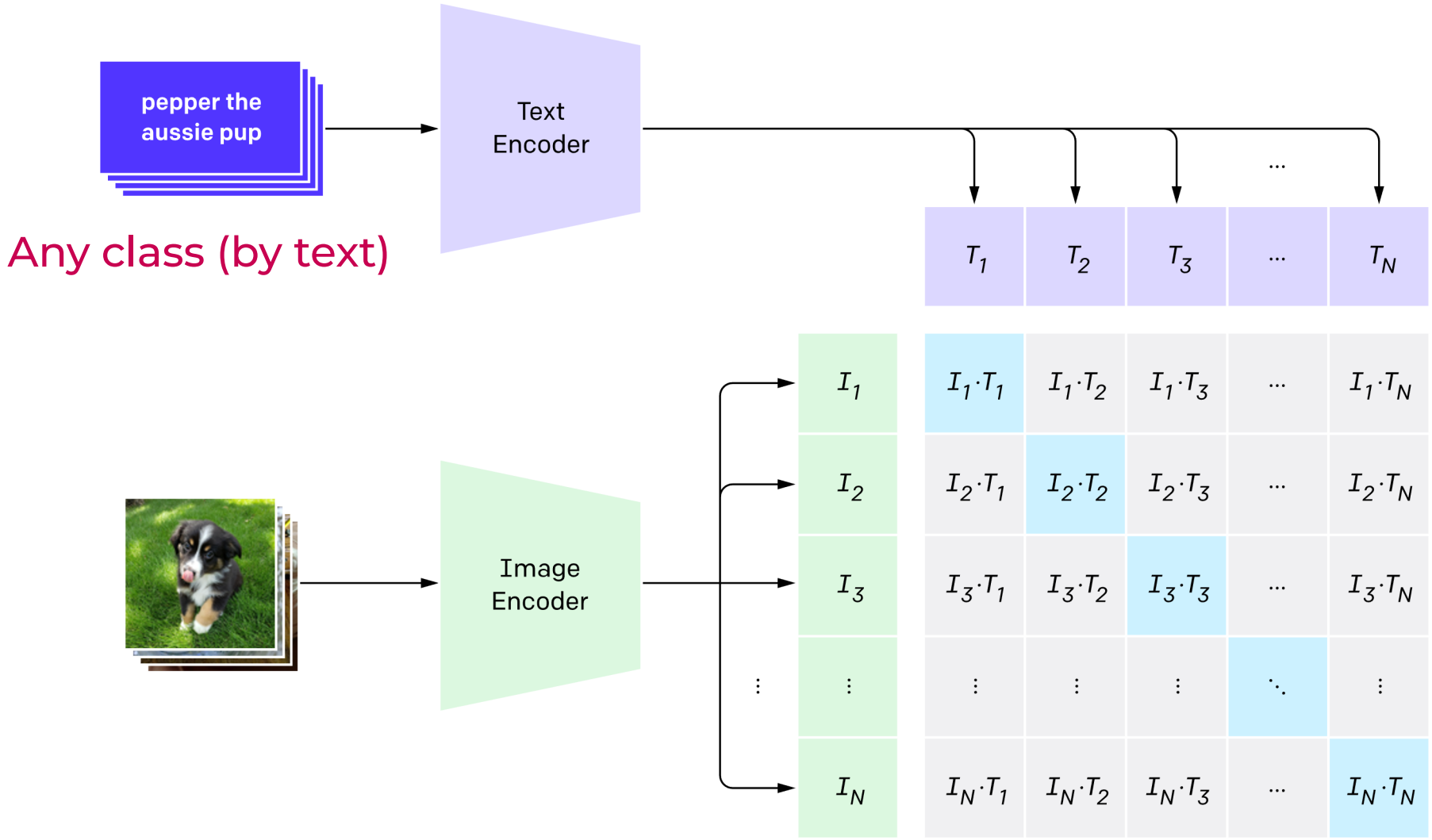
Aligning Vision and Language via Geometry and Topology

Sangwoo Mo

Assistant Professor
POSTECH IME

Vision-Language Models (e.g., CLIP) are Amazing!



Vision-Language Models (e.g., CLIP) are Amazing!





...but Only for **Standard Domains** (e.g., LAION)

Backend url:
https://knn5.laior
Index:
laion_5B

Clip retrieval works by converting the text query to a CLIP embedding , then using that embedding to query a knn index of clip image embedddings

Display captions ☑
Display full captions ☐
Display similarities ☐
Safe mode ☑
Hide duplicate urls ☑
Hide (near) duplicate images ☑
Search over
image
Search with multilingual clip ☐

french cat


french cat

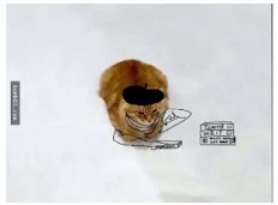

french cat


How to tell if your feline is french. He wears a b...


イケメン猫モデル「トキ・ナンタケット」がかっこいい - NAVER まとめ


Hilarious pics of funny cats! funnycatsgif.com


Hipster cat


網友挑戰「加幾筆畫出最創意貓咪圖片」，笑到岔氣之後我也手


cat in a suit Georgian sells tomatoes




French Bread Cat Loaf Metal Print

CLIP for **Specialist Domains?**



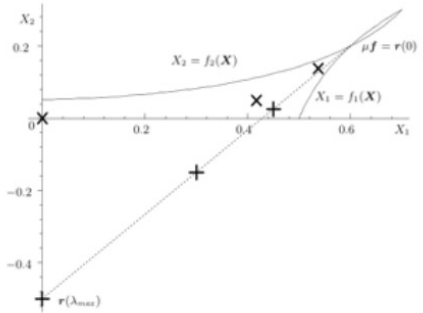
“many houses arranged neatly and some roads crossed them with some plants in the roadside”

(a) Remote sensing



“grey classic u-neck t-shirt. short sleeve t-shirt in heather grey. tonal logo flag at u-neck collar.”

(b) Fashion



“Fig. 2. variable 2-D initial probe distribution used for CFO runs reported here”

(c) Scientific figure



“simpsons wearing headphones and looking like a guy”

(d) Comics

CLIP for **Specialist Domains?**



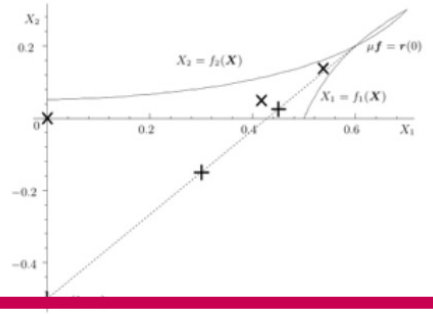
“many houses arranged neatly and some roads crossed them with some plants in the roadside”

(a) Remote sensing



“grey classic u-neck t-shirt. short sleeve t-shirt in heather grey. tonal logo flag at u-neck collar.”

