Learning Structured Representations by Embedding Class Hierarchy with *Fast Optimal Transport*

ICLR ‘25

Siqi Zeng, Sixian Du, Makoto Yamada, Han Zhao

ICLR
International Conference On Learning Representations

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

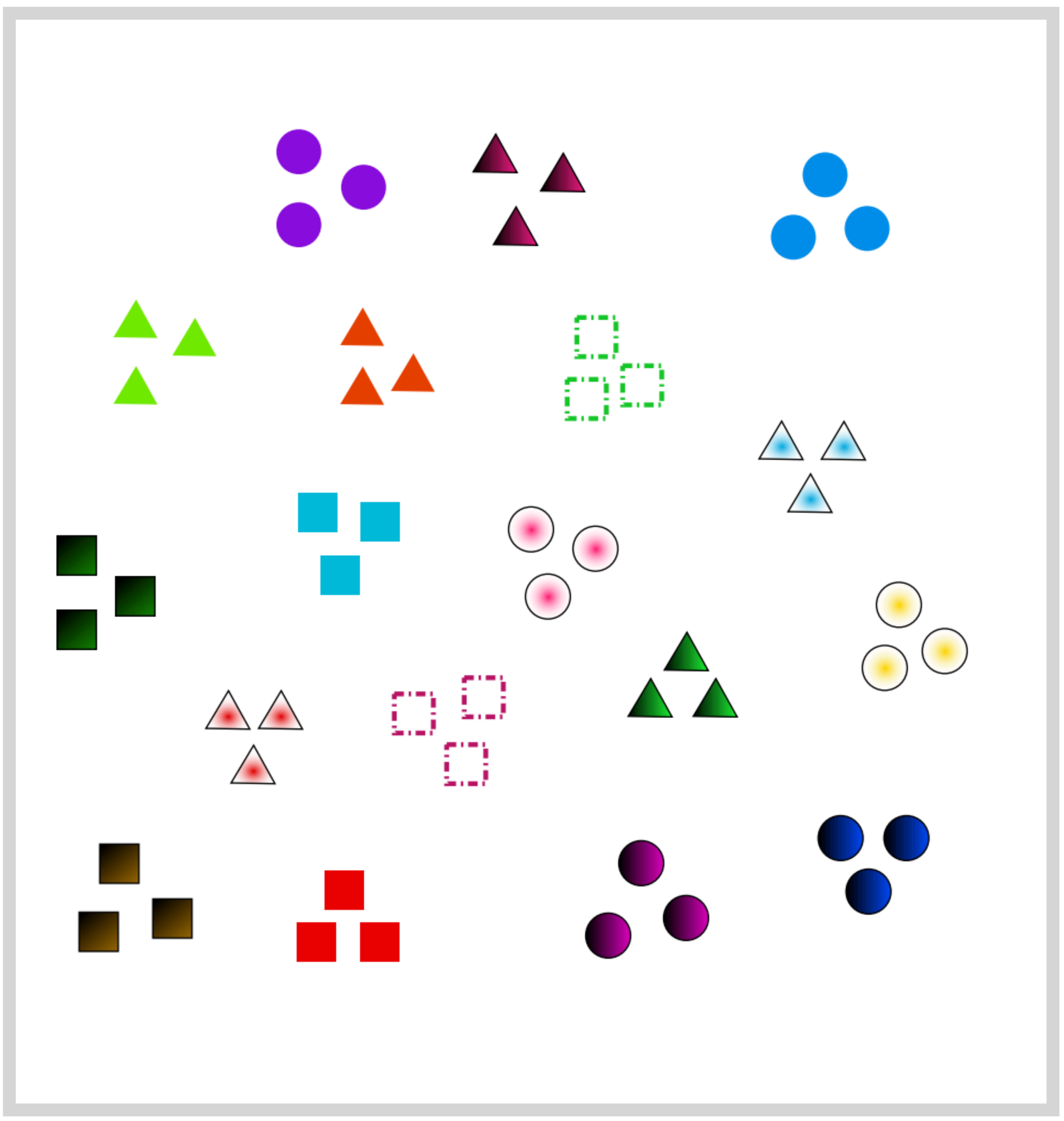
Stanford University

OIST

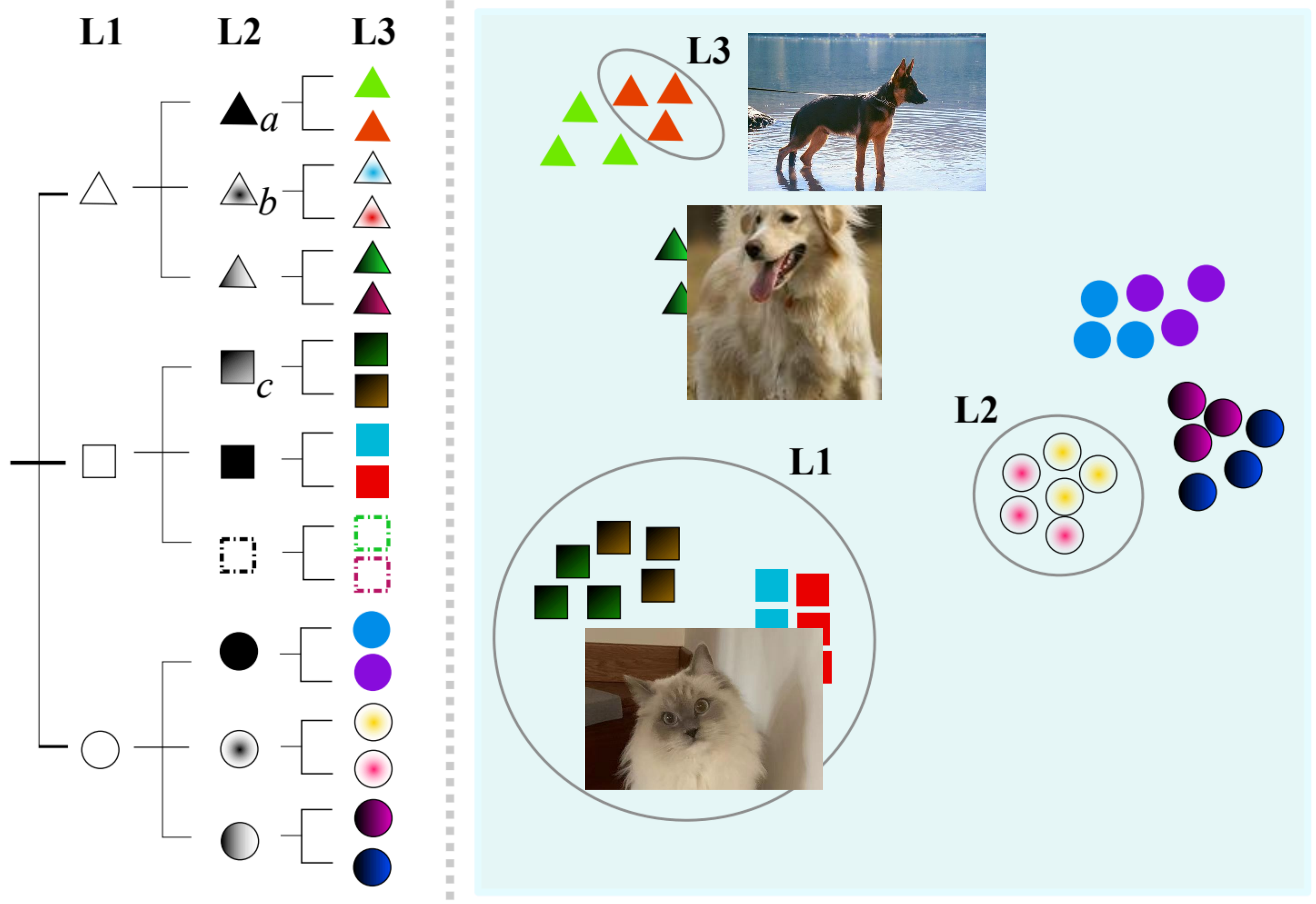
[Edited from slide of Zeng et al., 2022]

