VIS 2022

Dual Space Coupling Model Guided Overlap-Free Scatterplot

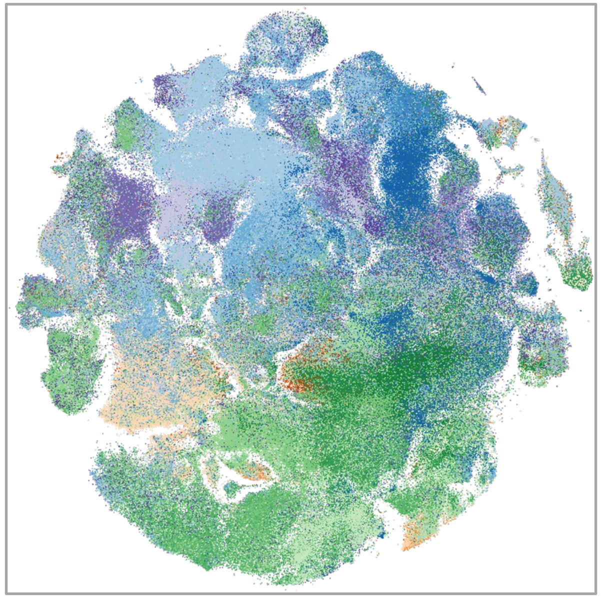
Zeyu Li, Ruizhi Shi, Yan Liu, Shizhuo Long, Ziheng Guo, Shichao Jia, and Jiawan, Zhang
Tianjin University

Agenda

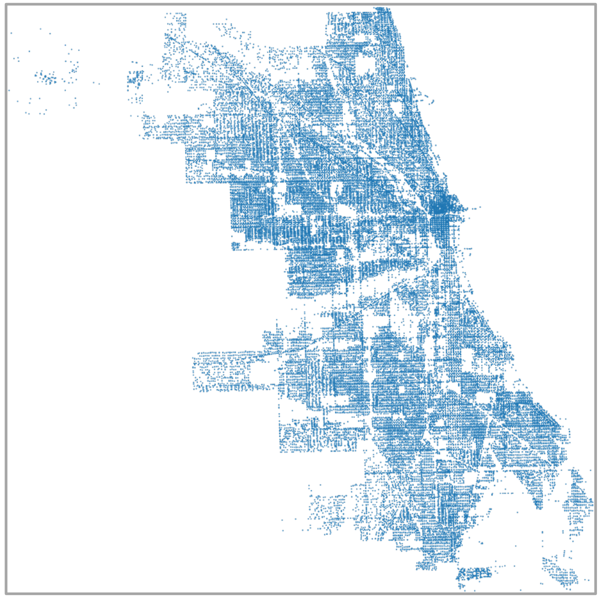
- Motivation
- Previous work
- Dual-space coupling model
- Methods
- Evaluation
- Conclusion

Motivation

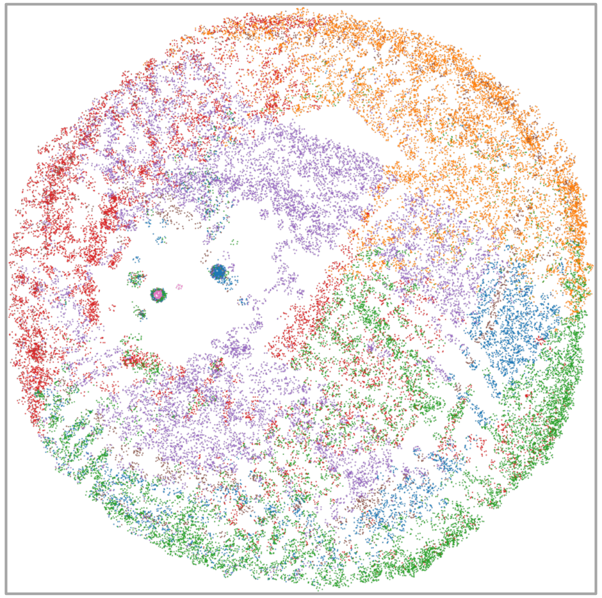
Example scatterplots created by different ways:



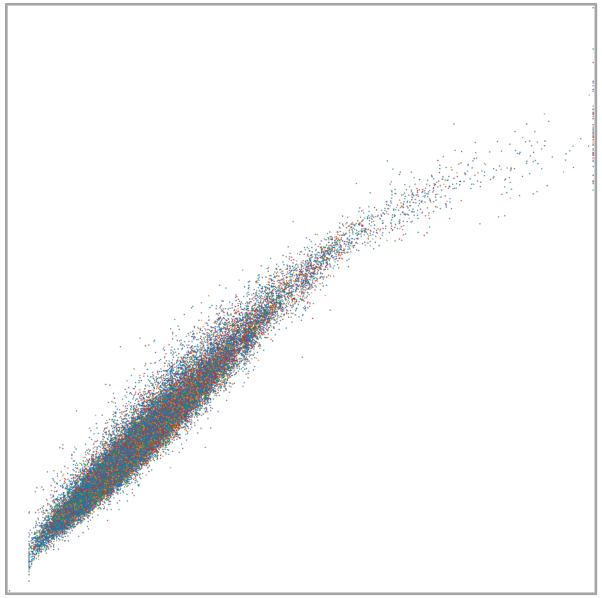
projection results of high-dimensional data



coordinates from geographic space

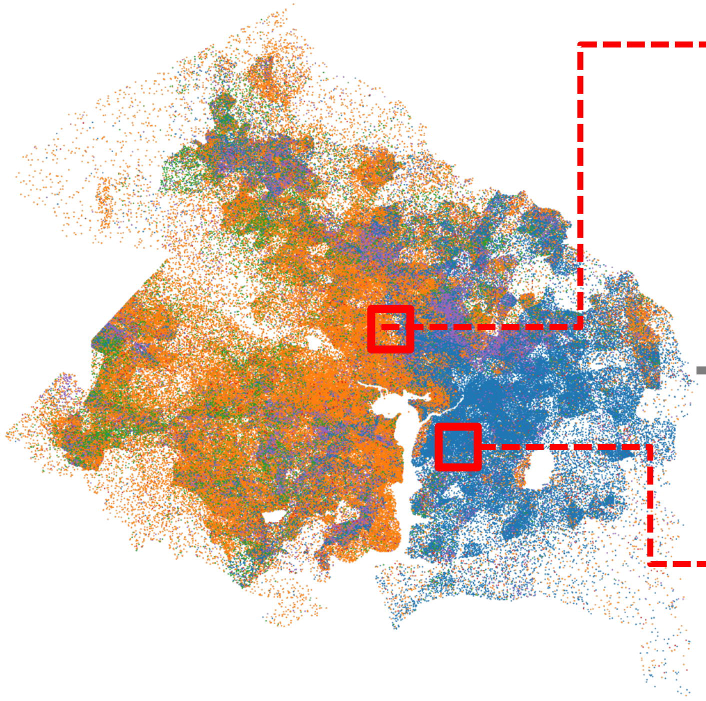


layout results of large-scale graphs

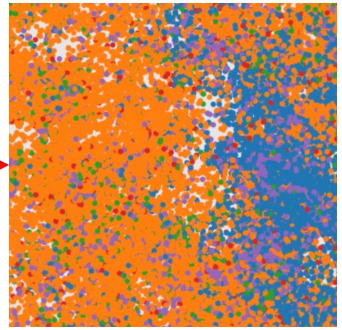


regular scatterplots with two semantic axes

Motivation

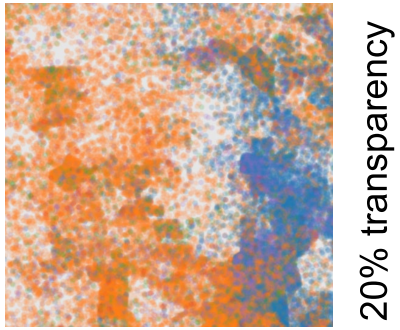


Looks like:



uniform density

In reality:



non-uniform density

Looks like:

99%

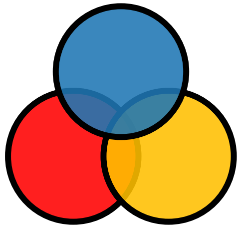
one pure class

In reality:

51% 22%

17% 10%

four mixed classes



Overlaps!

The overdraw problem severely damages visual tasks of scatterplots:

- density perception
- cluster identification
- shape examination
- trend analysis
- outlier identification
- similar data visual inspection

Previous Work

Data Space Methods		Visual Space Methods		Hybrid Methods	
data transformation	✔	data transformation	✖	data transformation	✔
view transformation	✖	view transformation	✔	view transformation	✔

Previous Work

Data Space Methods

data transformation
view transformation

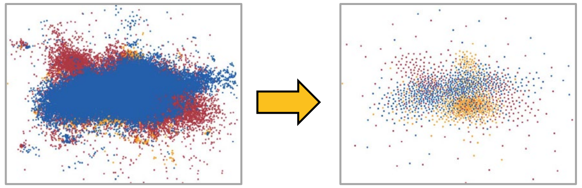
Visual Space Methods

data transformation
view transformation

Hybrid Methods

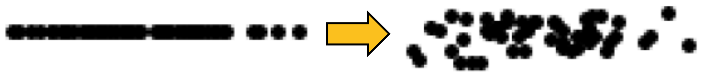
data transformation
view transformation

1. Data sampling or aggregation



- ineludible data loss
- cannot eliminate overlaps
- break one-to-one correspondence

2. Jitter



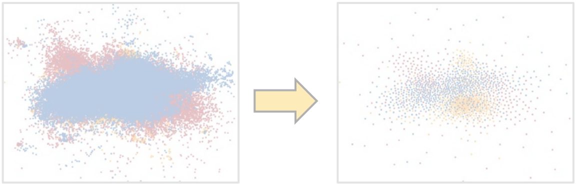
- cannot eliminate overlaps
- may disturb data features

Previous Work

Data Space Methods

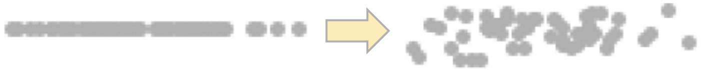
data transformation ✓
view transformation ✗

1. Data sampling or aggregation



- ineludible data loss and bias
- cannot eliminate overlaps
- break one-to-one correspondence

2. Jitter

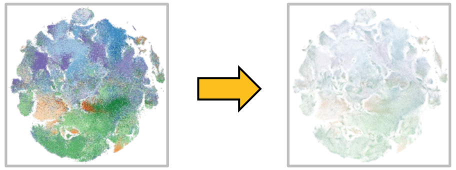


- cannot eliminate overlaps
- may disturb data features

Visual Space Methods

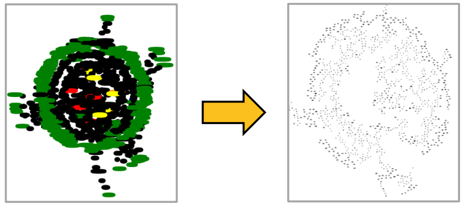
data transformation ✗
view transformation ✓

1. Appearance adjustment



- time-consuming
- color blending

2. Node dispersion

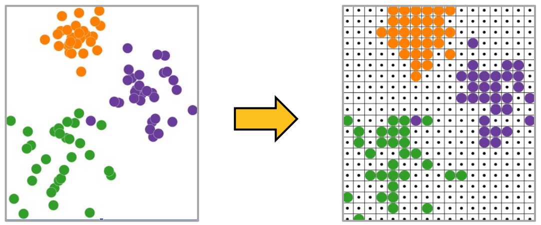


- poor scalability
- severe distortion
- cannot eliminate overlaps

Hybrid Methods

data transformation ✓
view transformation ✓

3. Subspace mapping methods



- shape and density distortion in high density regions

Previous Work

Data Space Methods

data transformation ✓
view transformation ✗

1. Data sampling or aggregation

- ineludible data loss and bias
- cannot eliminate overlaps
- break one-to-one correspondence

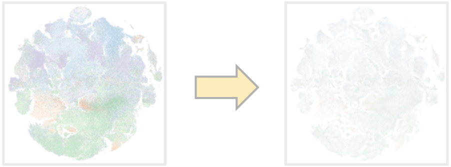
2. Jitter

- cannot eliminate overlaps
- may disturb data features
- cannot eliminate overlaps