(b) Fashion



“Fig. 2. variable 2-D initial probe distribution used for CFO runs reported here”

(c) Scientific figure

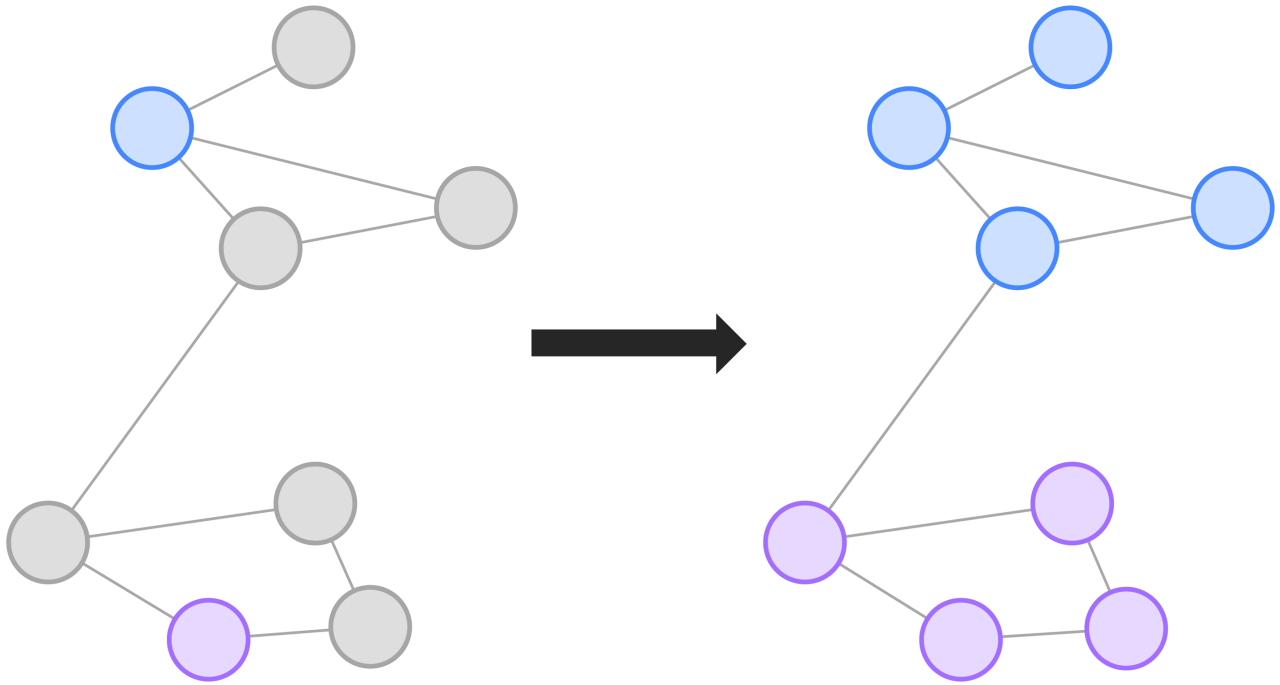


“simpsons wearing headphones and looking like a guy”

(d) Comics

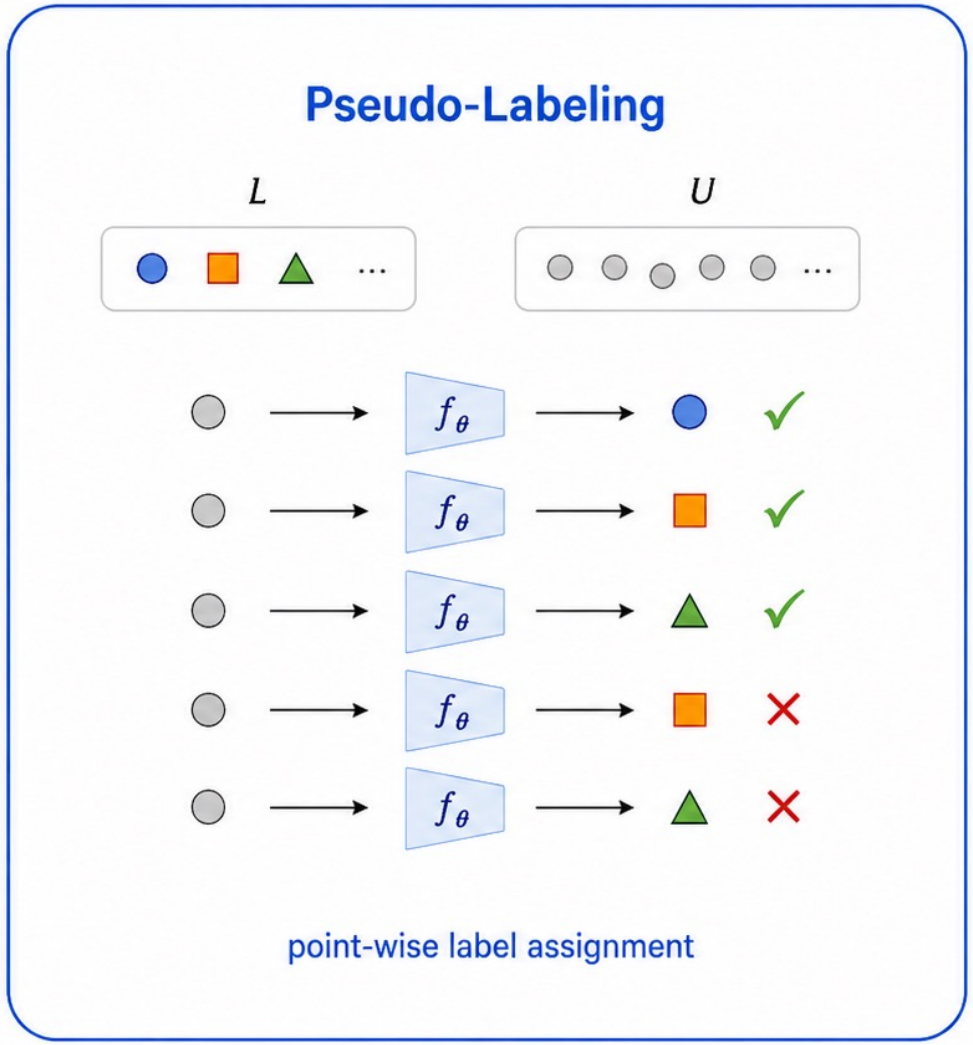
Challenge: We don’t have many **detailed captions**

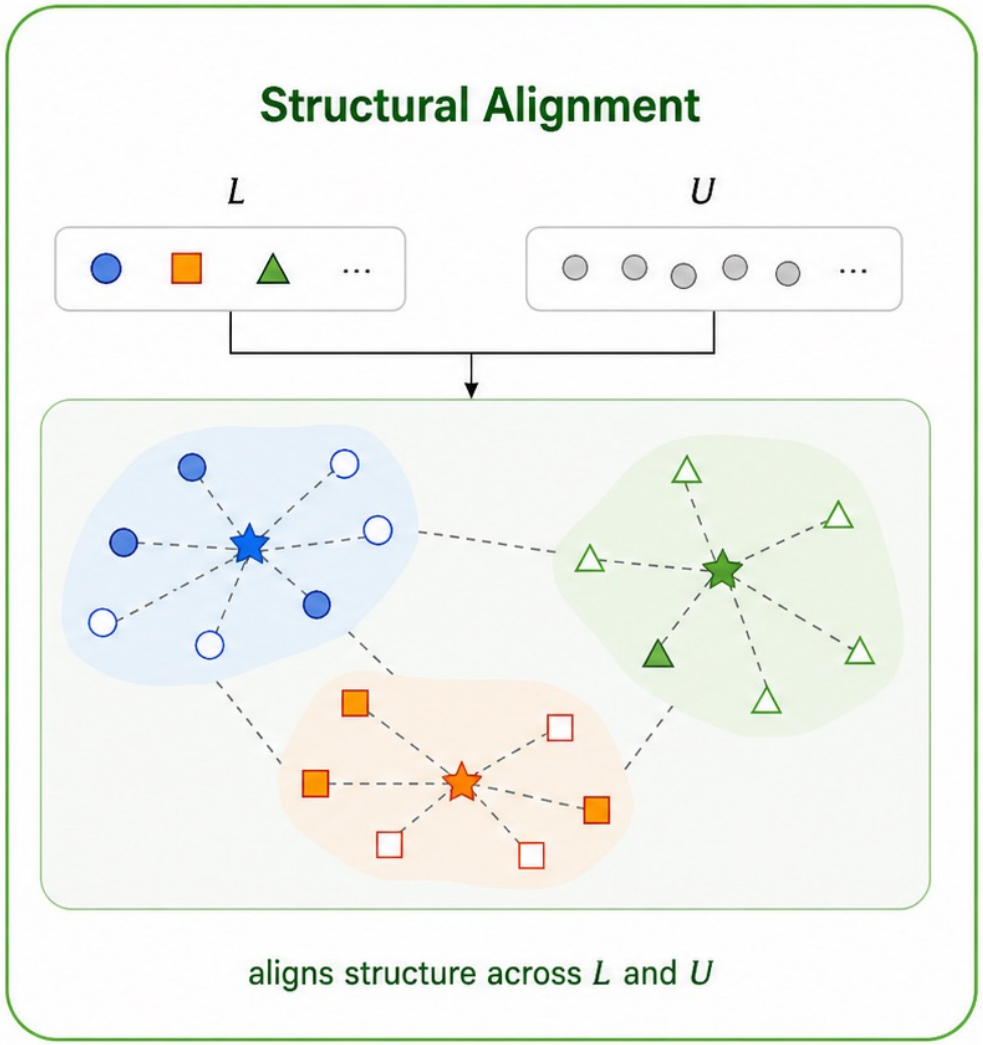
Solution: **Semi-Supervised Learning** (Semi-SL)