Background: Learning Structured Representations

Flat Representation



Structured Representation



[Edited from slide of Zeng et al., 2022]

Structured Representations w/ Euclidean CoPhenetic Correlation Coefficient

$$\text{CPCC}(d_{\mathcal{T}}, \rho) := \frac{\sum_{i<j}(d_{\mathcal{T}}(v_i, v_j) - \overline{d_{\mathcal{T}}})(\rho(v_i, v_j) - \overline{\rho})}{\sqrt{\sum_{i<j}(d_{\mathcal{T}}(v_i, v_j) - \overline{d_{\mathcal{T}}})^2}\sqrt{\sum_{i<j}(\rho(v_i, v_j) - \overline{\rho})^2}}$$

$\rightarrow \rho(v_i, v_j) :=$ ***The Euclidean distance between two class centroids, where v_i and v_j are fine classes.***

$\rightarrow d_{\mathcal{T}}(v_i, v_j) :=$ The shortest path between two vertices on the tree.

Final Objective: $\mathcal{L}(\mathcal{D}) = \sum_{(x,y)\in\mathcal{D}} \ell_{\text{Flat}}(y, \hat{y}) - \lambda \cdot \text{CPCC}(d_{\mathcal{T}}, \rho)$

Research Questions In Our Work

- **RQ1**: *What's the limitation of ℓ_2-CPCC and how to address it?*
  - **A**: EMD-CPCC!
- **RQ2**: *EMD is slow. Can we make it faster for CPCC?*
  - **A**: Yes, and we propose the FastFT method as an approximation of EMD for our learning setting.