Visual Space Methods

data transformation ✗
view transformation ✓

1. Appearance adjustment



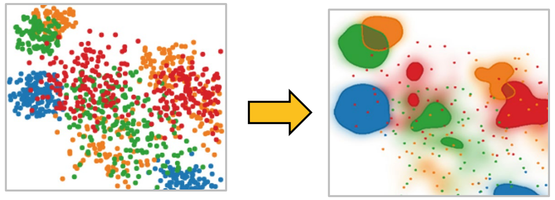
- time-consuming
- color blending

2. Node dispersion

- poor scalability
- severe distortion
- cannot eliminate overlaps

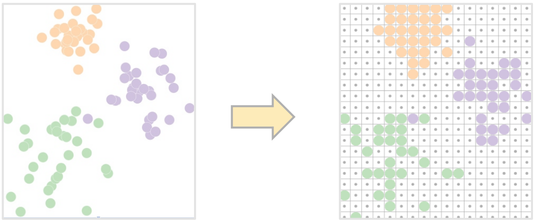
Hybrid Methods

data transformation ✓
view transformation ✓



- ineludible data loss
- break one-to-one correspondence

3. Subspace mapping methods



- shape and density distortion in high density regions

Dual-space coupling model - four criteria and a goal

$DS = \{x,\ y\}$ data set in data space, each data point is scale-free and immaterial

$NS = \{x,\ y,\ r\}$ visual node set in visual space, each visual node has a measurable radius

Four criteria that the overdraw solution should consider:

C1. Mutual Exclusion of Data Points: ------------ $\forall d_1,\ d_2 \in DS,\ \ d_1 \cap_{\mathbf{D}} d_2 = \emptyset$

C2. Mutual Exclusion of Visual Nodes: ----------- $\forall n_1,\ n_2 \in NS,\ \ n_1 \cap_{\mathbf{V}} n_2 = \emptyset$

C3. Data-Visual Space Bijection: -- $DS \leftrightarrow NS$

C4. Data-Visual Space Distribution Consistency: ------------- $F_{\mathbf{V}}(NS) \sim F_{\mathbf{D}}(DS)$

The **goal** of a desired overdraw solution:

$$argmax(similarity(F_{\mathbf{V}}(NS),\ F_{\mathbf{D}}(DS))),\ s.t.\ C1,\ C2,\ C3$$

$*$ $C1$ is not mandatory

Dual-space coupling model - metrics of distribution consistency

Local features:

- KNN preservation
- Displacement minimization

Related visual tasks:

- outlier identification
- similar data visual inspection

Global features:

- Shape preservation
- Density preservation

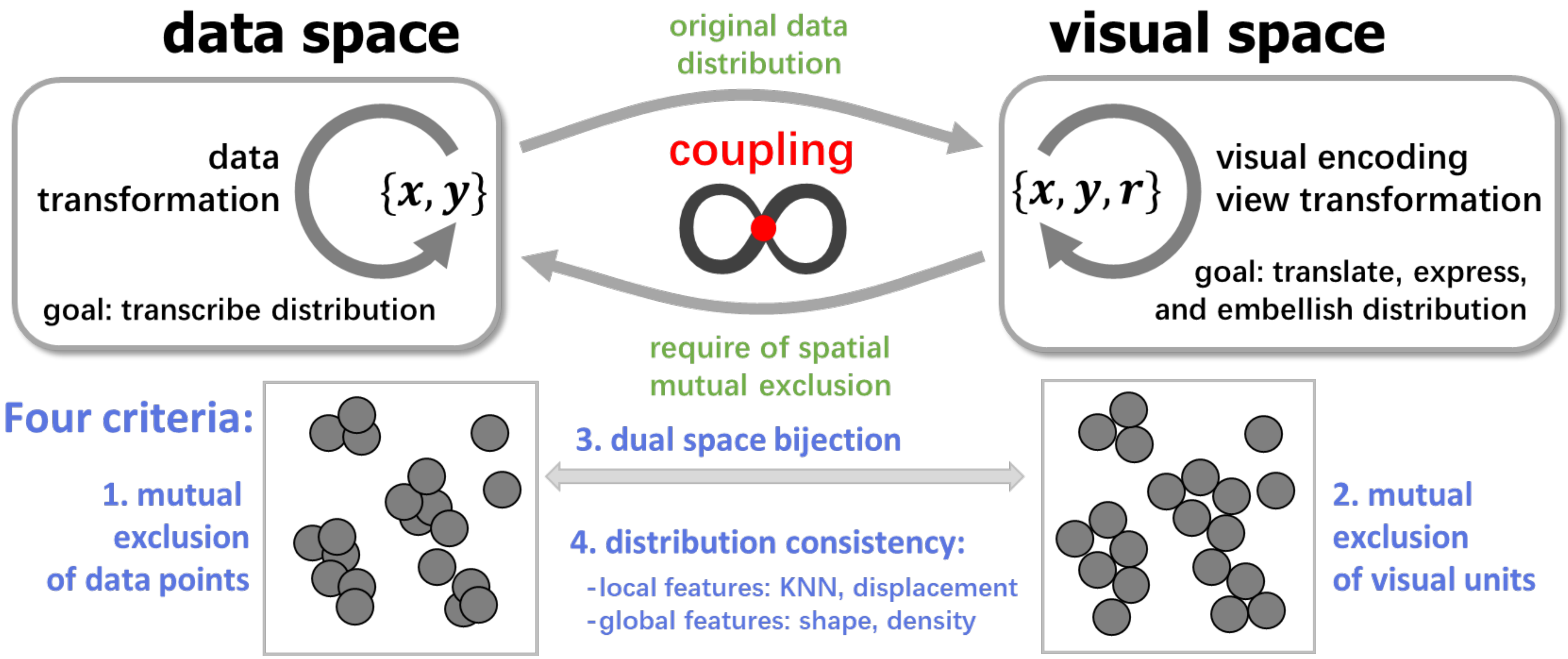
- density perception
- cluster identification
- shape examination
- trend analysis

An individual comprehensive metric:

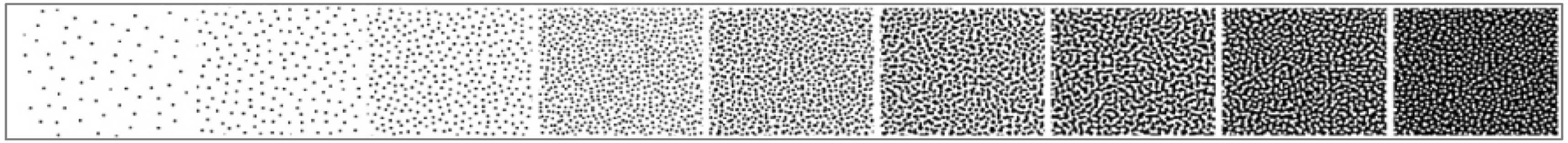
- Overall similarity

average similarity observed from multiple angles

Dual-space coupling model - overview



Methods - core idea and three key questions



FM halftoning

Image stippling

The **core idea** to reconstruct density distribution:

- simulate density by controlling the quantity of visual nodes in local area
- hypothesis: the filling rate of colored pixel ∝ perceived density

Three key questions raised by the core idea:

Q1. How to generate a set of circles that record the data distribution intactly?
Essence: **transcribe** the data distribution from data space to visual space

Q2. How to layout the circles to present the recorded distribution without overlaps?
Essence: **translate** the transcribed distribution into visual space

Q3. How to ensure no overlap occurs during necessary radius configuration?
Essence: **express** and **embellish** the distribution in visual space

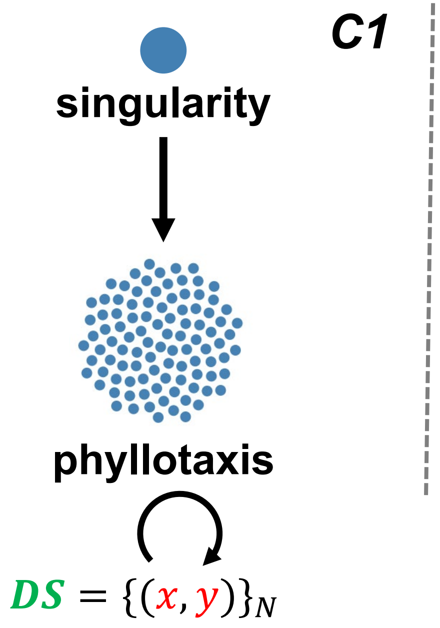
Methods - pipeline

	solution to Q1	solution to Q2	solution to Q3
Purpose	generate a set of circles that record the data distribution intactly	layout the circles without overlaps to present the recorded distribution	ensure no overlap occurs during necessary radius configuration
Essence	**Transcribe** the data distribution from data space to visual space	**Translate** the transcribed distribution into visual space	**Express** and **embellish** the distribution in visual space
Operation	data transformation	view transformation	visual encoding configuration
Method	**DistributionTranscriptor**	**PolarPacking**	**radius adjustment tool $f_{r_{draw}}$**
Satisfied criteria	*C3*: Bijection *C1*: mutual exclusion of data points	*C2*: mutual exclusion of visual nodes *C4*: distribution consistency	

Methods - geometry-based data distribution transcription

Purpose	Operation	Method	Satisfied criteria
generate a set of circles that record the data distribution intactly	data transformation	**DistributionTranscriptor**	*C3*: Bijection *C1*: mutual exclusion of data points

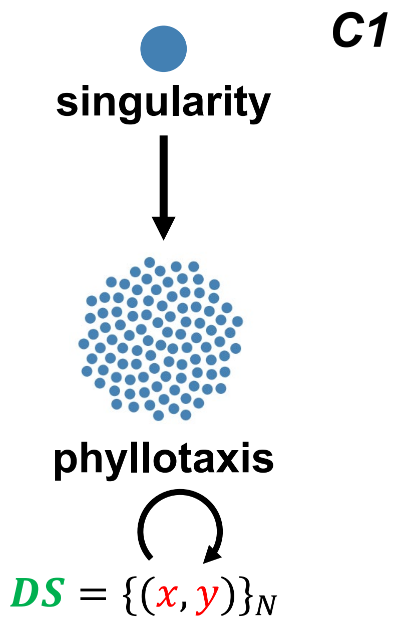
1. singularity spread



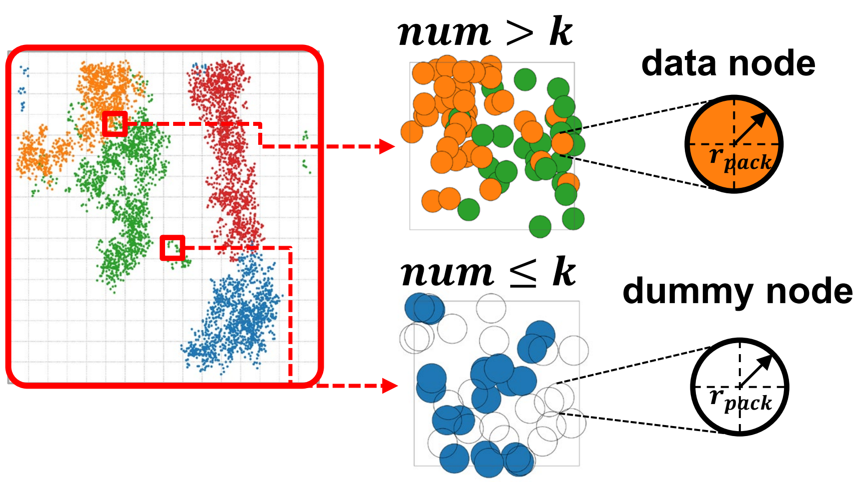
Methods - geometry-based data distribution transcription

Purpose	Operation	Method	Satisfied criteria
generate a set of circles that record the data distribution intactly	data transformation	DistributionTranscriptor	*C3*: Bijection *C1*: mutual exclusion of data points

