $t$	Real eigvals	$\subset$ MDT
AD	✓	✓	-	-	✓
ID	✓	✓	✓	-	✓
MVD	✓	~	-	✓	-
CR-DIFF	-	✓	-	-	-
COM-DIFF	-	-	-	✓	-
<b>MDT (ours)</b>	✓	✓	✓	-	-

Part 2

# **The MDT Framework**

# Core concept: *Multi-view Diffusion Trajectory* (MDT)



- $V = 1$ :  $\mathbf{W}^{(t)} = \mathbf{P}^t \rightarrow$  classical Diffusion  
Maps recovered
- $V > 1$ : each  $\mathbf{W}_i$  specifies the view diffusion occurs in at step  $i$
- Same base operators  $\mathcal{P}$ , but *different trajectories yield different geometries*  
 $\rightarrow \mathcal{P}$  spans a rich operator space

## Main technical challenge

Matrix multiplication is non-commutative  
 $\rightarrow$  the *order* of operators matters;  
the process becomes *time-inhomogeneous*

# Probabilistic interpretation

---

**An MDT trajectory**  $(\mathbf{W}_s)_{s=1}^t$  defines a *time-inhomogeneous Markov chain*  $(X_s)_{s=0}^t$  on the datapoints:

1-step transition (at step  $s$ ):  $[\mathbf{W}_s]_{ij} = \mathbb{P}(X_s = j \mid X_{s-1} = i)$

$t$ -step transition:  $[\mathbf{W}^{(t)}]_{ij} = \mathbb{P}(X_t = j \mid X_0 = i)$

- At each step  $s$ , the law of the walk can change (a different  $\mathbf{W}_s$  is in effect)
- The walk explores the data *across multiple views* over time
- The trajectory space is explored based on the transition law...  
→ evidently richer than choosing a single  $\mathbf{Q}$

# The MDT space

---

**MDT trajectory space:**  $\mathcal{P}^+ = \bigcup_{t \geq 0} \mathcal{P}^t$  ...all any-length trajectories over  $\mathcal{P}$

**MDT operator space:**  $\mathcal{W} = \{\mathbf{W}^{(t)} \mid t \geq 1, \mathbf{W}_i \in \mathcal{P}\}$  ...all induced MDT operators

## Designs

- **Discrete  $\mathcal{P}$ :** the number of length- $t$  trajectories grows with  $|\mathcal{P}|^t$
- **Continuous  $\mathcal{P}_{\text{cvx}}$ :** a smooth manifold of operators over the base set  $\mathcal{P}$

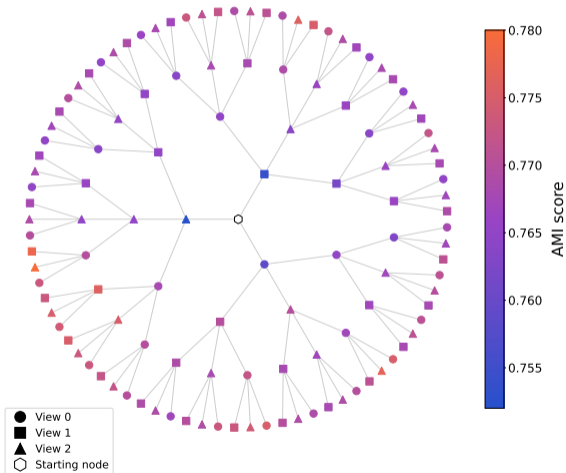
$$\mathcal{P}_{\text{cvx}} = \left\{ \sum_{v=1}^V a_v \mathbf{P}_v \mid a_v \geq 0, \sum_v a_v = 1 \right\}$$

- Each step is a convex combination of views
- $V - 1$  degrees of freedom per step under simplex constraint
- Smooth space  $\rightarrow$  gradient-based optimization possible
- Generalizes the PageRank random walk to multi-view

### Space richness

The MDT operator space  $\mathcal{W}$  is large and rich.  
We can *search*, *sample*, and *learn* MDTs.

# Discrete trajectory space as a $|\mathcal{P}|$ -ary tree



## Discrete MDTs as paths

- Root: identity matrix
- Depth is the trajectory length
- Branching at depth  $i$ :  
choose  $\mathbf{W}_i \in \mathcal{P}$
- Short trajectories: local structure
- Long trajectories: global structure  
common to multiple views
- No need to choose the fusion stage  
(this is decided by the optimization)
- Path  $\rightarrow$  geometry  $\rightarrow$  quality

# Basic properties

---

Assume every  $\mathbf{P} \in \mathcal{P}$  is row-stochastic with strictly positive diagonal, and the associated Markov chains are aperiodic and irreducible

## Stability of aperiodicity and irreducibility

Any product  $\mathbf{W}^{(t)} \in \mathcal{W}$  defines an aperiodic and irreducible chain

## Stationary distribution and convergence

- (i) Each  $\mathbf{W}^{(t)} \in \mathcal{W}$  admits a unique stationary distribution  $\pi_t$ : i.e.  $\pi_t^\top \mathbf{W}^{(t)} = \pi_t^\top$
- (ii) As  $t \rightarrow \infty$ ,  $\mathbf{W}^{(t)} \rightarrow \mathbf{1}\pi_\infty^\top$  (rank one)