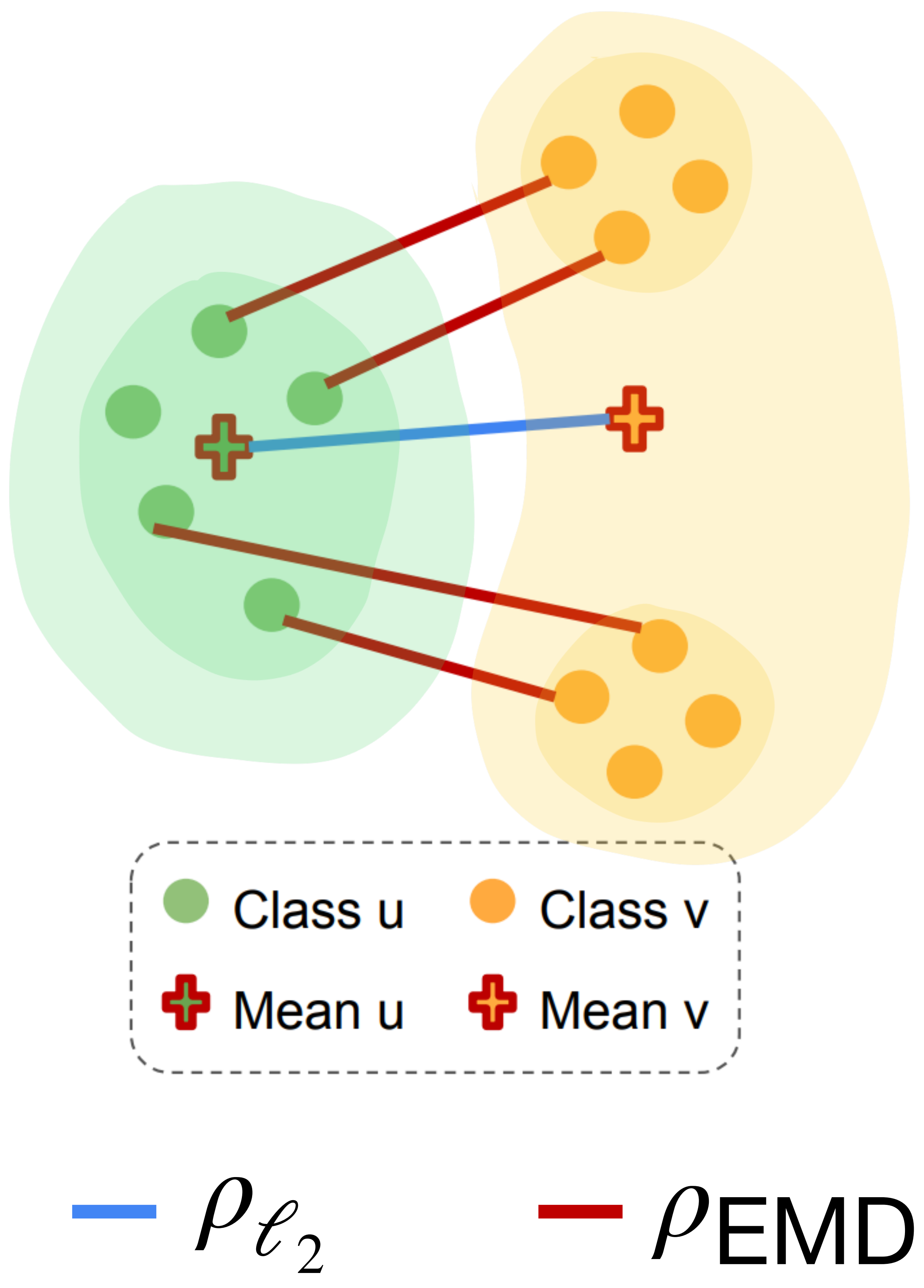
RQ1: Motivation of Using EMD-CPCC

$\rightarrow \rho_{\ell_2}(u, v) :=$ ***The Euclidean distance between two class centroids, where u and v are fine classes.***