1. singularity spread

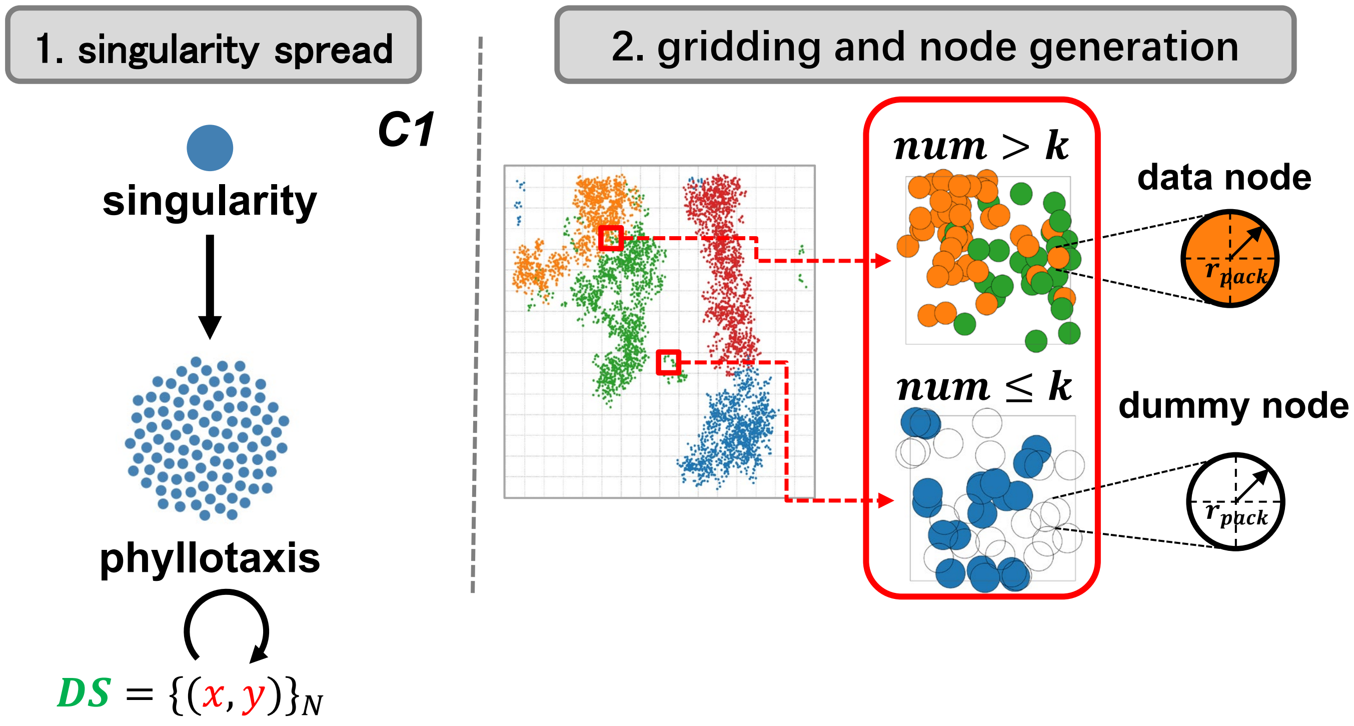


2. gridding and node generation



Methods - geometry-based data distribution transcription

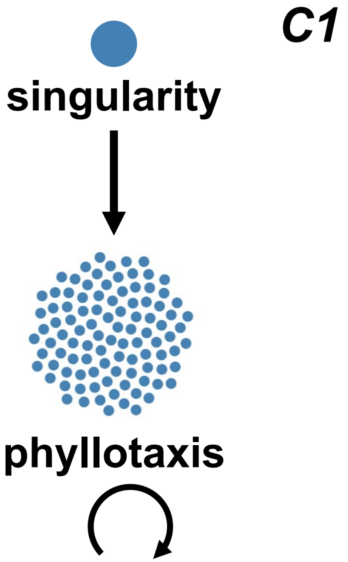
Purpose	Operation	Method	Satisfied criteria
generate a set of circles that record the data distribution intactly	data transformation	DistributionTranscriptor	*C3*: Bijection *C1*: mutual exclusion of data points



Methods - geometry-based data distribution transcription

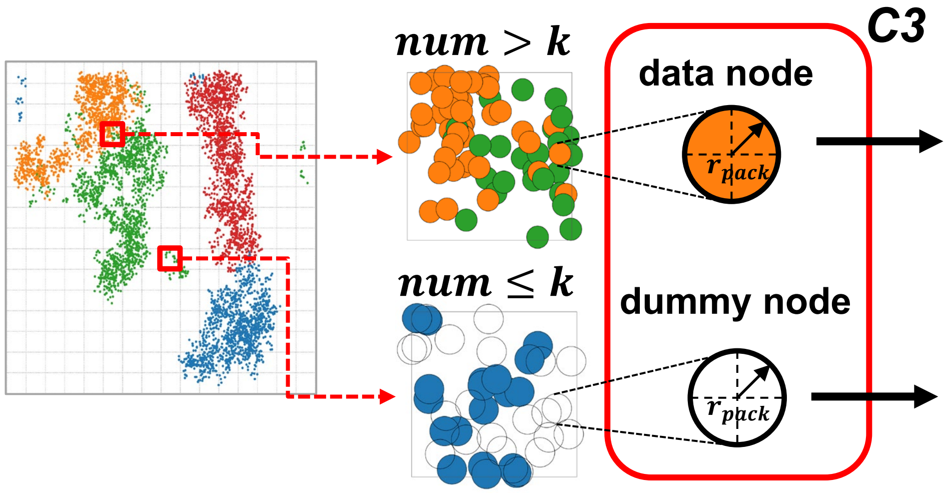
Purpose	Operation	Method	Satisfied criteria
generate a set of circles that record the data distribution intactly	data transformation	DistributionTranscriptor	*C3*: Bijection *C1*: mutual exclusion of data points

1. singularity spread



$DS = \{(x, y)\}_N$

2. gridding and node generation



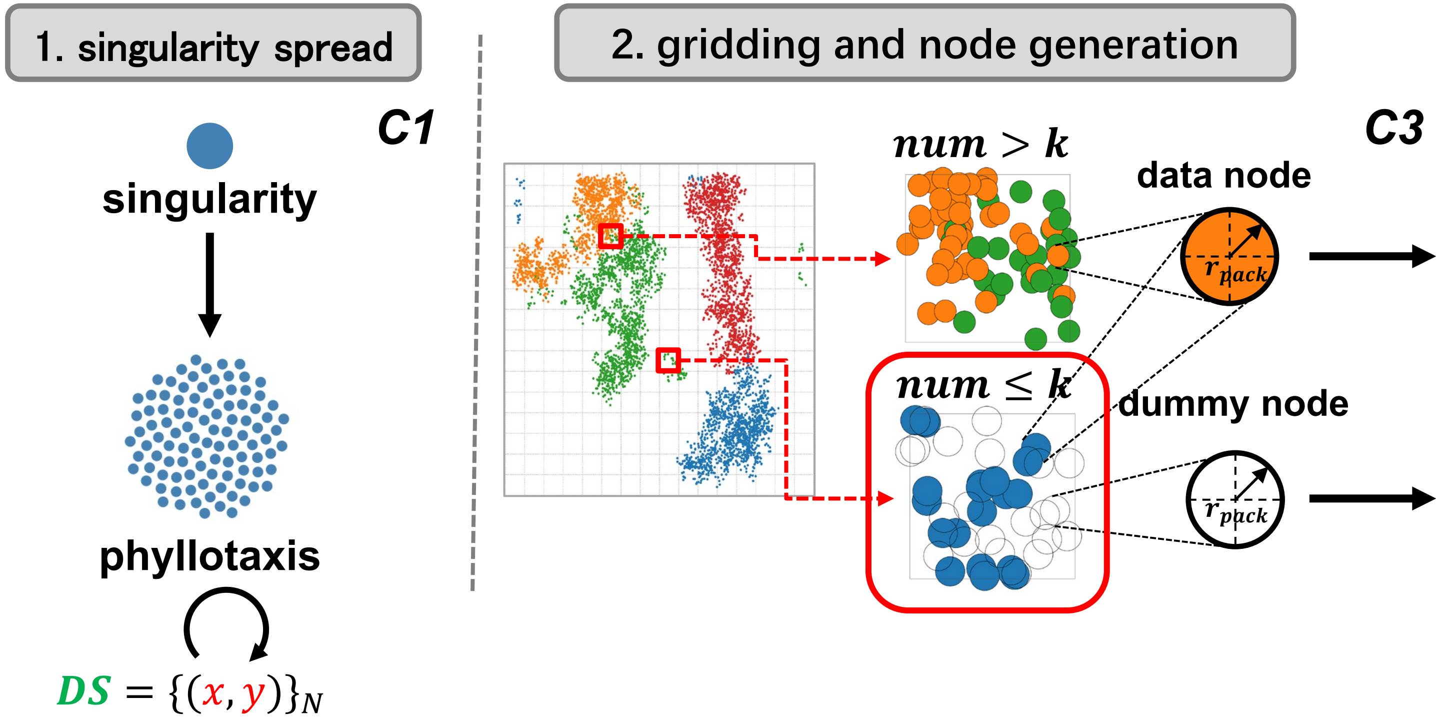
- Each corresponds to a data point.
- Coordinates inherit from the data points.

- Acting as **placeholders** in sparse grids.
- Coordinates are randomly sampled from the grid.

$$density \leftarrow \frac{num}{num_{\max}} \in [0, 1]$$

Methods - geometry-based data distribution transcription

Purpose	Operation	Method	Satisfied criteria
generate a set of circles that record the data distribution intactly	data transformation	DistributionTranscriptor	*C3*: Bijection *C1*: mutual exclusion of data points



Each non-sparse grid includes:

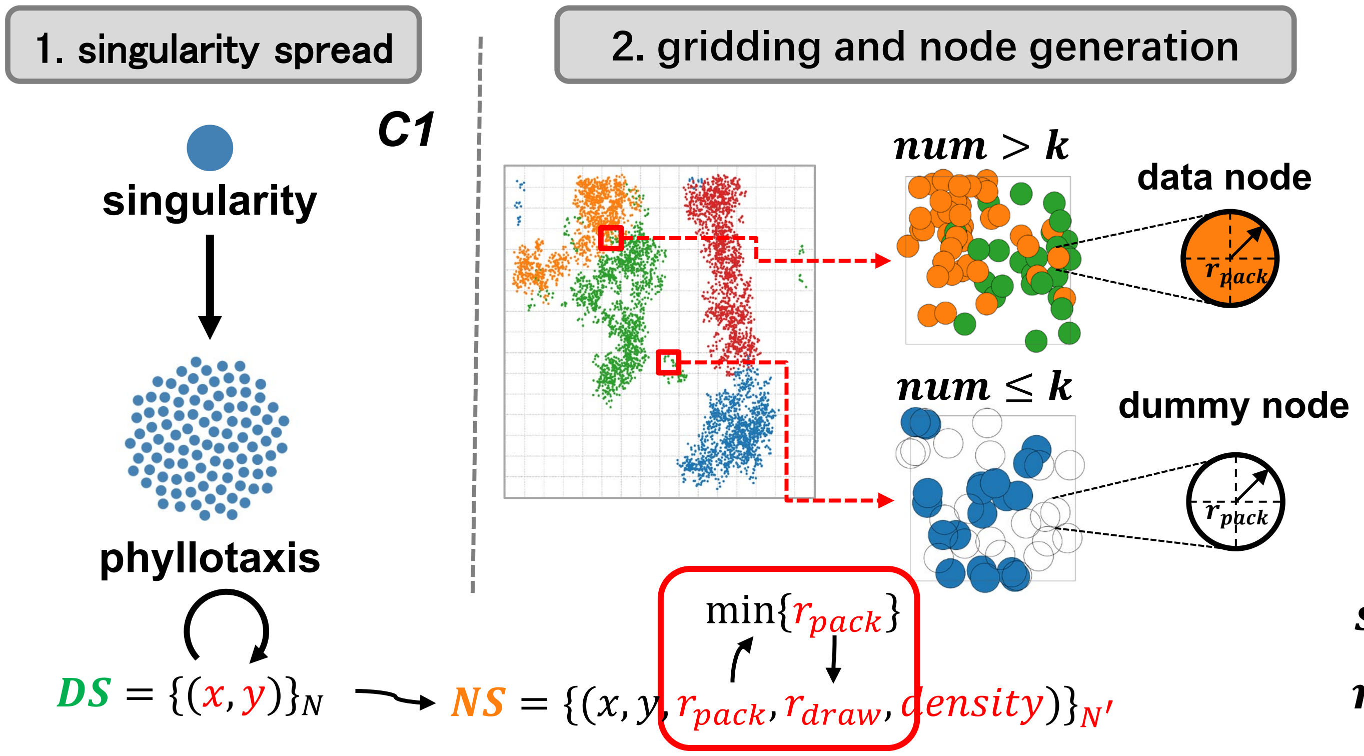
num data nodes $\quad num > k$

Each sparse grid includes:

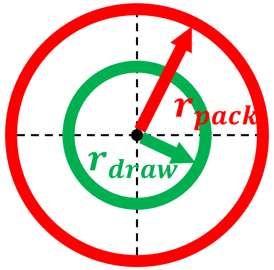
$$\left\{ \begin{array}{l} k - num \text{ dummy nodes} \\ num \text{ data nodes} \end{array} \right. \quad num \leq k$$

Methods - geometry-based data distribution transcription

Purpose	Operation	Method	Satisfied criteria
generate a set of circles that record the data distribution intactly	data transformation	DistributionTranscriptor	*C3*: Bijection *C1*: mutual exclusion of data points



node



r_{pack} depicts occupied space

r_{draw} is rendering radius

$$r_{pack} \leftarrow \sqrt{\frac{size^2}{\pi \times \max(k,\ num)}}$$

$size$ the length of the side of a grid

$max(k, num)$ the number of nodes in a grid

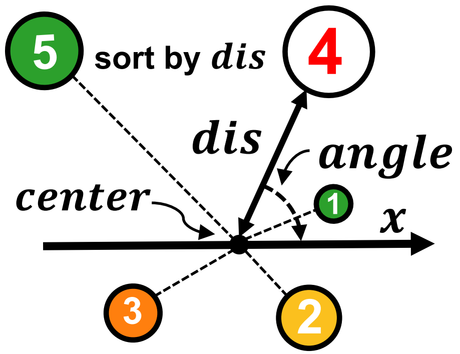
Methods - pipeline

	solution to Q1	solution to Q2	solution to Q3
Purpose	generate a set of circles that record the data distribution intactly	layout the circles without overlaps to present the recorded distribution	ensure no overlap occurs during necessary radius configuration
Essence	**Transcribe** the data distribution from data space to visual space	**Translate** the transcribed distribution into visual space	**Express** and **embellish** the distribution in visual space
Operation	data transformation	view transformation	visual encoding configuration
Method	**DistributionTranscriptor**	**PolarPacking**	**radius adjustment tool $f_{r_{draw}}$**
Satisfied criteria	*C3*: Bijection *C1*: mutual exclusion of data points	*C2*: mutual exclusion of visual nodes *C4*: distribution consistency	