L = image-text pairs / **U** = image-only

Two Perspectives on Utilizing Unlabeled Data





S-CLIP: Semi-supervised Vision-Language Learning using Few Specialist Captions



Sangwoo Mo



Minkyu Kim



Kyungmin Lee



Jinwoo Shin

NeurIPS 2023

Motivation: Naïve **Pseudo-Labeling (PL) Fails**



Nearest image
in the batch

Their captions are different!
(naive pseudo-labeling fails)



"two tennis courts and some buildings are in two sides of a road"

"the huge rectangle contains a baseball field, a swimming pool, two basketball courts and eight tennis courts"

Idea 1: Caption-level PL

Pseudo-label = **Combination of the captions**
(from visually similar images in the batch)



=

0.4 ×



"…tennis courts…"

+

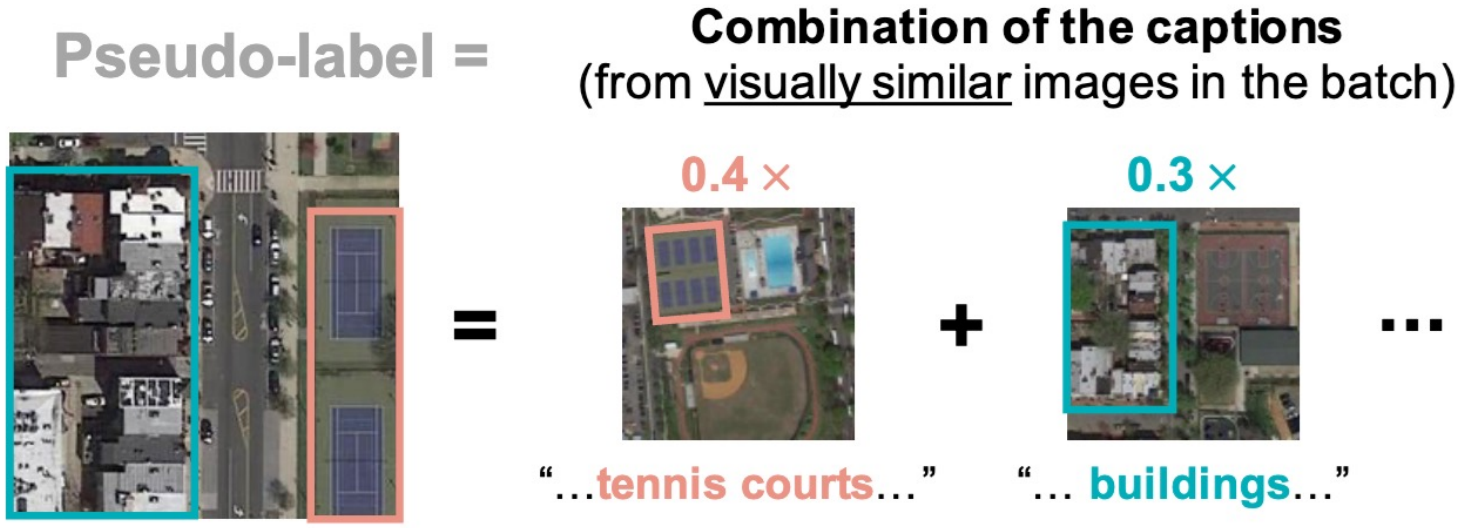
0.3 ×



"… buildings…"

…

Idea 1: **Caption-level PL**



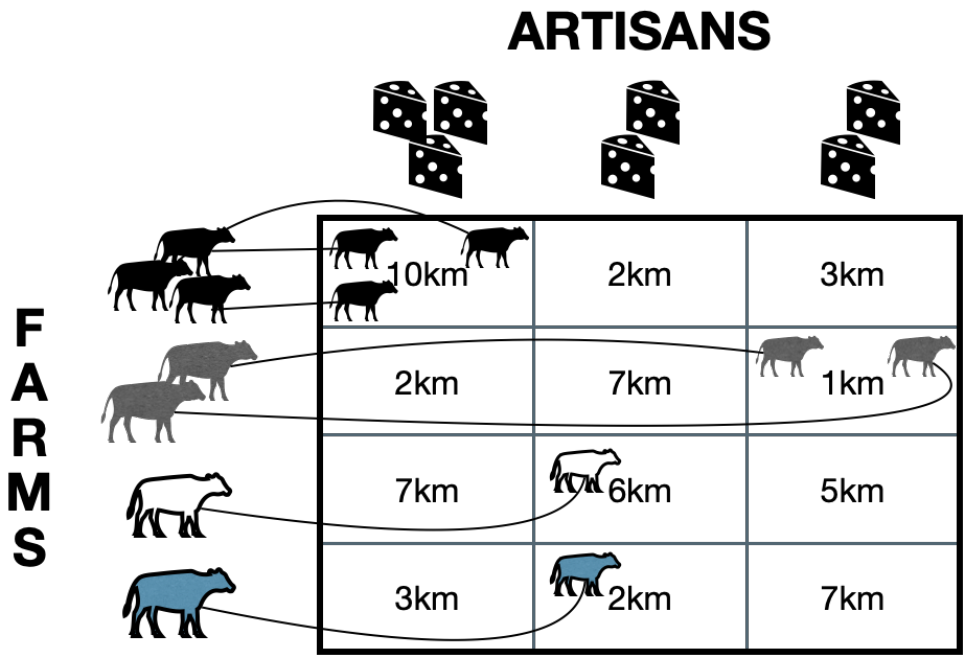
Optimal Transport (OT) between labeled & unlabeled images

$$\min_{\Gamma \in \Pi(\mathbf{p},\mathbf{q})} \langle \Gamma, \mathrm{C} \rangle - \lambda H(\Gamma), \quad \Pi(\mathbf{p},\mathbf{q}) = \{\Gamma \in \mathbb{R}_{+}^{M \times N} \mid \Gamma\, \mathbf{1}_N = \mathbf{p},\ \Gamma^{\top} \mathbf{1}_M = \mathbf{q}\}.$$

Idea 1: **Caption-level PL**

Optimal Transport (OT) between labeled & unlabeled images

$$\min_{\Gamma \in \Pi(\mathbf{p},\mathbf{q})} \langle \Gamma, \mathrm{C} \rangle - \lambda H(\Gamma), \quad \Pi(\mathbf{p},\mathbf{q}) = \{\Gamma \in \mathbb{R}_{+}^{M \times N} \mid \Gamma \mathbf{1}_N = \mathbf{p},\ \Gamma^{\top} \mathbf{1}_M = \mathbf{q}\}.$$



Algorithm 1 Sinkhorn-Knopp Algorithm (SK).