Our method: replacing ρ_{ℓ_2} with ρ_{EMD},

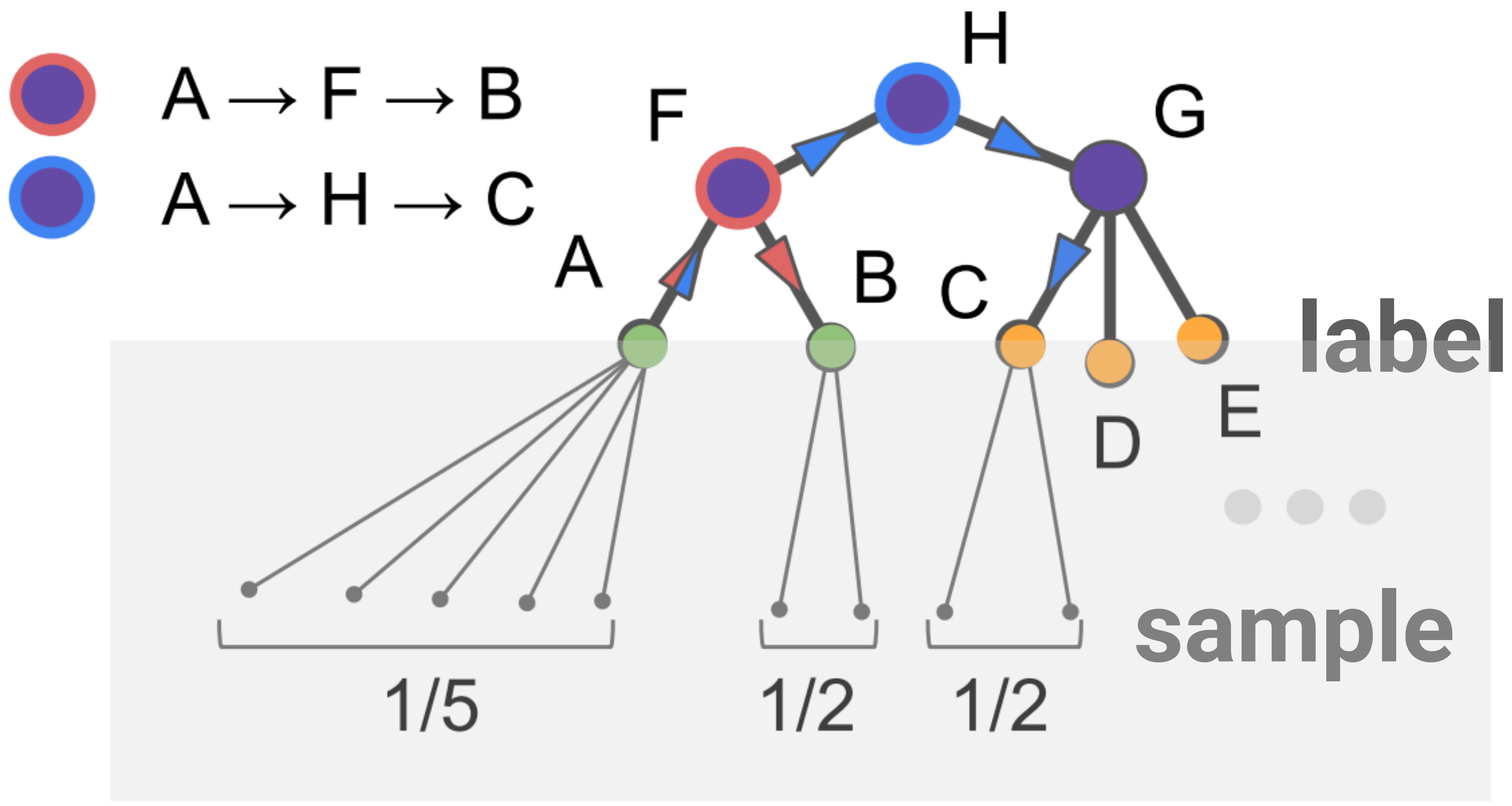
- We learn **more fine-grained structured representations** affected by distribution geometry
- EMD depends on **pairwise relationship for each source-target value pair** in the support
- We can also use other EMD approximation methods for ρ, all of them belongs to **O**ptimal**T**ransport**-CPCC** family.

Mis-representation of "multi-mode" distributions



RQ2: Fast Optimal Transport = FlowTree w/ Constructed Label Tree (FastFT)

- Tree-based approximation (FlowTree, TWD…) are linear time if tree exists, but building a tree is slow (ex., QuadTree, learning tree weights, …).
- **We have our label tree → Fast FlowTree = we run *FlowTree* on *Augmented Label Tree*!**
- ***Augmented Label Tree*:**
  - We extend label tree to sample tree by 1 level downwards
  - Every CPCC call only uses a subtree rooted by an internal node