## Caveats vs single-view DM

- $\mathbf{W}^{(t)}$  generally *not* self-adjoint in  $\langle \cdot, \cdot \rangle_{1/\pi_t}$ , and *not* reversible, i.e.  $\pi_t^\top \mathbf{W}^{(t)} \neq \mathbf{W}^{(t)} \pi_t$
- Decay of singular values of  $\mathbf{W}^{(t)}$  may be *non-monotonic* in  $t$
- Convergence rate depends on the trajectory chosen

# Trajectory-dependent diffusion geometry

---

## Trajectory-dependent diffusion distance

Let  $\pi_t$  be the stationary distribution of  $\mathbf{W}^{(t)}$ , then

$$\mathcal{D}_{\mathbf{W}^{(t)}}(x_i, x_j) = \sum_{k=1}^N \frac{1}{\pi_t(k)} (\mathbf{W}_{ik}^{(t)} - \mathbf{W}_{jk}^{(t)})^2$$

## Embedding

Since  $\mathbf{W}^{(t)}$  is generally not symmetric, we employ the SVD  $\mathbf{W}^{(t)} = \mathbf{U}_t \mathbf{\Sigma}_t \mathbf{V}_t^T$ :

$$\Psi^{(t)}(x_i) = e_i^T \mathbf{U}_t \mathbf{\Sigma}_t$$

## Diffusion distance preservation

$$\mathcal{D}_{\mathbf{W}^{(t)}}(x_i, x_j) = \left\| \Psi^{(t)}(x_i) - \Psi^{(t)}(x_j) \right\|^2$$

- Generalizes classical Diffusion Maps (eigen) to the inhomogeneous case (SVD)
- SVD truncation yields robust embeddings (denoising via top singular components)

Part 3

# **Sampling and Learning of Trajectories**

# Sampling and learning in the MDT space

---

## Sampling MDTs

Draw  $\mathbf{W}_i \sim \mu_i$  independently with  $\text{supp}(\mu_i) \subseteq \mathcal{P}$ , where  $\mu_i$  is a prob. distribution

### Use-cases

- (i) Random baseline (MDT-RAND)
- (ii) Data augmentation (sample  $V' > V$  views)
- (iii) Data summarization (sample  $V' < V$  views)

### Discrete-time Markov Jump Process over $\mathcal{P}$

If  $\mu_i = \mu$  for all  $i$  and the operators are i.i.d., the MJP process becomes time-homogeneous and the expected operator is:

$$\mathbb{E}[\mathbf{W}^{(t)}] = (\mathbb{E}_\mu[\mathbf{W}])^t$$

## Learning MDTs

Optimize an task-specific quality criterion  $Q$ :  
 $\mathbf{W}^* \in \arg \max_{\mathbf{W} \in \mathcal{W}} Q(\mathbf{W})$

### Strategies

- *Discrete* space  
→ Beam-Search,  $\epsilon$ -greedy, MCTS, DIRECT
- *Continuous* space  
→ ADAM, Bayesian/evolutionary methods

### What constrains the search

- Expressiveness of  $\mathcal{P}$
- Diffusion time  $t$
- The trajectory itself  
(while it's being developed)

# Random MDTs as a principled baseline

---

## Why random MDTs are a natural reference

- Parametrized only by  $\mathcal{P}$  and sampling distribution – minimal assumptions
- Unbiased mixture over view interactions
- Computationally cheap (no optimization)

## Performance Ratio to Random (PRR)

$$\text{PRR}(\text{method}, Q) = \frac{Q(\text{method})}{Q(\mathbb{E}_\mu[\mathbf{W}^{(t)}])}$$

where  $\mathbb{E}_\mu[\mathbf{W}^{(t)}]$  uses  $\mu$  uniform on the canonical set  $\mathcal{P}_c$ :

$$\mathbb{E}_\mu[\mathbf{W}^{(t)}] = \left( \frac{1}{|\mathcal{P}_c|} \sum_{\mathbf{P} \in \mathcal{P}_c} \mathbf{P} \right)^t$$

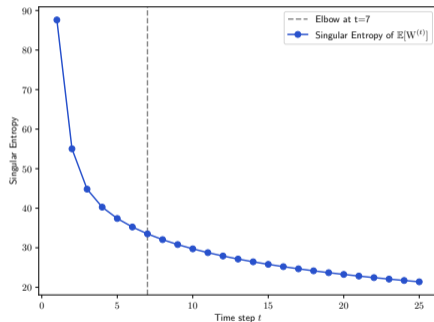
# Selecting the diffusion time $t$

## Singular Entropy (SE)

For a matrix  $\mathbf{A}$  with normalized singular values  $\tilde{\sigma}_i(\mathbf{A}) = \sigma_i(\mathbf{A}) / \sum_j \sigma_j(\mathbf{A})$ ,  $\sigma_1(\mathbf{A}) \geq \dots \geq \sigma_N(\mathbf{A})$ :

$$\text{SE}(\mathbf{A}) = - \sum_{i=1}^N \tilde{\sigma}_i(\mathbf{A}) \log \tilde{\sigma}_i(\mathbf{A})$$

- The spectrum of  $\mathbf{W}^{(t)}$  gets concentrated (non-monotonically) with  $t$   
→ dominant structure emerges
- Choose  $t$  at the *elbow point*
- For stability, evaluate *offline* and *generically*, on the expected operator under  $\mu$  uniform over the canonical set  $\mathcal{P}_c$ :  $\mathbb{E}_\mu[\mathbf{W}^{(t)}]$ .