Require: $\mathbf{a}, \mathbf{b}, \mathbf{C}, \lambda$

$\mathbf{u}^{(0)} = \mathbf{1}, \mathbf{K} = \exp(-\mathbf{C}/\lambda)$

for i in $1, \ldots, n_{it}$ **do**

$\mathbf{v}^{(i)} = \mathbf{b} \oslash \mathbf{K}^{\top}\mathbf{u}^{(i-1)}$ // Update right scaling

$\mathbf{u}^{(i)} = \mathbf{a} \oslash \mathbf{K}\mathbf{v}^{(i)}$ // Update left scaling

end for

return $\mathbf{T} = \mathrm{diag}(\mathbf{u}^{(n_{it})})\mathbf{K}\,\mathrm{diag}(\mathbf{v}^{(n_{it})})$

Idea 2: Keyword-level PL

Pseudo-label = ∈

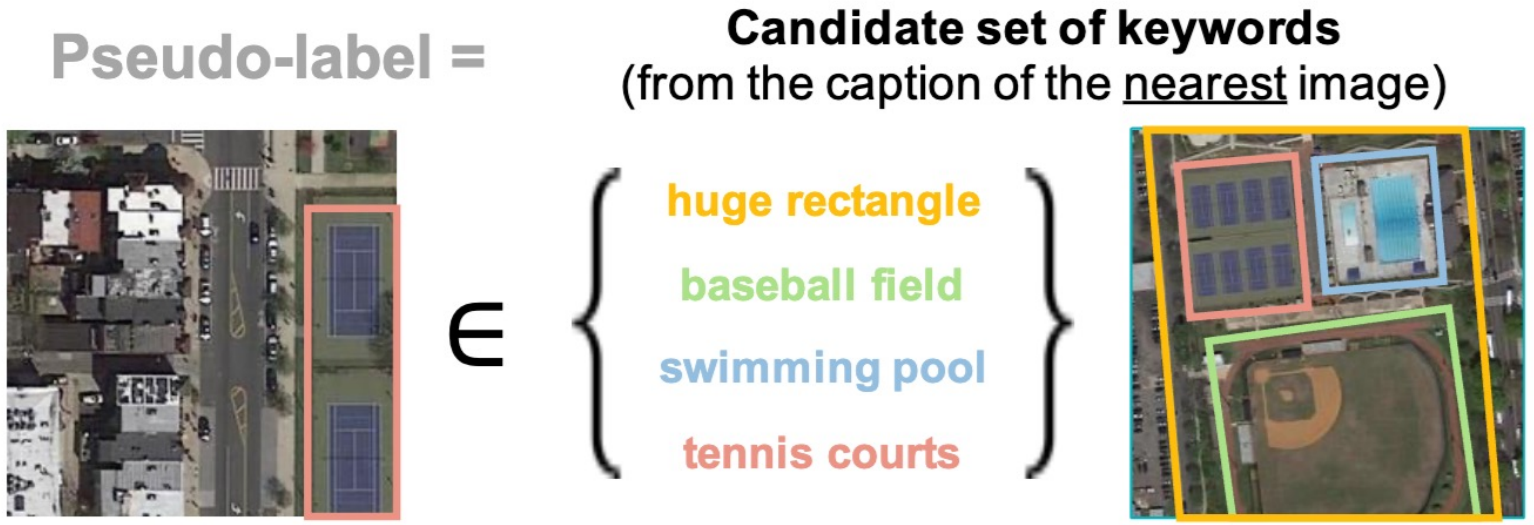
Candidate set of keywords
(from the caption of the nearest image)



{ huge rectangle, baseball field, swimming pool, tennis courts }



Idea 2: **Keyword-level PL**



Partial Label Learning (PLL) problem

→ Solved by **Expectation Maximization (EM)** algorithm

Idea 2: **Keyword-level PL**

Expectation Maximization (EM)

Jointly infer the label $\mathbf{z}$ (from candidates) and update the model θ

log of expectation of P(x|z)

$$\text{Goal: } \hat{\theta} = \arg\max_{\theta} \log\left(\sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z} \mid \theta)\right) \qquad f(\mathrm{E}[X]) \geq \mathrm{E}[f(X)]$$

1. E-step: compute

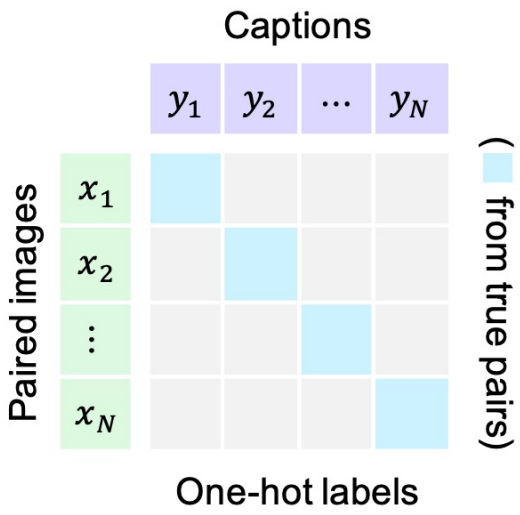
expectation of log of P(x|z)

$$\mathrm{E}_{z|x,\theta^{(t)}}\left[\log\left(p(\mathbf{x}, \mathbf{z} \mid \theta)\right)\right] = \sum_{\mathbf{z}} \log\left(p(\mathbf{x}, \mathbf{z} \mid \theta)\right) p\left(\mathbf{z} \mid \mathbf{x}, \theta^{(t)}\right)$$

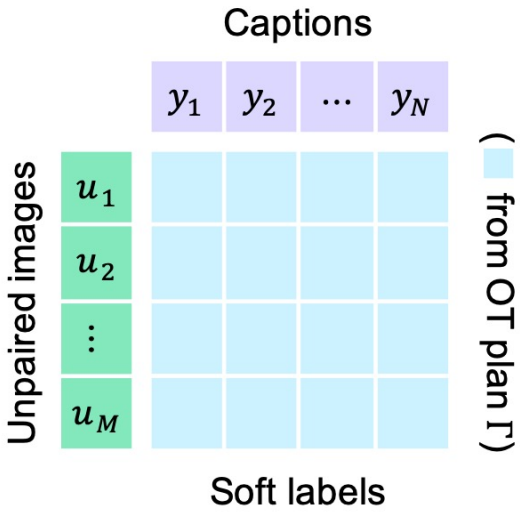
2. M-step: solve

$$\theta^{(t+1)} = \arg\max_{\theta} \sum_{\mathbf{z}} \log\left(p(\mathbf{x}, \mathbf{z} \mid \theta)\right) p\left(\mathbf{z} \mid \mathbf{x}, \theta^{(t)}\right)$$