Methods - spatial mutual exclusion guided view transformation

Purpose	Operation	Method	Satisfied criteria
layout the circles without overlaps to present the recorded distribution	view transformation	PolarPacking	*C2: mutual exclusion of visual nodes* *C4: distribution consistency*

1. build polar coordinates

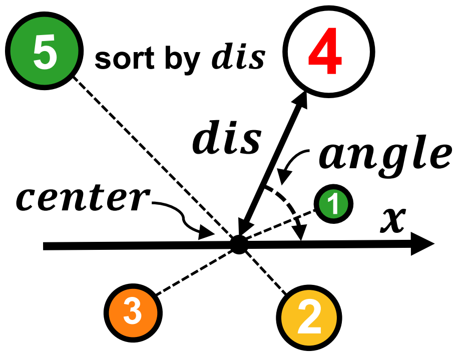


$$NS' = \{(x, y, r_{pack}, r_{draw}, density, dis, angle)\}_{N'}$$

Methods - spatial mutual exclusion guided view transformation

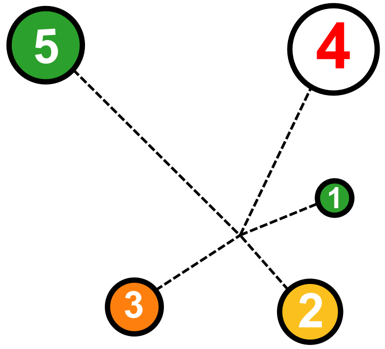
Purpose	Operation	Method	Satisfied criteria
layout the circles without overlaps to present the recorded distribution	view transformation	**PolarPacking**	*C2: mutual exclusion of visual nodes* *C4: distribution consistency*

1. build polar coordinates



$$NS' = \{(x, y, r_{pack}, r_{draw}, density, dis, angle)\}_{N'}$$

2. Pack nodes one by one



$$NS' = \{(x, y, r_{pack}, r_{draw}, density)\}_{N'}$$

Methods - spatial mutual exclusion guided view transformation

Purpose

layout the circles without overlaps to present the recorded distribution

Operation

view transformation

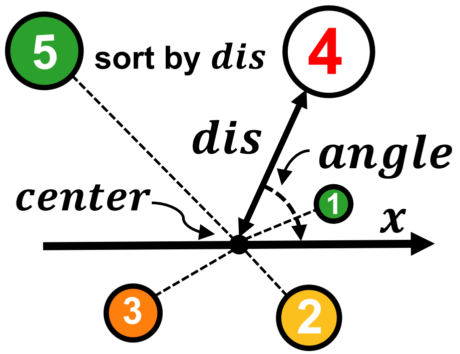
Method

PolarPacking

Satisfied criteria

C2: mutual exclusion of visual nodes
C4: distribution consistency

1. build polar coordinates



$$NS' = \{(x, y, r_{pack}, r_{draw}, density, dis, angle)\}_{N'}$$

2. Pack nodes one by one

$$NS' = \{(x, y, r_{pack}, r_{draw}, density)\}_{N'}$$

3. filter out dummy nodes

C4

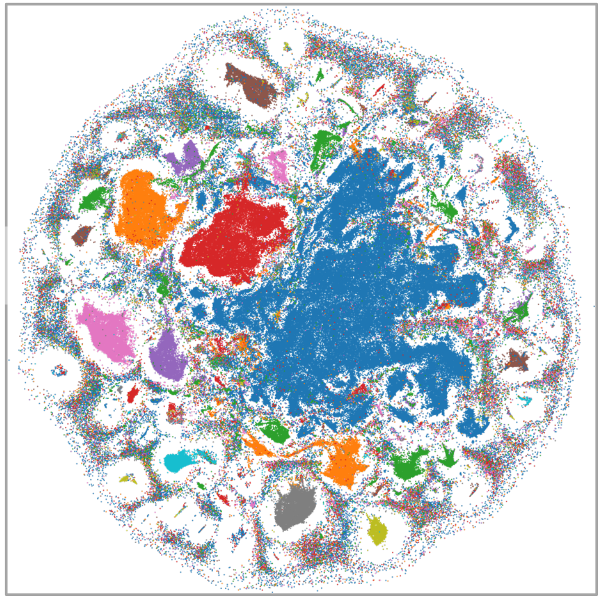
$$NS' = \{(x, y, r_{pack}, r_{draw}, density)\}_{N}$$

Methods - spatial mutual exclusion guided view transformation

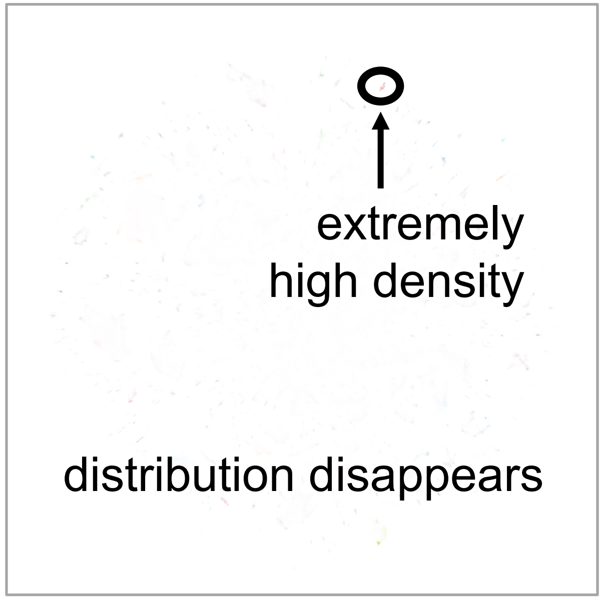
Purpose	Operation	Method	Satisfied criteria
ensure no overlap occurs during necessary radius configuration	visual encoding configuration	**radius adjustment tool** $f_{r_{draw}}$	None

default $r_{draw} = \min\{r_{pack}\}$

Original scatterplot



Our default results



high dynamic range (**HDR**) dataset

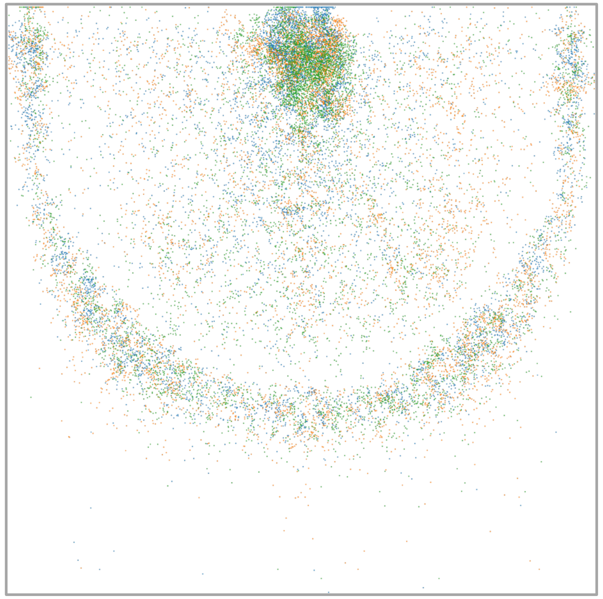
low contrast issue

Methods - spatial mutual exclusion guided view transformation

Purpose	Operation	Method	Satisfied criteria
ensure no overlap occurs during necessary radius configuration	visual encoding configuration	**radius adjustment tool** $f_{r_{draw}}$	None

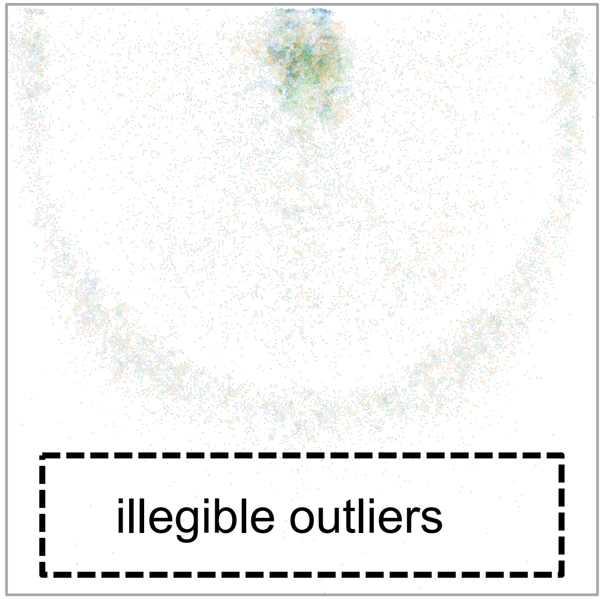
default $r_{draw} = \min\{r_{pack}\}$

Original scatterplot



dataset with **outliers**

Our default results



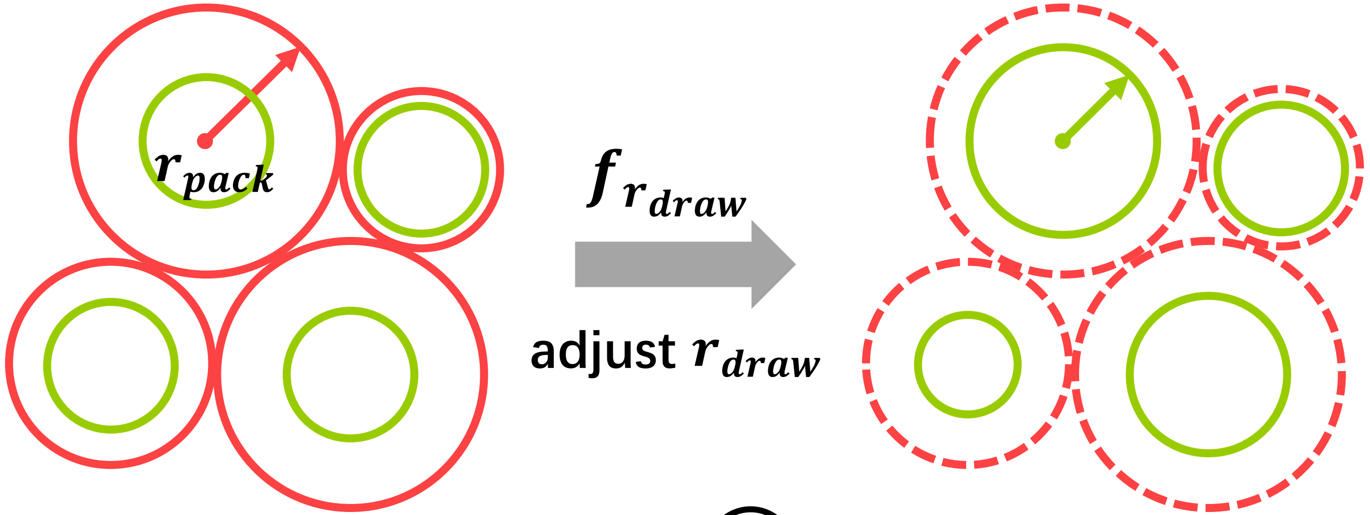
outlier invisible issue

Methods - pipeline

	solution to Q1	solution to Q2	solution to Q3
Purpose	generate a set of circles that record the data distribution intactly	layout the circles without overlaps to present the recorded distribution	ensure no overlap occurs during necessary radius configuration
Essence	**Transcribe** the data distribution from data space to visual space	**Translate** the transcribed distribution into visual space	**Express** and **embellish** the distribution in visual space
Operation	data transformation	view transformation	visual encoding configuration
Method	**DistributionTranscriptor**	**PolarPacking**	**radius adjustment tool $f_{r_{draw}}$**
Satisfied criteria	$C3$: Bijection $C1$: mutual exclusion of data points	$C2$: mutual exclusion of visual nodes $C4$: distribution consistency	