# Quality measures and MDT variants

---

## Unsupervised quality measures

- For clustering – *Multi-view Calinski–Harabasz* (MvCH):  $\sum_v \tau_v \cdot \text{CH}(\mathbf{X}_v, C(\mathbf{M}))$   
Ratio of between- to within-cluster scatter
- For manifold learning – *Contrastive* (CST):  $\sum_v \tau_v \sum_i \sum_{j \in \mathcal{N}_i(v)} -\log \frac{\exp(\mathbf{M}_{ij})}{\sum_{k \neq i} \exp(\mathbf{M}_{ik})}$   
Encourages local neighborhood agreement
- Weights are  $\tau_v = 1/V$  when no prior on view importance

Variant	Trajectory space	Quality	Optimizer	Task
MDT-RAND	$\mathcal{P}_c^+$	-	-	Any
MDT-CVX-RAND	$\mathcal{P}_{cvx}^+$	-	-	Any
MDT-CST	$\mathcal{P}_{cvx}^+$	CST	ADAM	Manifold learning
MDT-DIRECT	$\mathcal{P}_{cvx}^+$	CH	DIRECT	Clustering
MDT-BEAM	$\mathcal{P}_c^+$	CH	BEAM-SEARCH	Clustering

# MDT-based learning pipeline

---

---

## Algorithm – Generic MDT-based learning pipeline

---

**Input:**  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_V$ : aligned multi-view data

$\mathcal{P}$ -design: choice of base operator set  $\triangleright$  e.g., discrete  $\mathcal{P}_c$  or continuous  $\mathcal{P}_{cvx}$

$\mathcal{T}$ : a downstream task  $\triangleright$  e.g., clustering, manifold learning

$Q$ : a task-suited quality measure

$\mathcal{O}$ : an optimizer compatible with  $\mathcal{P}$   $\triangleright$  see MDT variants table

**Output:**  $\mathbf{W}^*$ : learned MDT operator;  $\mathcal{T}(\mathbf{W}^*)$ : task output

---

Build the canonical set of random walk operators  $\mathcal{P}_c = \{\mathbf{P}_v\}_{v=1}^V$

Construct the base operator set  $\mathcal{P}$  from  $\mathcal{P}_c$

Determine the diffusion time  $t$   $\triangleright$  e.g., SE-elbow of  $(\mathbb{E}_\mu[\mathbf{W}^{(t)}])_t$

Determine  $\mathbf{W}^* \in \arg \max_{\mathbf{W} \in \mathcal{P}^t} Q(\mathbf{W})$ , solved via  $\mathcal{O}$

Apply the downstream task to obtain  $\mathcal{T}(\mathbf{W}^*)$

**return**  $\mathbf{W}^*, \mathcal{T}(\mathbf{W}^*)$

---

Part 4

# **Experiments**

# Experimental setup

---

## Datasets (synthetic + real, $V \in \{2, \dots, 6\}$ )

- *Manifold*: Helix-A, Helix-B [Lindenbaum et al., 2020], Deformed-Plane (ours)
- *Clustering*: K-MvMNIST [Kuchroo et al., 2022], L-MvMNIST [Lindenbaum et al., 2020], Olivetti, Yale, 100Leaves, L-Isolet, MSRC, Multi-Feat, Caltech101-7

## Preprocessing

- Gaussian kernel, max-min bandwidth
- K-NN graph with  $K = \lceil \log N \rceil$
- Time parameter  $t$  via the elbow of SE

## Baselines compared against

- *Diffusion*: AD, ID, P-AD, MVD, CR-DIFF, COM-DIFF
- *Non-diffusion*: GCCA [Afshin-Pour et al., 2012], MVSC [Kumar et al., 2011]

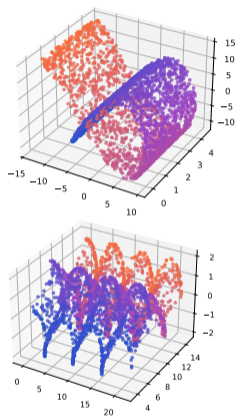
## Evaluation metrics

- *Manifold*: Trustworthiness, Continuity [Venna and Kaski, 2001]
- *Clustering*: Adjusted Mutual Information
- *Across methods*: PRR vs MDT-RAND