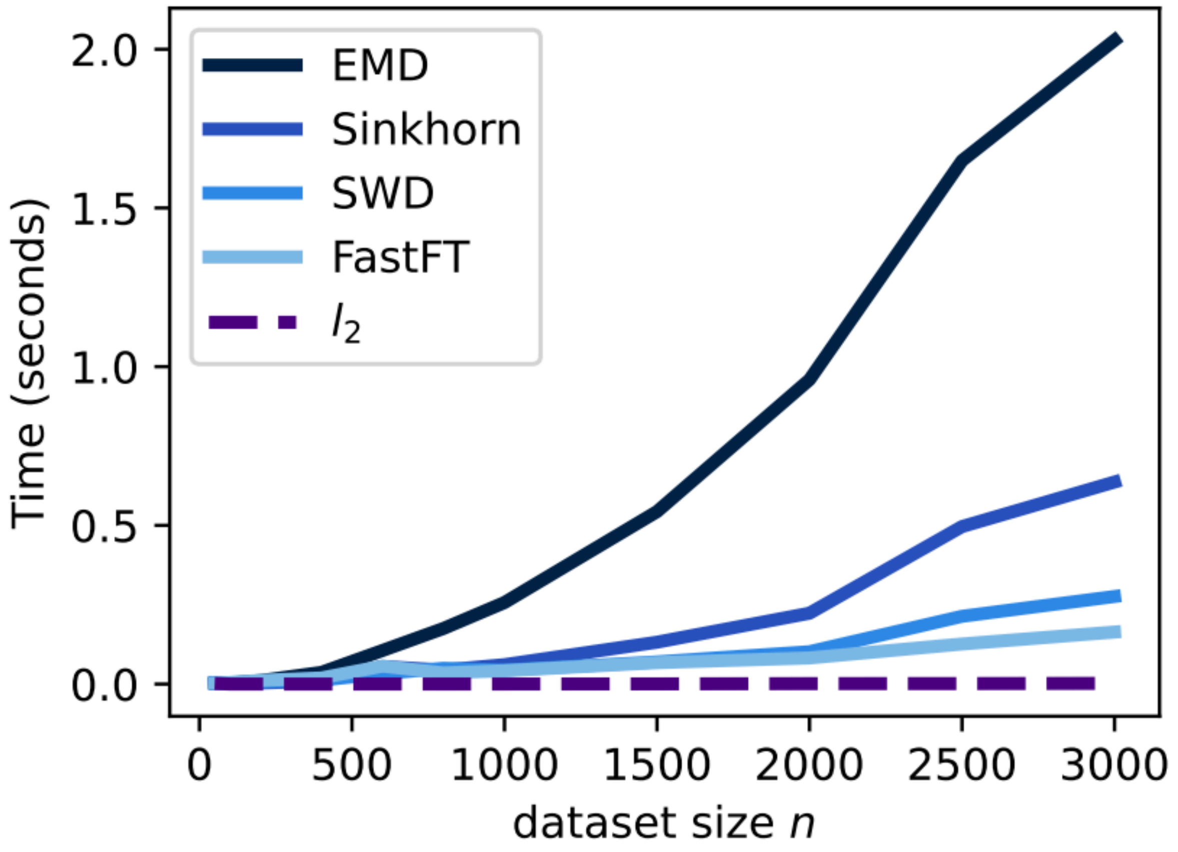


Theorem 3.2: With *Augmented Label Tree*, running *FlowTree* reduces to using Greedy Matching in *1d-EMD* w/o sorting, so the time complexity is linear $O(nd)$.

Results 1: Runtime Comparison

- Fast FT is ***linear-time***, and has better ***advantage for larger batch size***