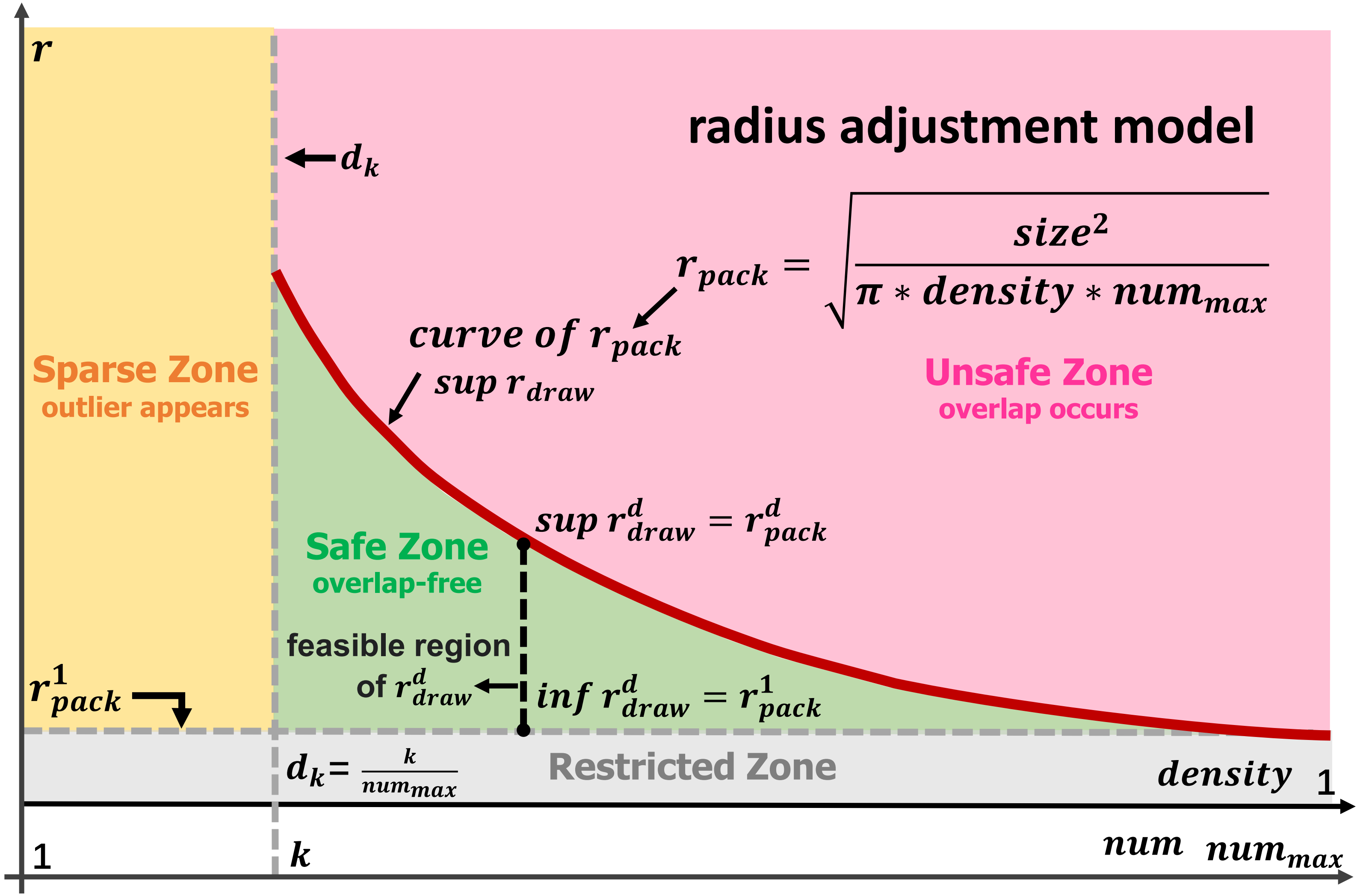
Methods - overlap-free oriented visual encoding configuration

Purpose	Operation	Method	Satisfied criteria
ensure no overlap occurs during necessary radius configuration	visual encoding configuration	**radius adjustment tool** $f_{r_{draw}}$	None

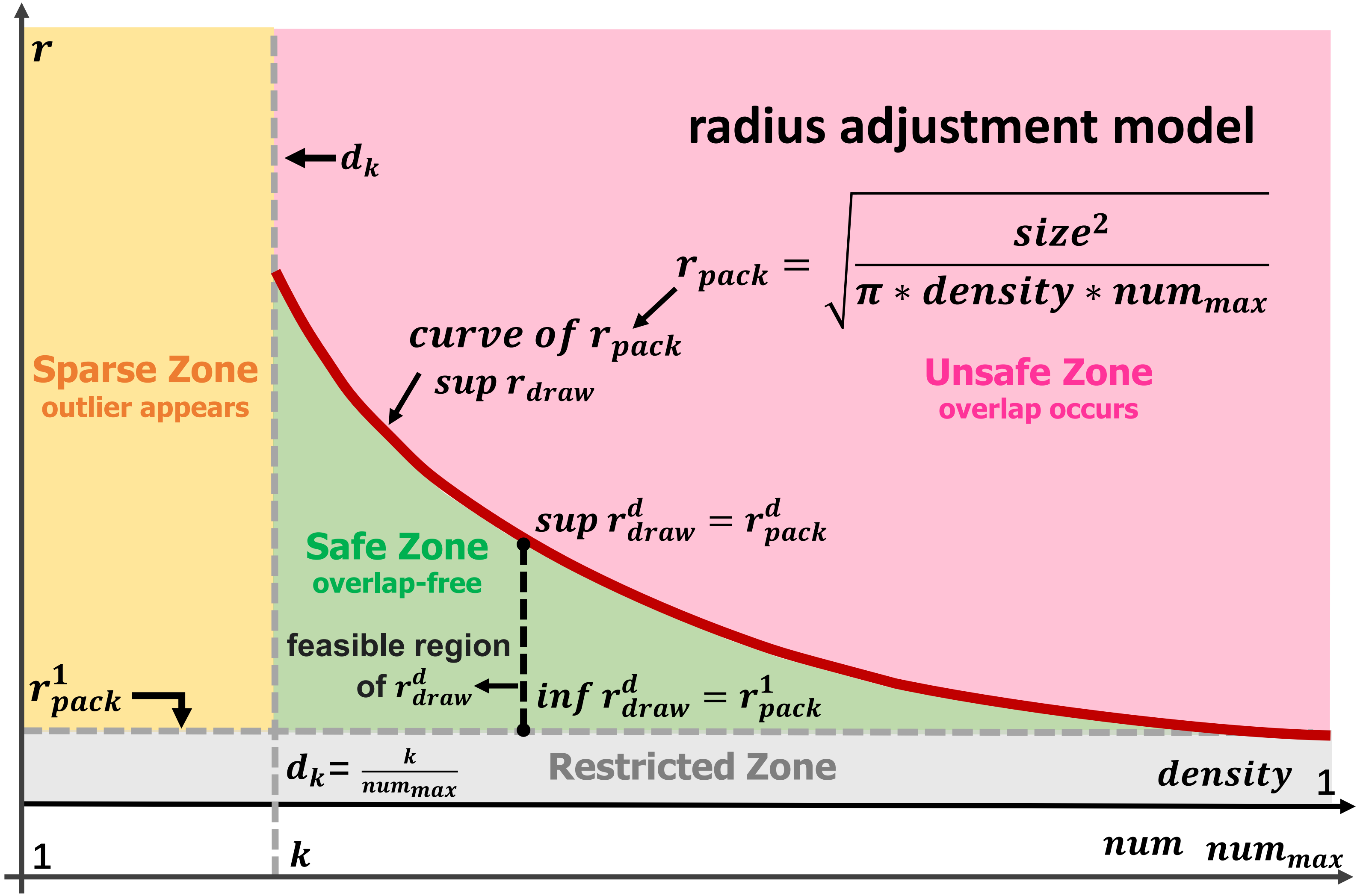


$$NS' = \{(x, y, r_{pack}, r_{draw}, density)\}_N$$

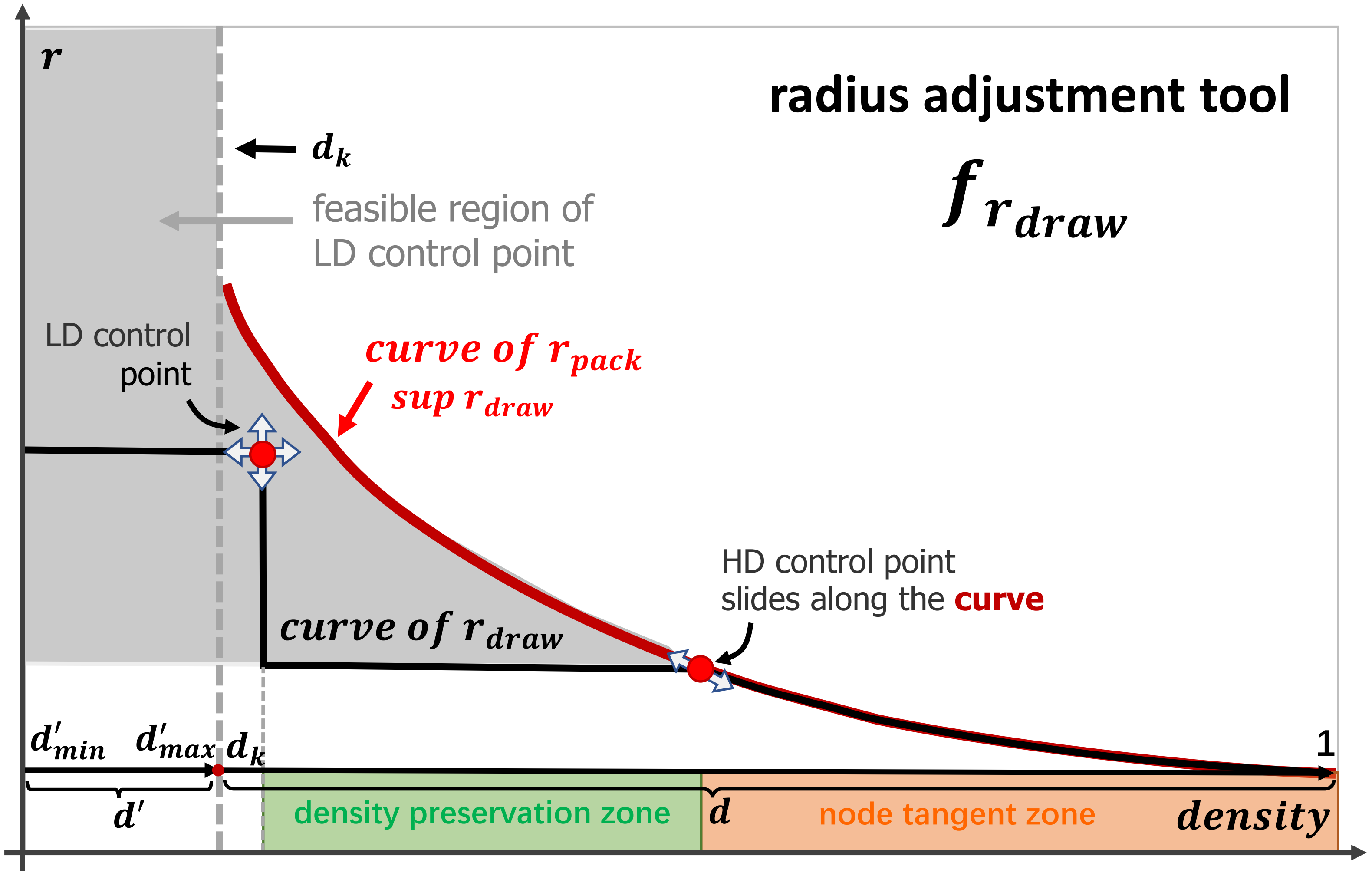
Methods - overlap-free oriented visual encoding configuration