*Higher values are better*

*100 runs per setting, mean  $\pm$  std reported*

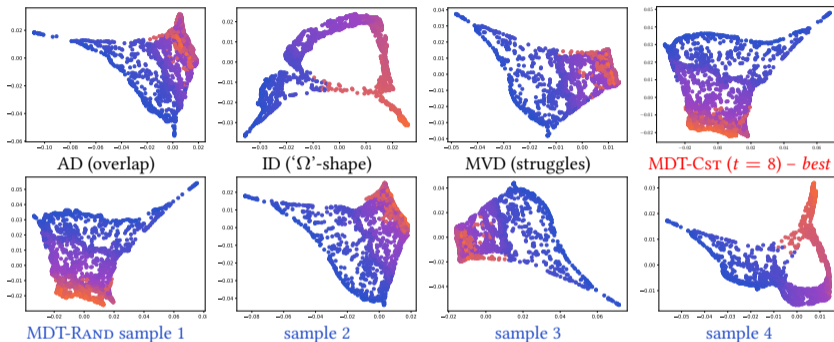
# Manifold learning results



**Deformed-Plane dataset:**

View 1 & 2

Two non-bijective  
non-linear deformations of  
the same 2D plane



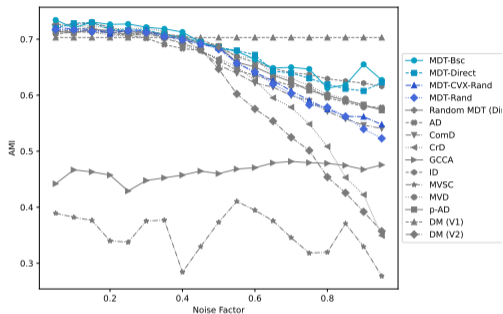
Method	Deformed Plane		Helix-A		Helix-B		PRR	
	Trust.	Cont.	Trust.	Cont.	Trust.	Cont.	Trust.	Cont.
AD	<u>95.95</u>	98.36	99.19	99.74	99.52	<b>100.00</b>	1.01	<u>1.00</u>
MVD	75.34	93.71	99.47	99.80	82.35	<u>99.88</u>	0.88	0.99
ID	89.80	96.09	<b>99.69</b>	<u>99.91</u>	<u>99.84</u>	<b>100.00</b>	0.99	0.99
MDT-Cst	<b>98.49</b>	<b>99.31</b>	<u>99.83</u>	<b>99.92</b>	<b>100.00</b>	<b>100.00</b>	<b>1.02</b>	<b>1.01</b>
MDT-Cvx-RAND	95.85	<u>98.54</u>	99.66	<u>99.86</u>	<b>100.00</b>	<b>100.00</b>	<u>1.01</u>	<u>1.00</u>
MDT-RAND	92.31	97.87	99.65	<u>99.86</u>	<b>100.00</b>	<b>100.00</b>	1.00	<u>1.00</u>

# Clustering results

Method	K-MvMNIST	L-MvMNIST	Olivetti	Yale	100Leaves	L-Isolet	MSRC	Multi-Feat	Caltech101-7	PRR
GCCA	46.0	43.5	69.7	46.8	87.8	70.8	64.7	63.6	26.4	0.84
MVSC	37.3	32.1	<b>80.2</b>	<b>59.4</b>	<b>93.2</b>	75.8	73.1	80.7	40.2	0.93
CR-DIFF	66.6	59.5	71.8	56.2	44.4	<b>79.8</b>	48.3	76.9	<b>65.9</b>	0.93
AD	66.3	61.0	75.1	58.0	80.0	77.2	50.0	83.0	56.7	0.99
MVD	<b>68.8</b>	61.6	75.8	56.6	82.4	8.0	48.8	82.5	56.1	0.90
ID	<u>68.8</u>	61.5	75.0	52.4	73.8	<u>79.8</u>	35.2	83.4	45.8	0.93
P-AD	<u>68.8</u>	<u>62.1</u>	75.5	57.7	69.9	78.8	41.0	81.9	58.2	0.97
MDT-DIRECT	68.5	<b>62.2</b>	<u>76.9</u>	<u>59.4</u>	<u>91.4</u>	77.9	<b>76.2</b>	<u>85.4</u>	63.6	<b>1.08</b>
MDT-BEAM	68.4	61.8	75.1	58.6	61.7	79.6	73.7	<b>87.3</b>	<u>64.1</u>	1.03
MDT-CVX-RAND	68.5	61.7	76.5	58.7	86.2	78.0	<u>74.6</u>	84.8	63.5	<u>1.07</u>
MDT-RAND	68.2	61.3	76.0	57.1	70.5	77.8	61.3	80.3	58.6	1.00

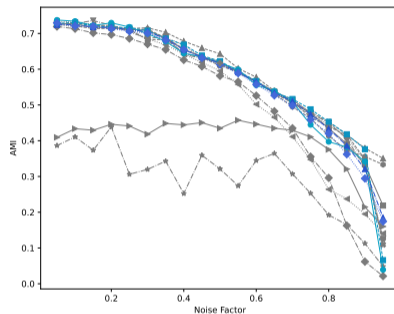
- The table reports average AMI and PRR scores across 9 datasets
- **MDT-DIRECT** wins on 5/9 datasets, ranks 2nd in one case
- Convex variants (MDT-CVX-RAND, MDT-DIRECT) consistently strong → design of  $\mathcal{P}$  matters
- MDT-BEAM adapts  $t$  per dataset; can be misled when CH disagrees with labels (100Leaves)

# Clustering results: Robustness to noise



K-MvMNIST – View 1: original images

View 2: added Gaussian noise (std)