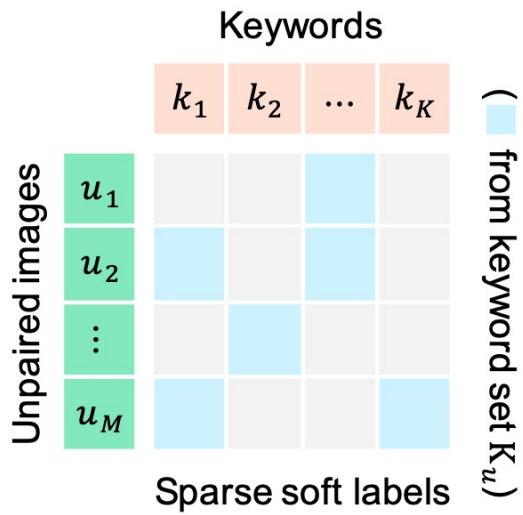
Training Objective of S-CLIP



(a) CLIP loss $\mathcal{L}_{\text{CLIP}}$



(b) Caption loss $\mathcal{L}_{\text{caption}}$



(c) Keyword loss $\mathcal{L}_{\text{keyword}}$

Performance ↑

Zero-shot classification

Method	Data	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
CLIP (original)	-	45.3	50.5	65.5	58.9	47.8
CLIP (fine-tune)	L	58.3±0.3	63.5±3.4	76.5±3.2	61.9±1.2	63.1±1.3
Hard-PL [16]	L=U	56.6±3.5 (-1.7)	61.6±2.2 (-1.9)	78.1±2.5 (+1.6)	63.9±2.1 (+2.0)	63.2±2.6 (+0.1)
Soft-PL [21]	L=U	62.5±0.8 (+4.2)	65.7±2.7 (+2.2)	83.7±2.7 (+7.2)	65.7±0.6 (+3.8)	68.0±0.7 (+4.9)
S-CLIP (ours)	L=U	**66.9**±1.7 (+8.6)	**66.7**±1.6 (+3.2)	**86.9**±2.0 (+10.4)	**66.2**±1.1 (+4.3)	**73.0**±0.3 (+9.9)

Image-text retrieval

		Image→text R@5			Text→image R@5		
Method	Data	RSICD	UCM	Sydney	RSICD	UCM	Sydney
CLIP (original)	-	9.4	34.3	36.2	10.1	24.8	51.7
CLIP (fine-tune)	L	15.4±1.7	41.3±1.8	47.1±6.5	15.1±1.0	40.9±1.6	56.1±2.4
Hard-PL [16]	L=U	16.1±0.2	40.8±2.9	43.1±3.0	15.7±0.7	40.5±3.0	47.7±5.3
Soft-PL [21]	L=U	17.0±0.9	43.2±3.9	42.0±4.3	16.5±0.1	42.9±3.3	50.2±4.9
S-CLIP (ours)	L=U	**18.4**±0.6	**45.7**±1.4	**50.0**±3.0	**16.8**±1.2	**43.5**±1.5	**55.1**±2.0

Robustness ↑

Labeled & unlabeled images come from **different** datasets

Method	Data	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
CLIP (original)	-	45.3	50.5	65.5	58.9	47.8
CLIP (fine-tune)	L	58.3±0.3	63.5±3.4	76.5±3.2	61.9±1.2	63.1±1.3
Hard-PL [16]	L≠U	55.8±1.2 (-2.5)	61.0±6.2 (-2.5)	76.3±1.1 (-0.2)	62.5±2.7 (+0.6)	62.3±1.4 (-0.8)
Soft-PL [21]	L≠U	56.1±3.5 (-2.2)	62.4±1.8 (-1.1)	79.5±3.6 (+3.0)	63.3±2.6 (+1.4)	62.4±2.1 (-0.7)
S-CLIP (ours)	L≠U	**65.3**±2.1 (+7.0)	**66.0**±0.4 (+2.5)	**86.3**±1.1 (+9.8)	**68.3**±2.2 (+6.4)	**70.8**±2.5 (+7.7)

POSTECH

Topology-Aware Representation Alignment for Semi-Supervised Vision-Language Learning



Junwon You



Mihyun Jang



Sangwoo Mo†



Jae-Hun Jung†

† Equal advising

Under Review

Motivation: **Structural Alignment** Matters



Image caption :

"Eight storage tanks are surrounded by green meadows."

Predicted concepts :

tanks, storage, lawn, meadows

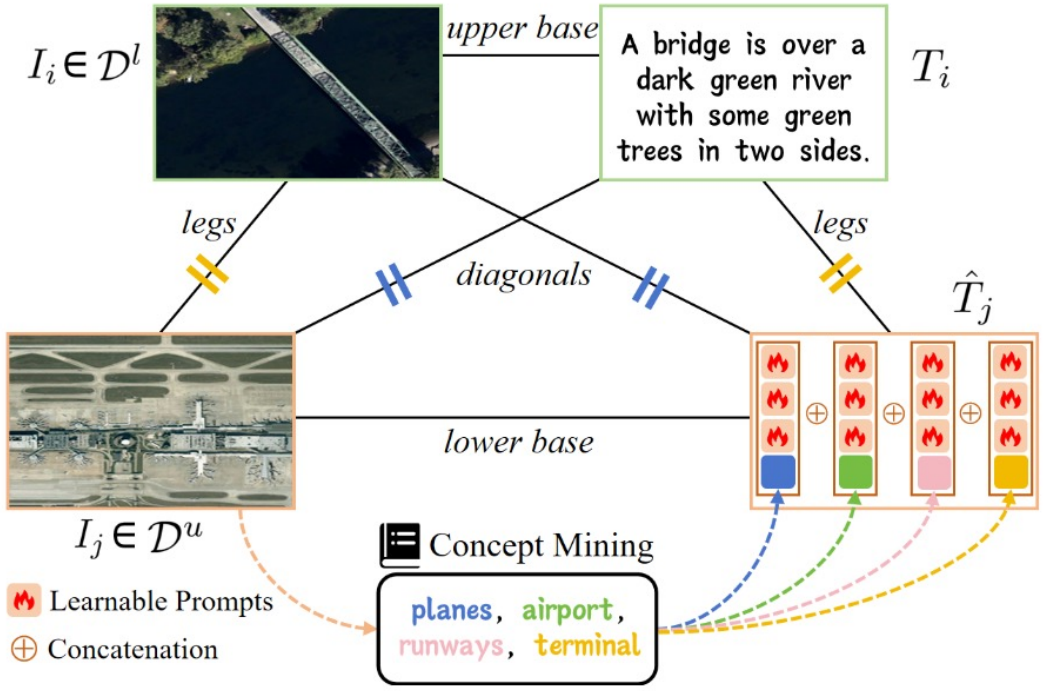


Image caption :

"Blue seriously stretchy dark wash mid-rise jegging."