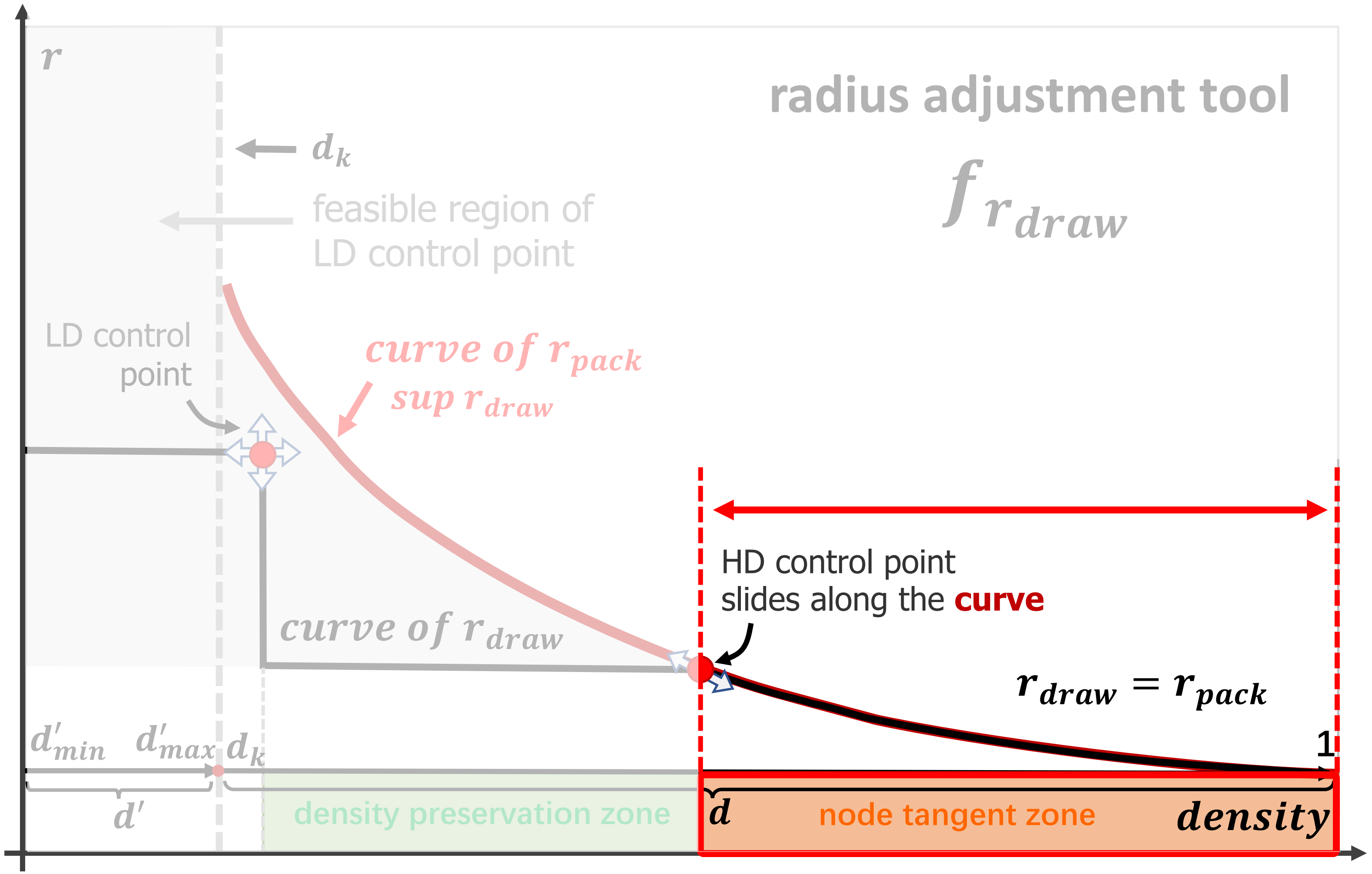
Methods - overlap-free oriented visual encoding configuration



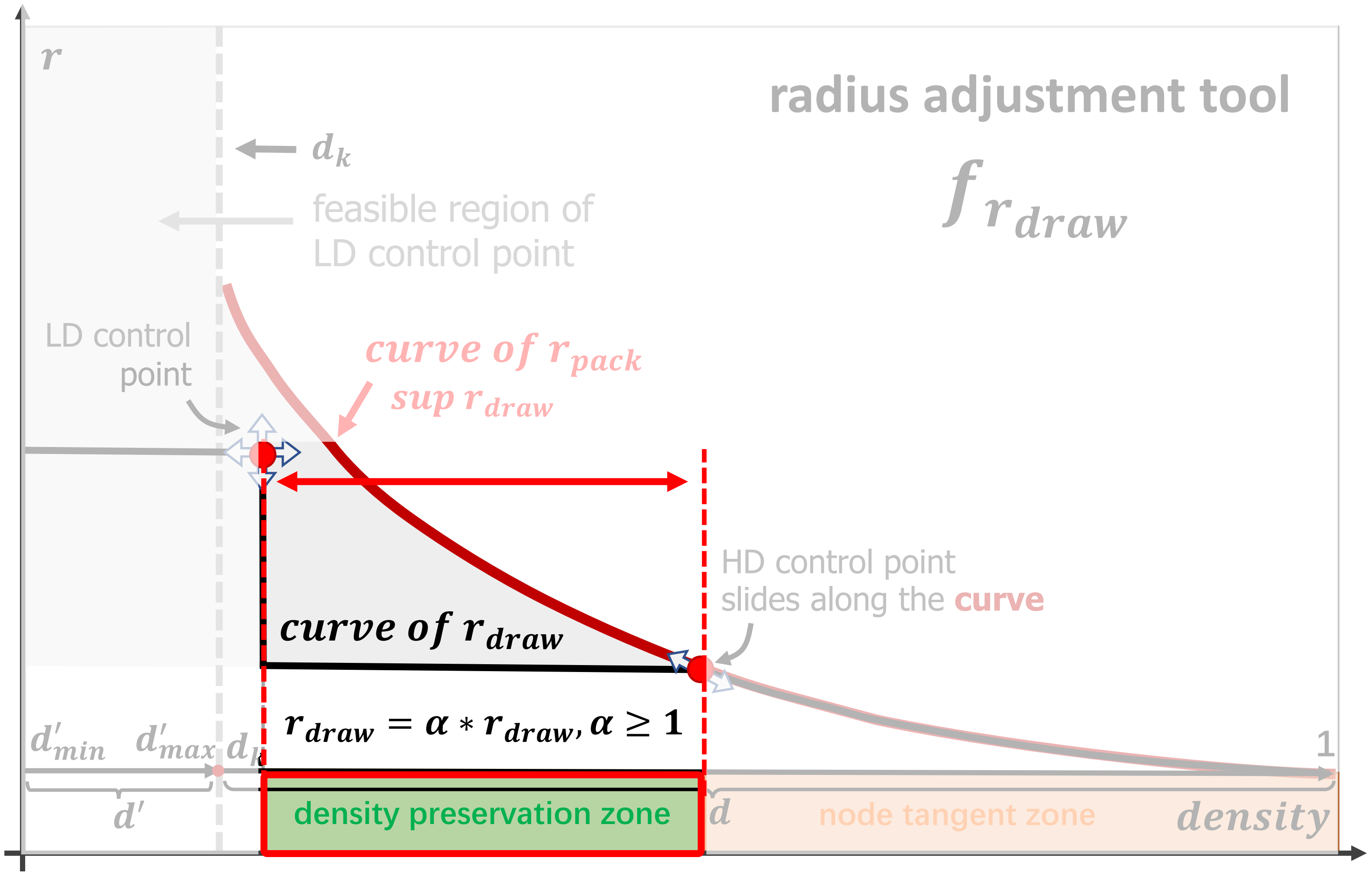
Methods - overlap-free oriented visual encoding configuration



Methods - overlap-free oriented visual encoding configuration



Methods - overlap-free oriented visual encoding configuration

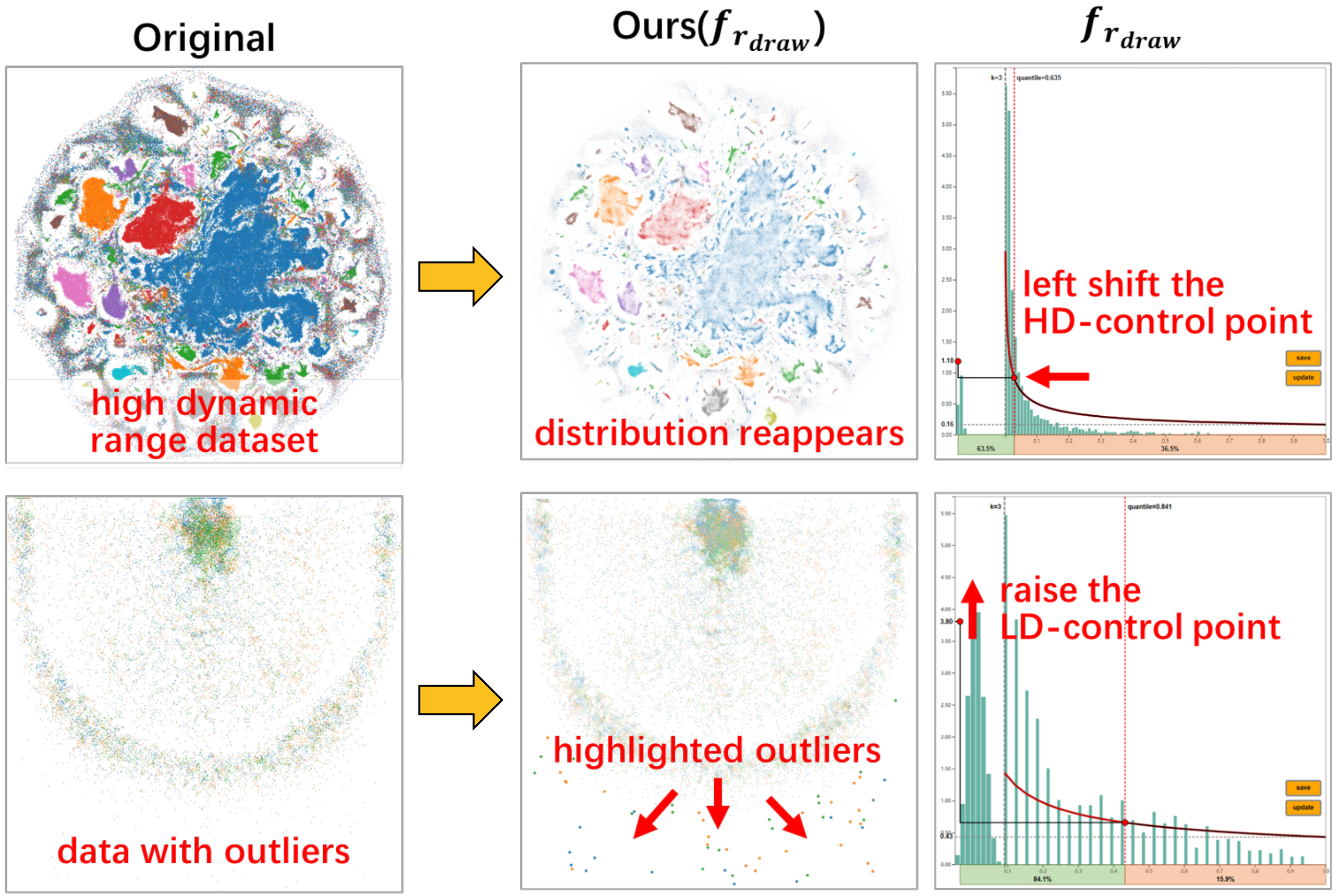


Methods - overlap-free oriented visual encoding configuration

Two examples of applying our $f_{r_{draw}}$ to **improve the visual quality** of a scatterplot.

Solve ***low contrast*** issue faced by HDR datasets by moving the HD-control point to the left

Solve ***outlier invisible*** issue by raising the LD-control point

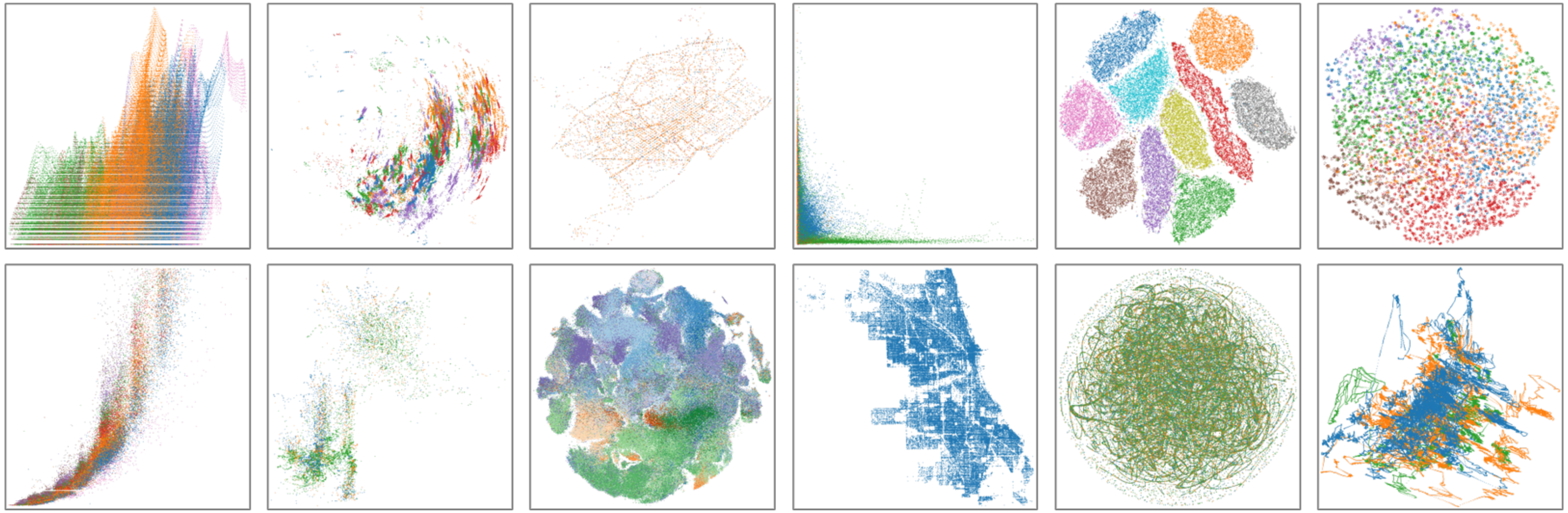


Evaluation - quantitative evaluation

- Competing Algorithms
  - node dispersion methods: *PFS', PRISM, Gtree,* and *RWordle-L* ⟶ Sampled3k
  - subspace-mapping methods: *HaGrid* and *DGrid* ⟶ Full datasets