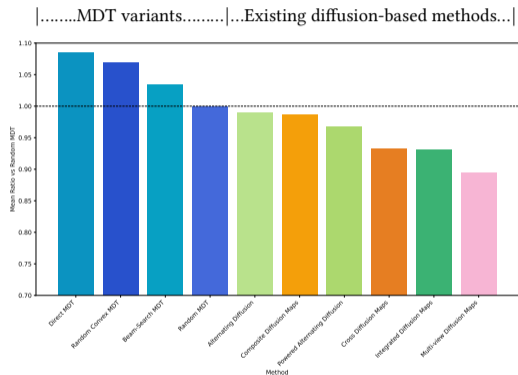
L-MvMNIST – View 1: images with Gaussian noise

View 2: a % of pixels are dropped

## Findings

- K-MvMNIST: methods that weight views (ID, MDT-DIRECT, MDT-BEAM) decay more gracefully than uniform-treatment ones
- L-MvMNIST: all methods drop similarly – no view can compensate
- Optimized MDT variants leverage informative views while reducing impact of noisier ones

# Clustering results: Random MDTs are hard to beat



Average PRR score across 9 datasets,  
wrt MDT-RAND as neutral baseline

## Findings

- All three learned MDT variants beat the MDT-RAND baseline  
→ MDT learning has an effect
- All existing diffusion-based methods do *not* beat MDT-RAND on average  
→ Some of them look into a tiny slice of the MDT operator space

## Implications

MDT-RAND is a fair, principled, cheap yet hard-to-beat reference for future multi-view diffusion benchmarks

# Clustering results: MDTs as a data augmentation tool

---

## Recipe

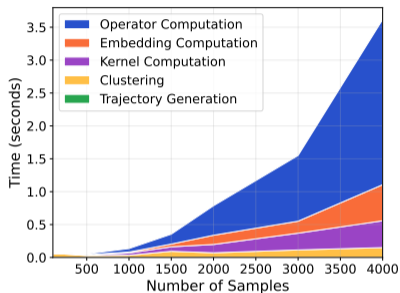
- Sample  $V' > V$  random MDTs from  $\mathcal{P}_{\text{cvx}}$ , no optimization (below, we sample  $V' = 10$  views)
- Use them as augmented input views for *non-diffusion* multi-view methods

Method	K-MvMNIST	L-MvMNIST	Olivetti	Yale	100Leaves	L-Isolet	Multi-Feat	Caltech101-7
GCCA	46.0	43.5	69.7	46.8	87.8	70.8	63.6	26.4
GCCA + MDT	<b>69.3</b>	<b>61.4</b>	80.0	<u>59.2</u>	92.0	<u>77.3</u>	83.0	<b>56.7</b>
MVSC	37.3	32.1	<u>80.2</u>	59.4	<u>93.2</u>	75.8	80.7	40.2
MVSC + MDT	<u>68.0</u>	<u>57.7</u>	<b>80.2</b>	58.8	<b>95.1</b>	<b>77.7</b>	<b>90.2</b>	<u>55.4</u>

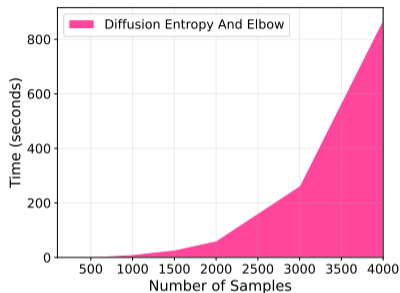
## Findings

- Substantial gains on noisy/hard datasets (e.g. +23% AMI for GCCA on K-MvMNIST)
- Non-diffusion methods benefit from MDT-augmented views  
→  $\mathcal{W}$  contains *many* high-quality data representations, not specific to diffusion-based pipelines
- Cheap:  $V'$  samples without optimization

# Runtime analysis



Pipeline cost breakdown vs.  $N$



Time-selection (SE) vs.  $N$

## Findings

- Dominant cost: SVD of  $\mathbf{W}^{(t)}$  – standard for diffusion methods
- Sampling random MDTs is essentially free once  $\mathcal{P}$  is built
- DIRECT and Beam-Search add overhead, but scale well with  $t \cdot |\mathcal{P}|$  when modest

Part 5

# **Conclusions**

## Multi-view Diffusion Trajectories (MDTs)

A unified, flexible **framework** for operator-based **multi-view diffusion geometry** through **time-inhomogeneous random walks** intertwining the views

- **Framework:** rich MDT operator space encompassing a number of existing approaches; discrete and convex designs.
- **Theory:** ergodicity, stationary distributions, trajectory-dependent diffusion distances and embeddings via SVD
- **Sampling:** random MDTs as a principled neutral baseline, and as a data-augmentation tool.
- **Learning:** unsupervised trajectory optimization via task-related internal quality metrics
- **MDT in practice:** Results in manifold learning and clustering
- **Empirical finding:** random MDTs outperform sophisticated existing methods

# Thank you!

Questions?