Predicted concepts :

denim, jeggings, skinny, knee



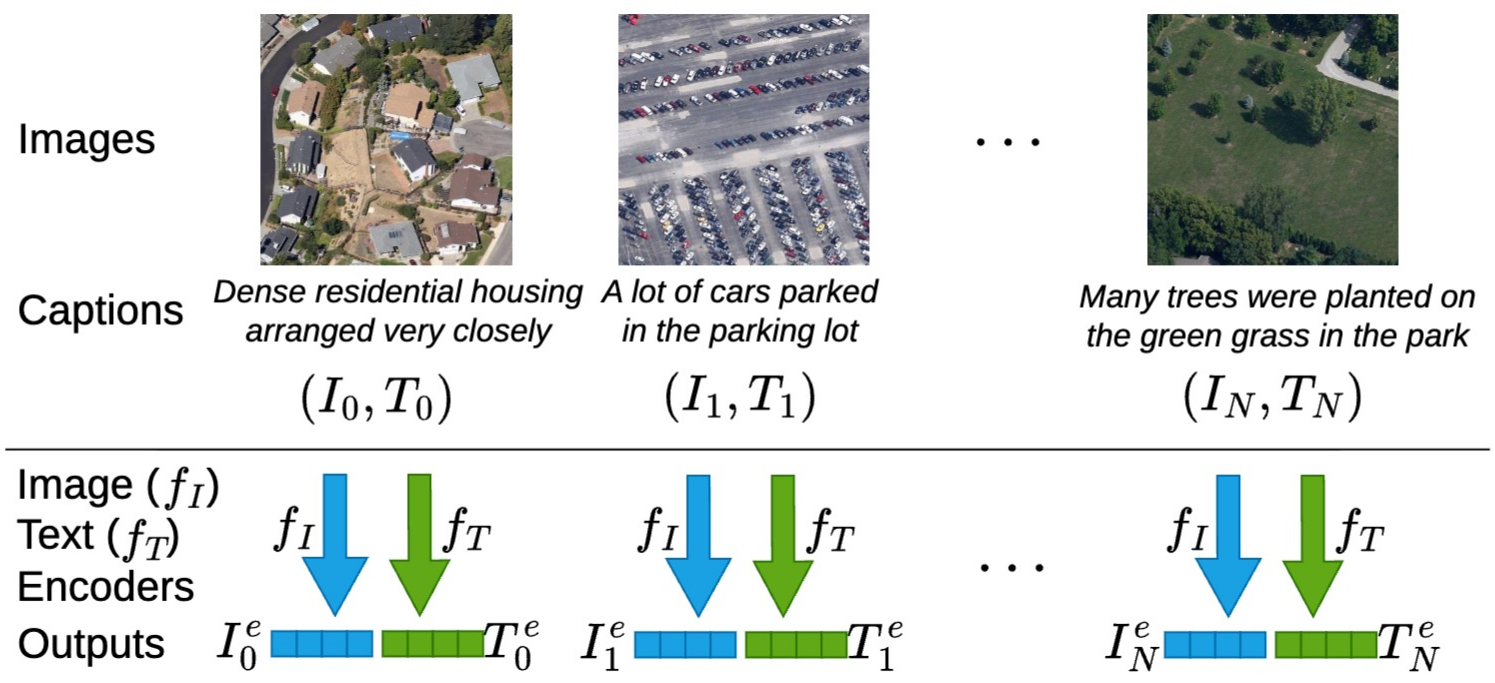
Gan et al. Semi-Supervised CLIP Adaptation by Enforcing Semantic and Trapezoidal Consistency. ICLR 2025.

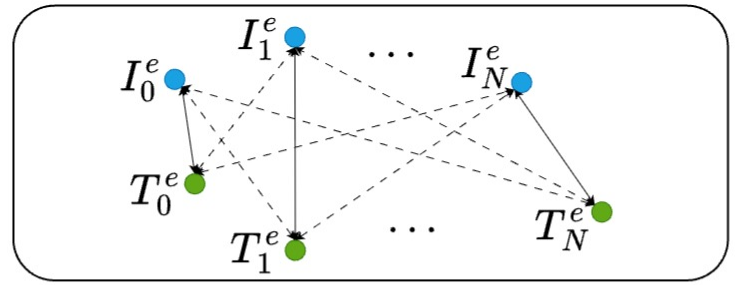
Motivation: **Structural Alignment** Matters

Method	Data	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
CLIP (original)	-	59.2	60.2	81.2	69.0	59.6
CLIP (fine-tuned)	L	75.7±2.2	80.1±6.8	92.2±1.5	70.7±5.5	79.8±3.8
Oracle (fully supervised fine-tuned)	L + U	87.5±1.8	67.9±2.3	94.2±1.5	75.7±2.9	89.3±1.2
Hard-PL (Lee et al., 2013)	L=U	78.0±2.0 (+2.3)	72.4±5.2 (-7.7)	92.3±2.6 (+0.1)	71.7±4.1 (+1.0)	81.4±2.2 (+1.6)
Soft-PL (Assran et al., 2021)	L=U	81.7±0.6 (+6.0)	78.9±5.2 (-1.2)	95.3±0.8 (+3.1)	72.0±2.5 (+1.3)	85.8±1.6 (+6.0)
S-CLIP (Mo et al., 2023)	L=U	81.4±1.8 (+5.7)	81.3±3.4 (+1.2)	95.9±1.8 (+3.7)	75.1±2.0 (+4.4)	86.4±1.7 (+6.6)
SEMICLIP (ours)	L=U	**83.8**±1.0 (+8.1)	**85.6**±3.4 (+5.5)	**96.6**±1.2 (+4.4)	**75.3**±1.9 (+4.6)	**87.4**±0.5 (+7.6)

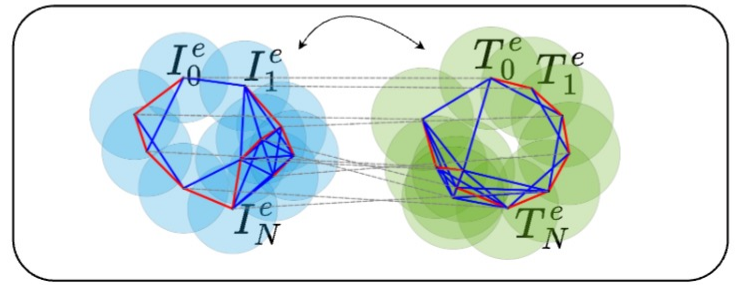
Gan et al. Semi-Supervised CLIP Adaptation by Enforcing Semantic and Trapezoidal Consistency. ICLR 2025.