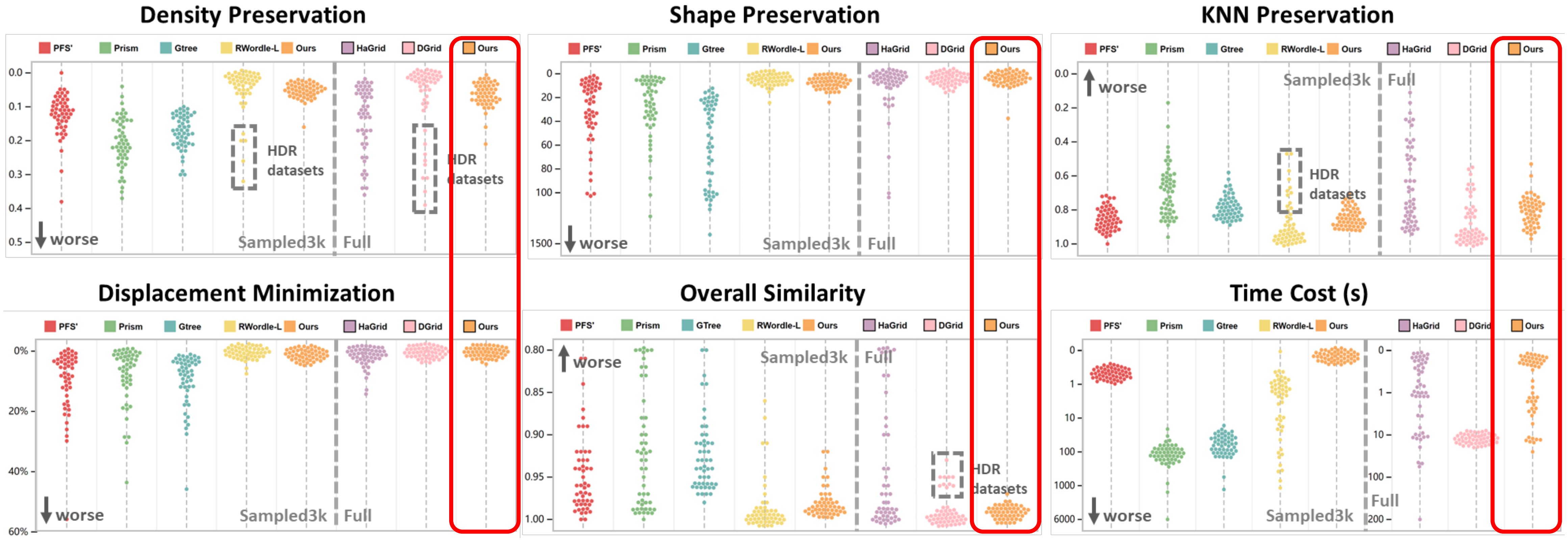
- Datasets : 50 real-world datasets, data scale ranges from 4k to 1M

  12 example datasets:



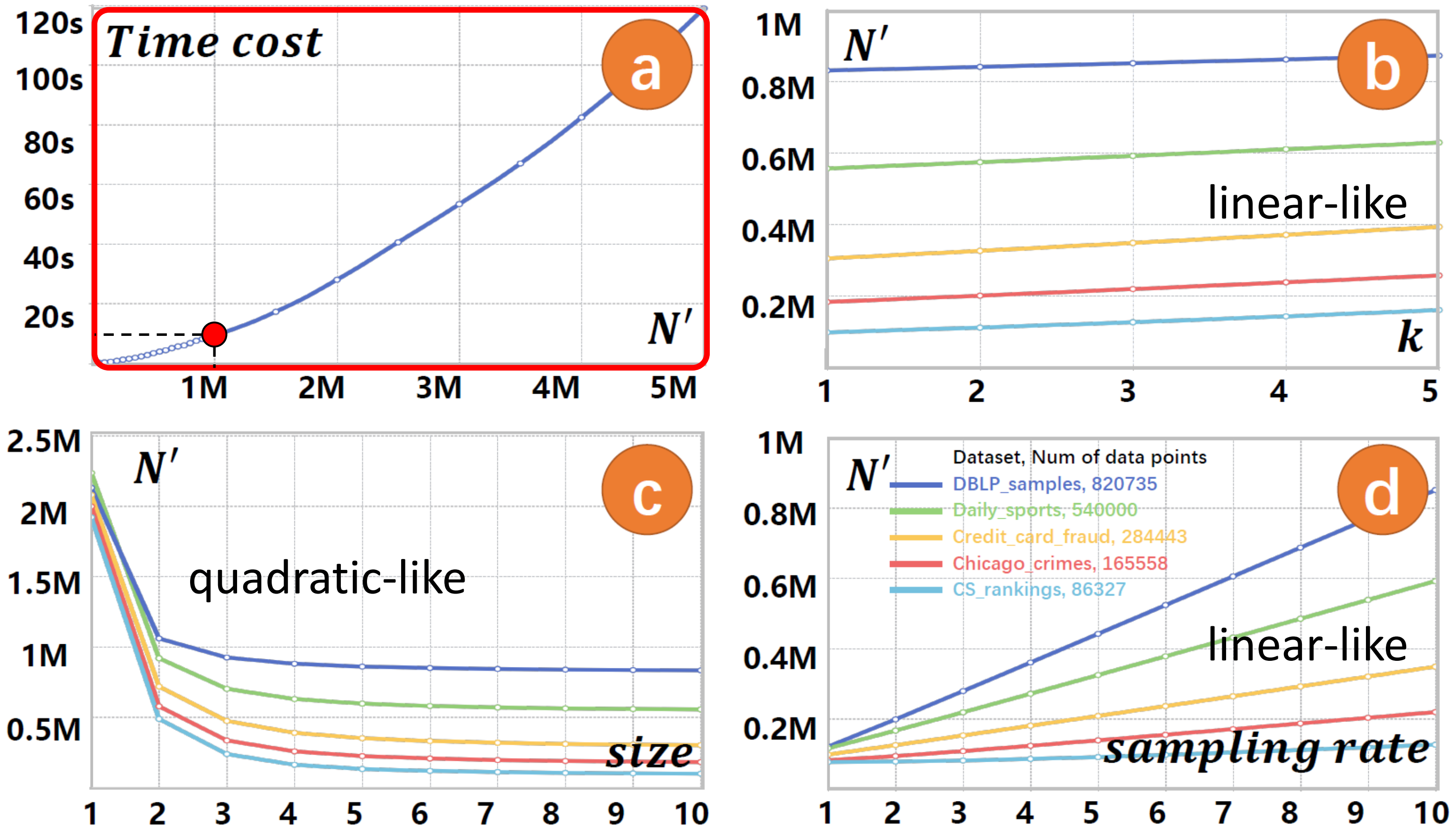
Evaluation - quantitative evaluation

- Our method achieves the best or near the best scores on all metrics compared with the state-of-the-art algorithms.
- Our method takes great advantage on computational efficiency (average time cost: 1/4.6 of Hagrid, 1/47.6 of DGrid).
- Our method presents strong adaptability to high dynamic range(HDR) datasets.



Evaluation - quantitative evaluation

Impact of parameters on time cost:



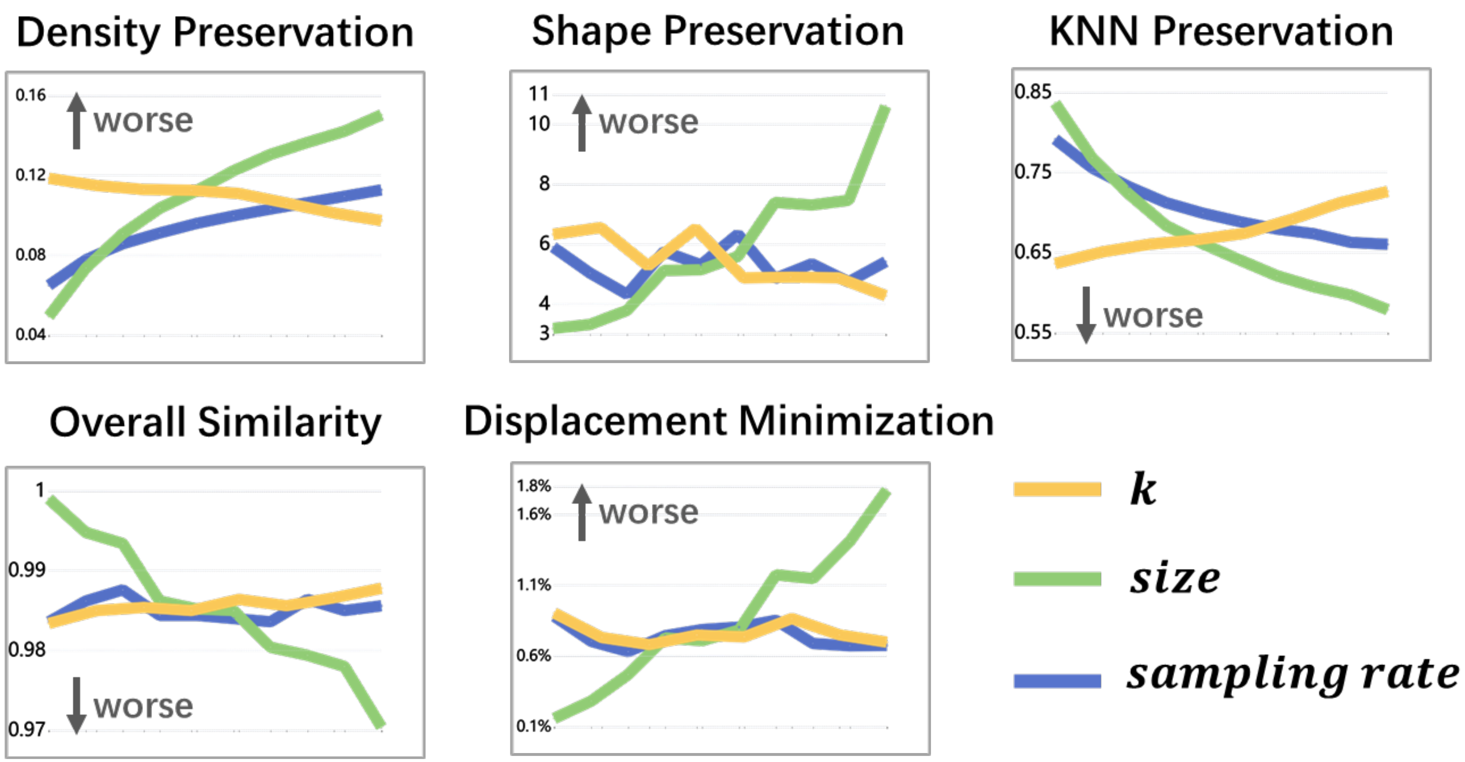
Time Complexity: $O(N'\sqrt{N'})$

N' is the number of nodes to be packed, including **data nodes** and **dummy nodes**.