Code is available online

**See also our recent works:**

*Parametrized Power-Iteration Clustering (2026).*

G. Debaussart-Joniec, H. Sevi, M. Jonckheere, and A. Kalogeratos, ICML (*to appear*)

*Generalized Dirichlet Energy and Graph Laplacians for Clustering Directed and Undirected Graphs (2025).*

H. Sevi, G. Debaussart-Joniec, M. Hacini, M. Jonckheere, and A. Kalogeratos, *Under journal review*

This work: *Multi-view Diffusion Geometry via Intertwined Diffusion Trajectories (2025).*

G. Debaussart-Joniec and Argyris Kalogeratos, *Under journal review*

Appendix

# **Supplementary slides**

Reference material for Q&A

# A1. Designs for the operator set $\mathcal{P}$

---

**Discrete designs** – countable  $\mathcal{P}$ :

- Canonical set  $\mathcal{P}_c = \{\mathbf{P}_v\}_{v=1}^V$
- Identity  $\mathbf{I}_N$  – “idle” step
- Uniform rank-one  $\mathbf{\Xi} = \frac{1}{N}\mathbf{1}\mathbf{1}^\top$  – teleportation
- PageRank-style  $\mathbf{P}_{\text{PR},v}(a) = a\mathbf{P}_v + (1-a)\mathbf{\Xi}$
- Smoothed  $\mathbf{P}_v^{t'}$  for small  $t' \in \mathbb{N}^*$

**Continuous designs** –  $\mathcal{P}$  a continuous set:

$$\mathcal{P}_{\text{cvx}} = \left\{ \sum_{v=1}^V a_v \mathbf{P}_v \mid a_v \geq 0, \sum_v a_v = 1 \right\}$$

- Each step is a convex combination of views.
- $V - 1$  degrees of freedom per step under simplex constraint.
- Smooth space  $\Rightarrow$  [gradient-based optimization possible](#).

Naturally generalises the PageRank random walk to multi-view.

## A2. Unification: existing methods are special MDT cases

---

**Encompassed by MDT (with the canonical set  $\mathcal{P}_c$ ):**

- **Alternating Diffusion**:  $\mathbf{W}_i = \mathbf{P}_{(i \bmod V)+1}$ ,  $\mathbf{W}^{(Vt)} = \text{AD at time } t$
- **Integrated Diffusion**:  $\mathbf{W}^{(t_1+t_2)} = \mathbf{P}_1^{t_1} \mathbf{P}_2^{t_2}$   
(ID = AD on iterated operators  $\{\mathbf{P}_v^{t_v}\}$ )
- **PageRank-MDT**:  $\mathbf{W}_i = \sum_v a_{i,v} \mathbf{P}_{\text{PR},v}(\alpha)$   
– interpolates between view-specific and global diffusion

**Not in MDT (use backward operators or larger composite matrices):**

- MVD– builds a  $2N \times 2N$  operator
- CR-DIFF, ■ COM-DIFF– rely on  $\mathbf{P}_v^\top$

Fixed designs optimize over a *tiny slice* of  $\mathcal{W}$ .  
Searching the full space should do at least as well.

## A2. Existing operators in detail

[supplementary]

**Alternating Diffusion** [Katz et al., 2019]:  $\mathbf{Q}_{\text{AD}} = \mathbf{P}_1 \mathbf{P}_2$

**Integrated Diffusion** [Kuchroo et al., 2022]:  $\mathbf{Q}_{\text{ID}} = \mathbf{P}_1^{t_1} \mathbf{P}_2^{t_2}$  (denoise per view, then alternate)

**Multi-View Diffusion** [Lindenbaum et al., 2020]:

$$\mathbf{K}_{\text{MVD}} = \begin{bmatrix} 0 & \mathbf{K}_1 \mathbf{K}_2 \\ \mathbf{K}_2 \mathbf{K}_1 & 0 \end{bmatrix}, \quad \mathbf{Q}_{\text{MVD}} = \mathbf{D}_{\text{MVD}}^{-1} \mathbf{K}_{\text{MVD}} \quad (2N \times 2N)$$

**Cross-Diffusion** [Wang et al., 2012]:

$$\mathbf{Q}_{\text{CR-DIFF},1}^{(t+1)} = \mathbf{P}_1 \mathbf{Q}_{\text{CR-DIFF},2}^{(t)} \mathbf{P}_2^{\text{T}}, \quad \mathbf{Q}_{\text{CR-DIFF},2}^{(t+1)} = \mathbf{P}_2 \mathbf{Q}_{\text{CR-DIFF},1}^{(t)} \mathbf{P}_1^{\text{T}}$$

**Composite Diffusion** [Shnitzer et al., 2018]:

$$\mathbf{Q}_{\text{COM-DIFF},1} = \mathbf{P}_2 \mathbf{P}_1^{\text{T}} + \mathbf{P}_1 \mathbf{P}_2^{\text{T}} \text{ (sym., real eigs.)}, \quad \mathbf{Q}_{\text{COM-DIFF},2} = \mathbf{P}_2 \mathbf{P}_1^{\text{T}} - \mathbf{P}_1 \mathbf{P}_2^{\text{T}} \text{ (anti-sym.)}$$

**Caveat.** Iterating row-stochastic matrices:  $\mathbf{PQ}^{\text{T}}$  is generally *not* stochastic; CR-DIFF and COM-DIFF therefore depart from the random-walk framework.