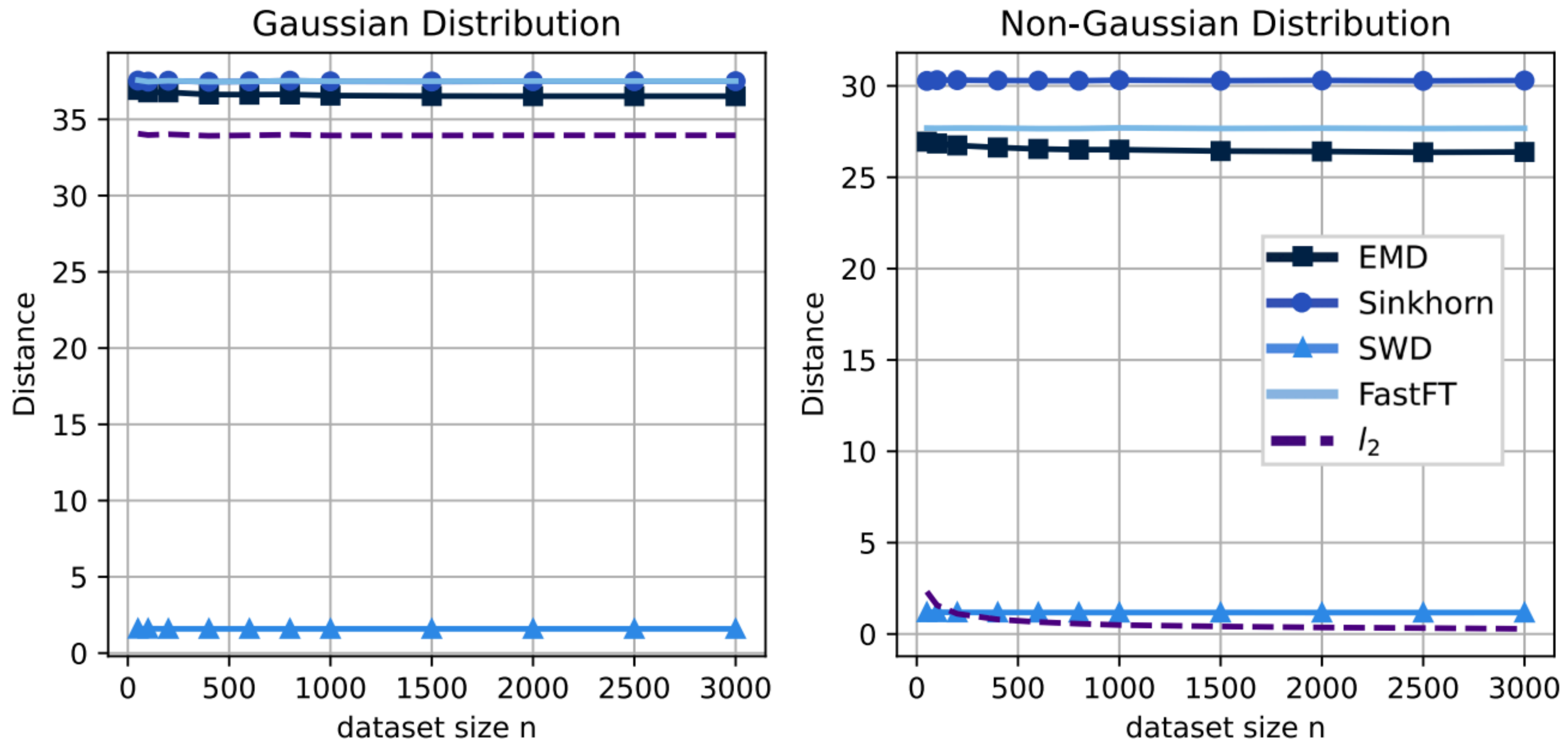
Method	Time Complexity
ℓ_2	$O((nk+k^2)d)$
EMD	$\tilde{O}(n(d+n^2\log n))$
Sinkhorn	$\tilde{O}(n^2(d+I))$
TWD	$\tilde{O}(n^2dI)$
SWD	$\tilde{O}(np(d+\log n))$
FT	$\tilde{O}(nd\log(d\Phi))$
FastFT	$\tilde{\Theta}(nd)$



Dataset	Objective	Per Epoch (s)	Dataset	Objective	Per Epoch (s)
	Flat	85.88 (3.69)		Flat	147.67 (4.54)
	ℓ_2	83.78 (9.80)		ℓ_2	149.39 (1.24)
CIFAR10	$\boxed{\text{FastFT}}$	$\boxed{85.21\ (4.26)}$	*INAT*	$\boxed{\text{FastFT}}$	$\boxed{132.01\ (3.24)}$
(b=128)	EMD	87.70 (2.43)	(b=1024)	EMD	234.68 (4.40)
	Sinkhorn	85.65 (4.15)		Sinkhorn	165.69 (4.60)
	SWD	92.90 (4.65)		SWD	152.52 (8.90)