Idea: **Topology-Aware Alignment** Can be Even Better



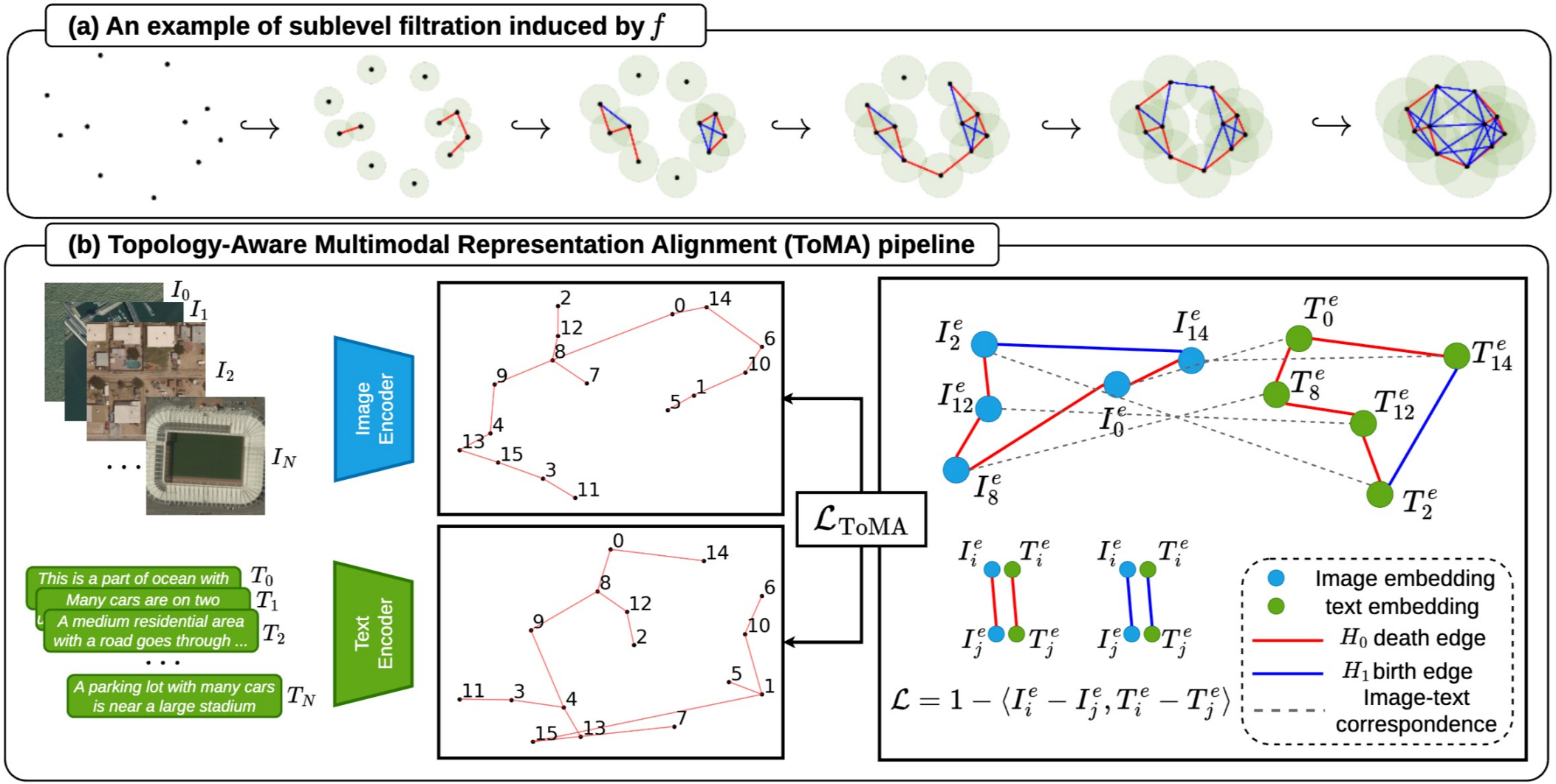


Pair-wise alignment (prior work)

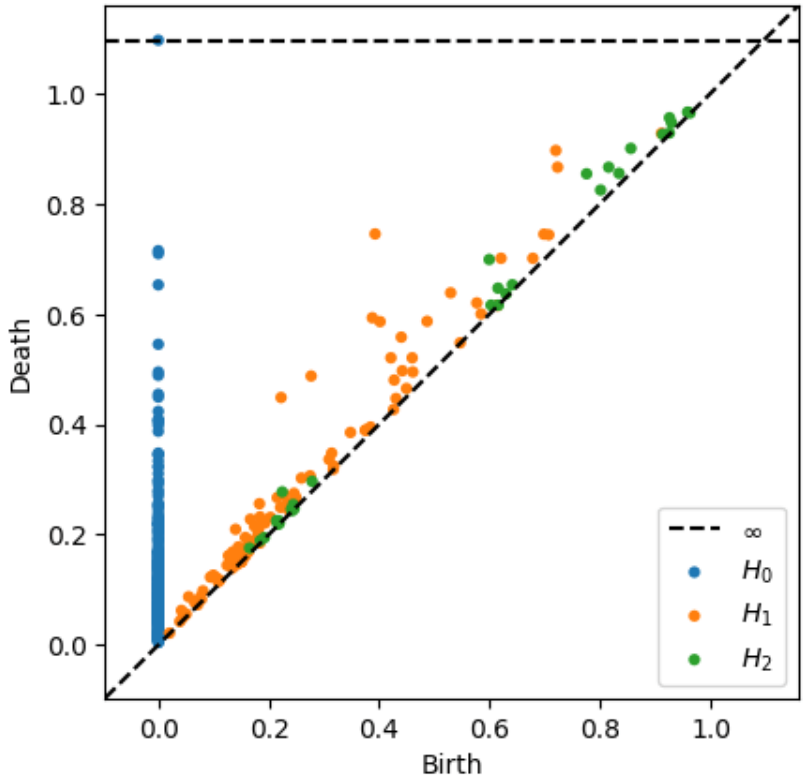


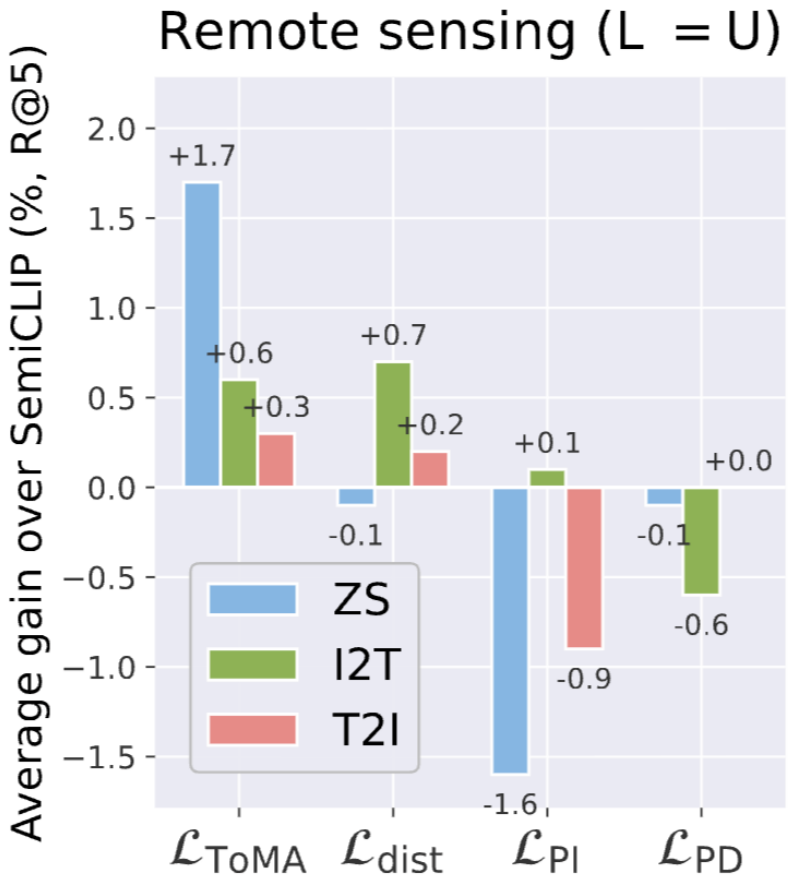
Topology-aware alignment (ours)

ToMA: Quantify the **Contribution of Image-Text Pairs** to the Global Structure (H_0, H_1)