## A4. Sketch: stability of aperiodicity & irreducibility

[supplementary]

**Claim.** If every  $\mathbf{P} \in \mathcal{P}$  is row-stochastic with strictly positive diagonal and defines an aperiodic, irreducible chain, then any  $\mathbf{W}^{(t)} \in \mathcal{W}$  does too.

**Aperiodicity.** Positive diagonal entries of each  $\mathbf{W}_i$  allow “self-transitions” at every step. Their left-product preserves the existence of 1-step self-loops:

$$[\mathbf{W}^{(t)}]_{ii} \geq \prod_{s=1}^t [\mathbf{W}_s]_{ii} > 0.$$

Hence gcd of return times is 1  $\Rightarrow$  aperiodic.

**Irreducibility.** For each  $\mathbf{W}_s$ , there exists  $\tau_s \in \mathbb{N}$  such that  $[\mathbf{W}_s^{\tau_s}]_{ij} > 0$  for every  $i, j$ . Choosing trajectory segments alternating well-mixing operators preserves reachability in the product. Formally, given any pair  $(i, j)$ , there is a finite path through intermediate states whose probability under  $\mathbf{W}^{(t)}$  is bounded below by a product of strictly positive entries.

**Consequence.**  $\mathbf{W}^{(t)}$  admits a unique stationary  $\pi_t$ ; iterating beyond  $t$  keeps the chain ergodic, hence  $\mathbf{W}^{(t)} \rightarrow \mathbf{1}\pi_\infty^\top$  as  $t \rightarrow \infty$ .

Full proofs in Appendix C of the paper.

## A5. Sketch: SVD-based diffusion-map isometry

[supplementary]

**Claim.** For  $\mathbf{W}^{(t)} \in \mathcal{W}$  with SVD  $\mathbf{W}^{(t)} = \mathbf{U}_t \boldsymbol{\Sigma}_t \mathbf{V}_t^\top$  and  $\boldsymbol{\Psi}^{(t)}(x_i) = e_i^\top \mathbf{U}_t \boldsymbol{\Sigma}_t$ ,

$$\mathcal{D}_{\mathbf{W}^{(t)}}(x_i, x_j)^2 = \left\| \boldsymbol{\Psi}^{(t)}(x_i) - \boldsymbol{\Psi}^{(t)}(x_j) \right\|_2^2.$$

**Sketch.** Write the trajectory distance using rows of  $\mathbf{W}^{(t)}$ :

$$\mathcal{D}_{\mathbf{W}^{(t)}}(x_i, x_j)^2 = (e_i - e_j)^\top \mathbf{W}^{(t)} \boldsymbol{\Pi}^{-1} \mathbf{W}^{(t)\top} (e_i - e_j),$$

where  $\boldsymbol{\Pi} = \text{diag}(\pi_t)$  encodes the stationary weighting.

Substituting the SVD and using  $\mathbf{V}_t^\top \boldsymbol{\Pi}^{-1} \mathbf{V}_t \approx \mathbf{I}$  (under the assumed weighting; see paper for the precise identity):

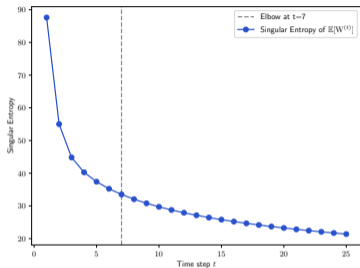
$$\mathcal{D}_{\mathbf{W}^{(t)}}(x_i, x_j)^2 = (e_i - e_j)^\top \mathbf{U}_t \boldsymbol{\Sigma}_t^2 \mathbf{U}_t^\top (e_i - e_j) = \left\| e_i^\top \mathbf{U}_t \boldsymbol{\Sigma}_t - e_j^\top \mathbf{U}_t \boldsymbol{\Sigma}_t \right\|_2^2.$$

The right-hand side is exactly  $\left\| \boldsymbol{\Psi}^{(t)}(x_i) - \boldsymbol{\Psi}^{(t)}(x_j) \right\|_2^2$ . □

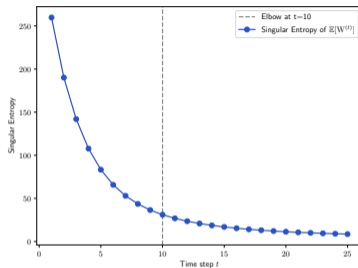
**Why SVD, not eigendecomposition?**  $\mathbf{W}^{(t)}$  is generally non-symmetric (time-inhomogeneous product), so eigendecomposition may be complex-valued; SVD always exists in  $\mathbb{R}$ .

# A5. Singular Entropy curves – time selection

[supplementary]



Helix-A: SE of  $\mathbb{E}_\mu[\mathbf{W}^{(t)}]$  vs.  $t$ . Dashed line: elbow ( $t^*$ ).



K-MvMNIST (noise factor 0.5): SE of  $\mathbb{E}_\mu[\mathbf{W}^{(t)}]$  vs.  $t$ .