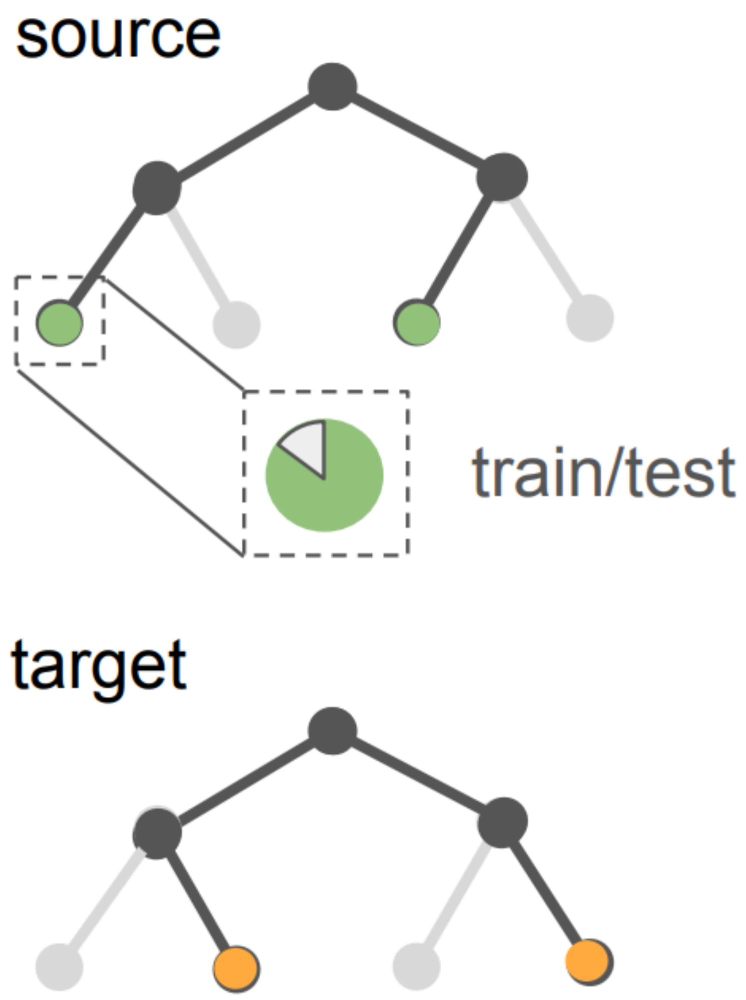
Results 2: Approximation Error of EMD

- **FastFT is a good approximation of EMD**
- EMD reduces to ℓ_2 for two Gaussians
- SWD is closer to ℓ_2 for multi-mode scenario



Results 3: Hierarchical Classification & Retrieval

Setup



Coarse

Dataset	Objective	sAcc	tAcc	sMAP	tMAP	Dataset	sAcc	tAcc	sMAP	tMAP
CIFAR10	Flat	99.58	87.30	99.22	89.66	*INAT*	94.63	38.81	70.41	34.00
	FastFT	99.61	**87.79**	**99.91**	93.04		**94.66**	39.43	72.90	35.63
	EMD	**99.61**	87.45	99.88	**93.07**		94.64	**41.01**	73.87	35.58
	Sinkhorn	99.61	87.41	99.87	92.60		94.30	36.94	68.75	34.92
	SWD	99.56	87.63	99.36	90.12		94.52	39.38	**75.13**	**38.11**

Fine

Dataset	Objective	sAcc	tAcc	sMAP	Dataset	sAcc	tAcc	sMAP
CIFAR10	ℓ_2	96.96	55.71	99.22	*INAT*	88.62	26.78	56.10
	FastFT	96.90	55.99	99.24		88.49	**27.10**	56.21
	EMD	**97.05**	56.12	99.24		**88.68**	26.78	**56.83**
	Sinkhorn	96.95	54.89	99.27		88.08	26.77	51.56
	SWD	96.96	**59.21**	**99.45**		88.46	26.78	54.19

- **OT-CPCC > Flat on coarse** level, **OT-CPCC > ℓ_2-CPCC on fine** level.
- For generalization performance, there's **no single OT method always achieves the best.**

Conclusion and Future Work

- **Contribution**
  - Identify limitation of ℓ_2-CPCC for learning structured representations, propose OT-CPCC
  - Propose FastFT, a linear-time EMD approximation algorithm with low approximation error
  - Improved generalization, and preserve tree information better

Thanks for listening!

Paper **Code**

Contact: siqi6@illinois.edu