Evaluation - quantitative evaluation

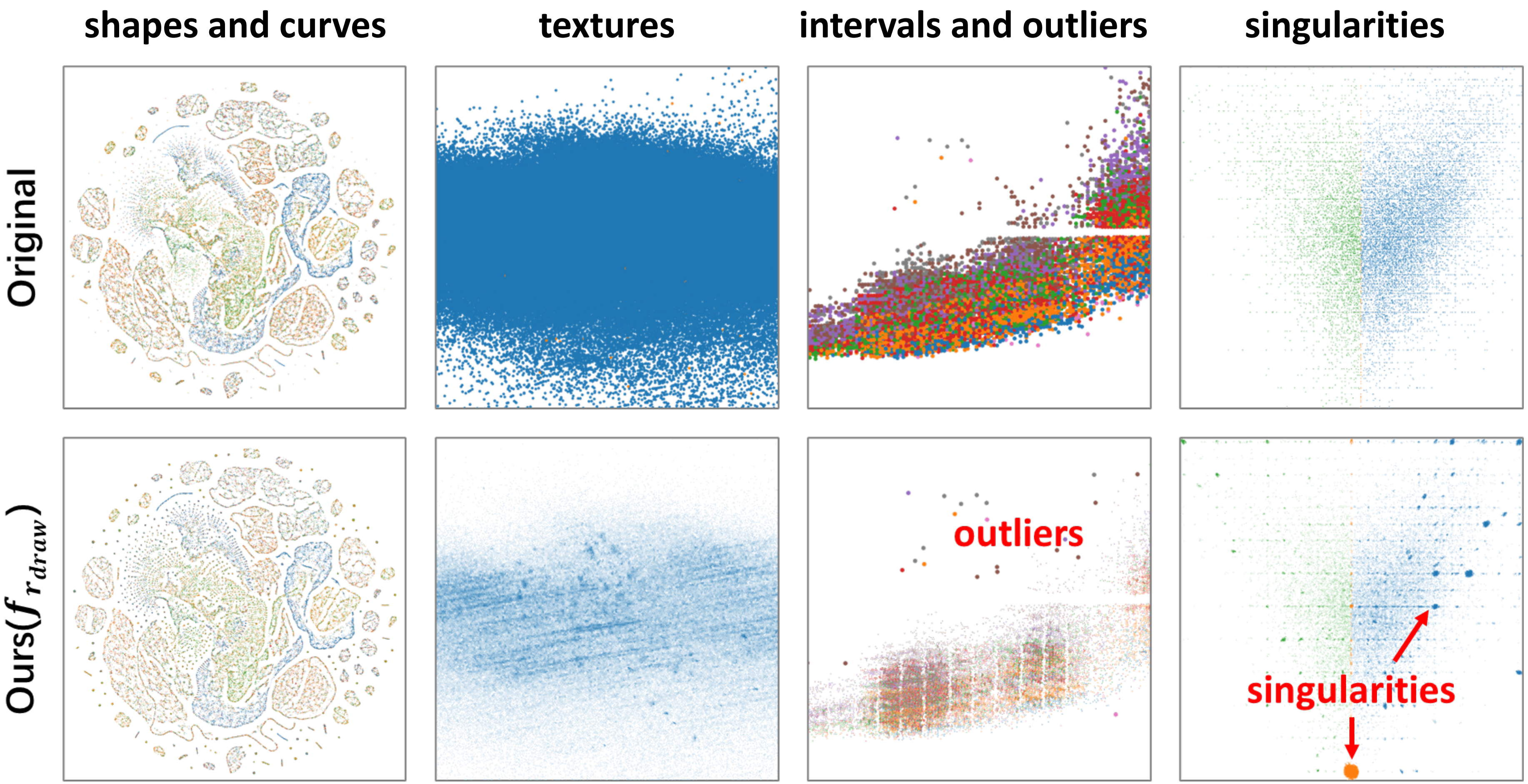
Impact of parameters on metrics:

- ***Size*** has a larger impact than ***k*** and ***sampling rate***, and all metrics get worse as it raises.
- ***Size*** controls the global resolution of the captured structures.
- Our method is fairly robust on parameters.



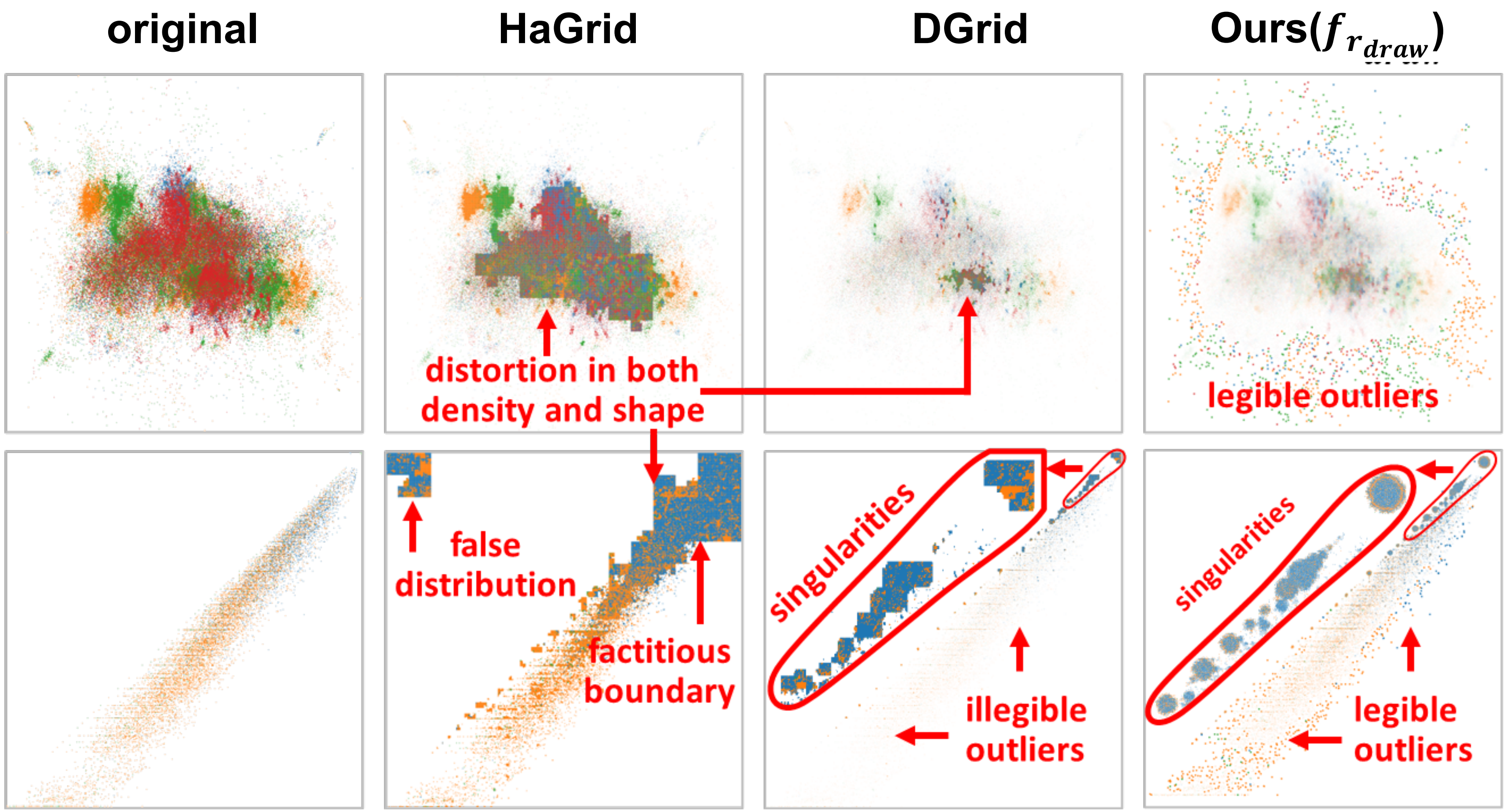
Evaluation - qualitative evaluation

Our method can maintain data distribution and reveal details hidden by overdraw.



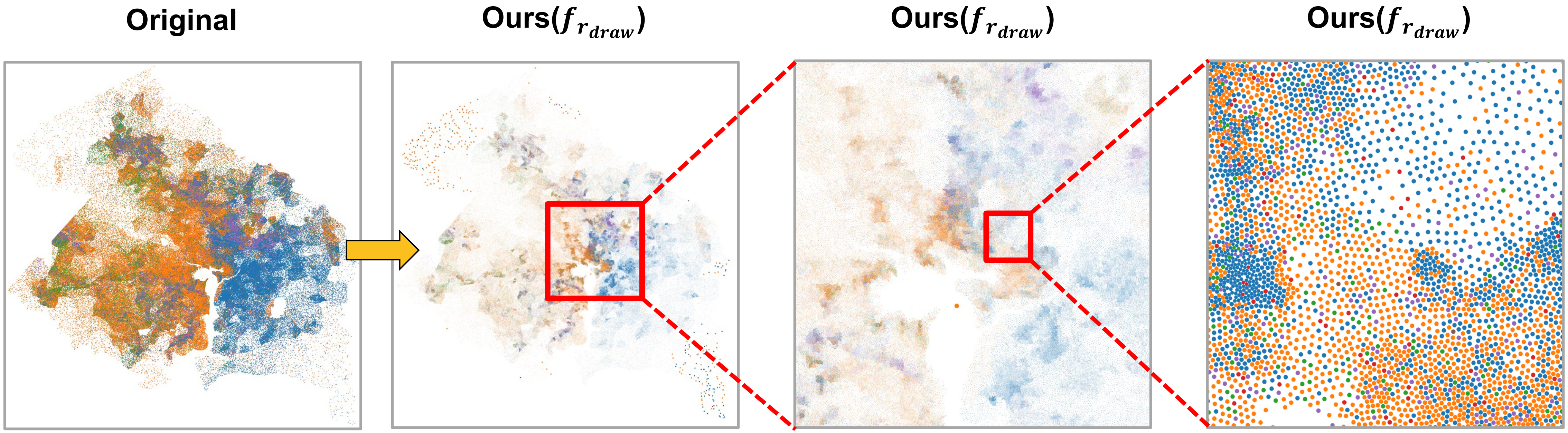
Evaluation - qualitative evaluation

Our method can overcome the crowded issue faced by state-of-the-art methods.



Evaluation - qualitative evaluation

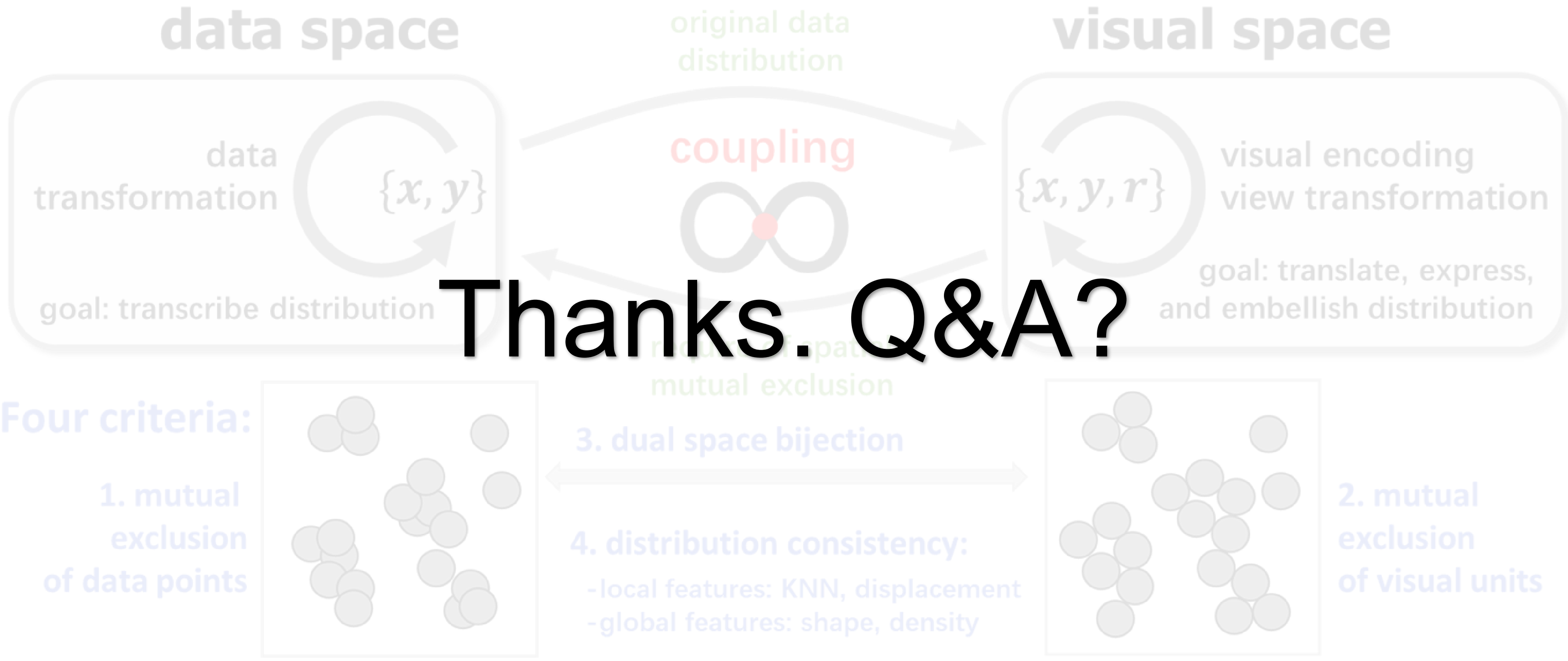
Our method can present rich and complete details at the micro scale.



Conclusion

- We contribute a dual space coupling model to represent the complex relationship within and between data space and visual space analytically to solve the scatterplot overdraw problem.
- The proposed model introduces a new design space for promising overlap removal algorithm and interaction paradigm.
- We also develop an overlap-free scatterplot visualization method on the basis of the model, which shows competitive advantages compared with the state-of-the-art methods.

Thanks. Q&A?


data space
original data distribution
visual space
data transformation
$\{x, y\}$
coupling
$\{x, y, r\}$
visual encoding view transformation
goal: transcribe distribution
goal: translate, express, and embellish distribution
mutual exclusion
Four criteria:
1. mutual exclusion of data points
3. dual space bijection
4. distribution consistency:
-local features: KNN, displacement
-global features: shape, density
2. mutual exclusion of visual units