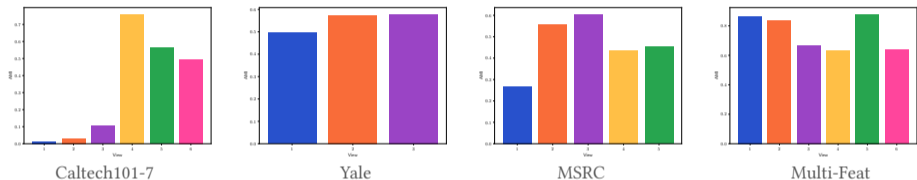
## Reading the curves

- Plateau at large  $t \rightarrow$  approaching rank-one regime; SE  $\rightarrow 0$
- Elbow  $\rightarrow$  best balance between noise reduction and information retention
- Evaluating on the *expected* operator  $\mathbb{E}_\mu[\mathbf{W}^{(t)}]$  removes per-trajectory fluctuations

Elbow detection via KNEED [Satopaa et al., 2011].

## A6. Per-view AMI – the difficulty of each view

[supplementary]

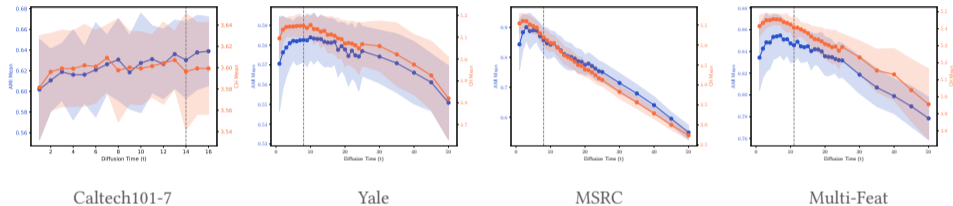


### Findings

- Some datasets have a clearly dominant view (Caltech101-7, Yale)
- Others have views of comparable quality (Multi-Feat) – fusion gains come from *combining*, not from pickin
- Identifying the best view in an unsupervised way is generally infeasible  $\Rightarrow$  a method that adapts to view quality is needed

## A8. Sensitivity to the internal criterion

[supplementary]



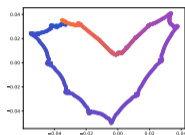
### Findings

- CH is robust but not universally best – DBS or SIL can be preferable for some datasets
- Optimizing one internal criterion does *not* automatically maximize another (or AMI)
- On 100Leaves, the true-label structure disagrees with internal-metric structure  
→ explains the MDT-BEAM miss

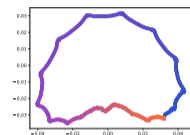
# A10. Detailed Helix embeddings

[supplementary]

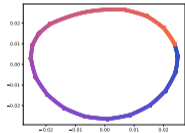
## Helix-A (left col.) – Helix-B (right col.)



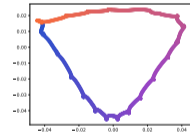
AD



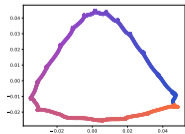
ID



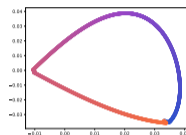
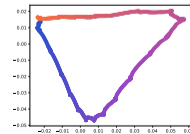
MVD



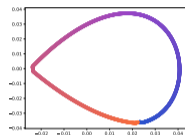
MDT-Csr



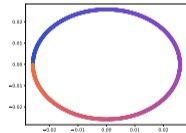
MDT-RAND (two samples)



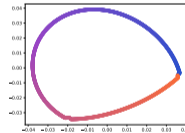
AD



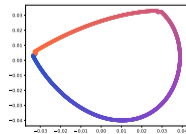
ID



MVD



MDT-Csr



MDT-RAND (two samples)

# References I

---

- B. Afshin-Pour, G.-A. Hossein-Zadeh, S. C. Strother, and H. Soltanian-Zadeh. Enhancing reproducibility of fmri statistical maps using generalized canonical correlation analysis in npairs framework. *NeuroImage*, May 2012.
- R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.
- L. Haghverdi, F. Buettner, and F. J. Theis. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998, 2015.
- O. Katz, R. Talmon, Y.-L. Lo, and H.-T. Wu. Alternating diffusion maps for multimodal data fusion. *Information Fusion*, 45, 2019.
- M. Kuchroo, A. Godavarthi, A. Tong, G. Wolf, and S. Krishnaswamy. Multimodal Data Visualization and Denoising with Integrated Diffusion, 2022.
- A. Kumar, P. Rai, and H. Daume. Co-regularized Multi-view Spectral Clustering. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, 2011.
- O. Lindenbaum, A. Yeredor, M. Salhov, and A. Averbuch. Multi-view diffusion maps. *Information Fusion*, 55:127–149, 2020.
- V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. Finding a “kneedle” in a haystack: Detecting knee points in system behavior. In *International Conference on Distributed Computing Systems Workshops*, 2011.
- T. Shnitzer, M. Ben-Chen, L. Guibas, R. Talmon, and H.-T. Wu. Recovering hidden components in multimodal data with composite diffusion operators, 2018.
- J. Venna and S. Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *International Conference on Artificial Neural Networks*, pages 485–491, 2001.
- B. Wang, J. Jiang, W. Wang, Z.-H. Zhou, and Z. Tu. Unsupervised metric fusion by cross diffusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2997–3004, 2012.