Direct Representation Alignment Beyond Persistence Diagram Matching



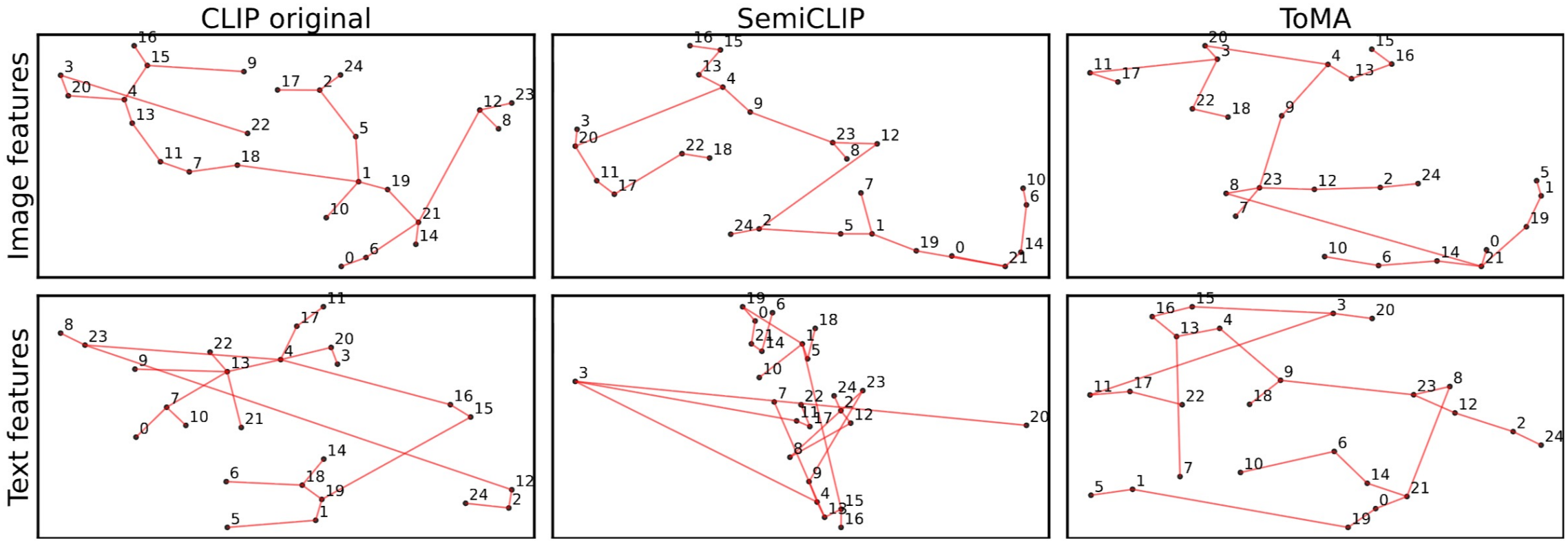


Performance ↑

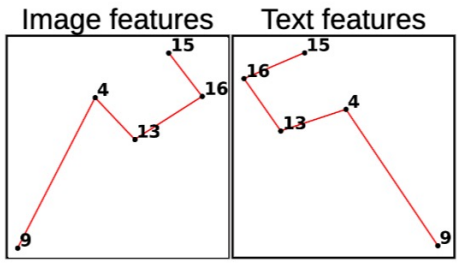
Method	Data	RSICD-CLS	UCM-CLS	WHU-RS19	RSSCN7	AID
CLIP (original)	-	59.2	60.2	81.2	69.0	59.6
CLIP (fine-tuned)	L	75.7±2.2	80.1±6.8	92.2±1.5	70.7±5.5	79.8±3.8
CLIP (oracle)	L + U	87.5±1.8	67.9±2.3	94.2±1.5	75.7±2.9	89.3±1.2
Hard-PL [56]	L=U	78.0±2.0 (+2.3)	72.4±5.2 (-7.7)	92.3±2.6 (+0.1)	71.7±4.1 (+1.0)	81.4±2.2 (+1.6)
Soft-PL [57]		81.7±0.6 (+6.0)	78.9±5.2 (-1.2)	95.3±0.8 (+3.1)	72.0±2.5 (+1.3)	85.8±1.6 (+6.0)
S-CLIP [4]		81.4±1.8 (+5.7)	81.3±3.4 (+1.2)	95.9±1.8 (+3.7)	75.1±2.0 (+4.4)	86.4±1.7 (+6.6)
SemiCLIP [5]		81.4±1.1 (+5.7)	84.3±3.2 (+4.2)	95.6±0.3 (+3.4)	**76.2**±2.7 (+5.5)	84.0±1.2 (+4.2)
ToMA (ours)		**84.8**±1.3 (+9.1)	86.1±3.4 (+6.0)	**97.9**±0.4 (+5.7)	74.2±1.7 (+3.5)	**87.3**±2.2 (+7.5)
ToMA-domain (ours)		83.8±3.2 (+8.1)	**87.5**±4.4 (+7.4)	97.8±0.6 (+5.6)	73.9±1.0 (+3.2)	86.8±2.8 (+7.0)

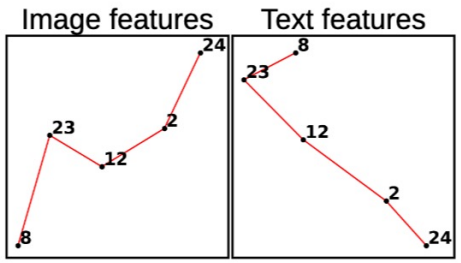
		Image→Text R@5			Text→Image R@5		
Method	Data	RSICD	UCM	Sydney	RSICD	UCM	Sydney
CLIP (original)	-	12.6	46.7	44.8	13.9	39.5	44.8
CLIP (fine-tuned)	L	25.7±2.3	55.2±0.8	46.6±3.0	24.9±0.6	56.3±1.6	49.4±2.0
CLIP (oracle)	L + U	34.3±1.3	63.5±1.3	63.6±2.4	32.0±1.3	66.6±1.1	65.4±2.5
Hard-PL [56]	L=U	27.3±1.1	54.8±1.6	52.3±3.6	24.2±1.3	54.6±1.0	55.2±6.9
Soft-PL [57]		27.3±0.9	54.8±2.2	52.3±2.0	26.1±1.7	55.6±4.0	56.3±5.3
S-CLIP [4]		27.5±1.1	57.0±1.5	51.1±4.3	25.6±0.6	57.3±4.1	50.6±2.6
SemiCLIP [5]		27.9±0.7	57.3±2.6	52.3±4.3	25.8±0.2	59.4±1.2	58.6±2.4
ToMA (ours)		**29.0**±1.5	**61.4**±4.6	48.9±2.1	**27.6**±1.2	60.8±3.0	56.3±3.5
ToMA-domain (ours)		28.7±1.2	59.2±3.9	**54.0**±2.9	27.0±0.9	**63.0**±2.5	**62.6**±2.9

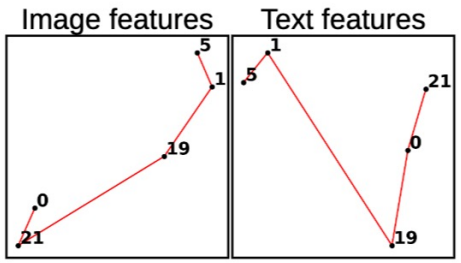
Representational Alignment ↑

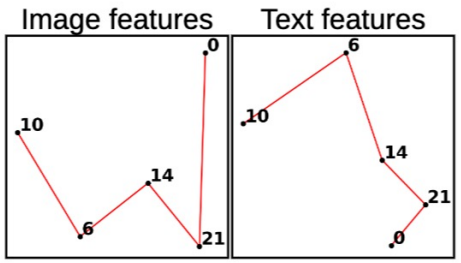


Representational Alignment ↑











T_{15}: a parking lot with many cars is near a large stadium .



T_{24}: Lots of houses arranged neatly in the residential area .



T_5: A road go across another roads with some cars on the roads and surrounded by plants .



T_{10}: many green plants and barelands are in two sides of a curved river .



T_{16}: a parking lot and a road are near a large stadium .



T_2: It is a medium residential area .

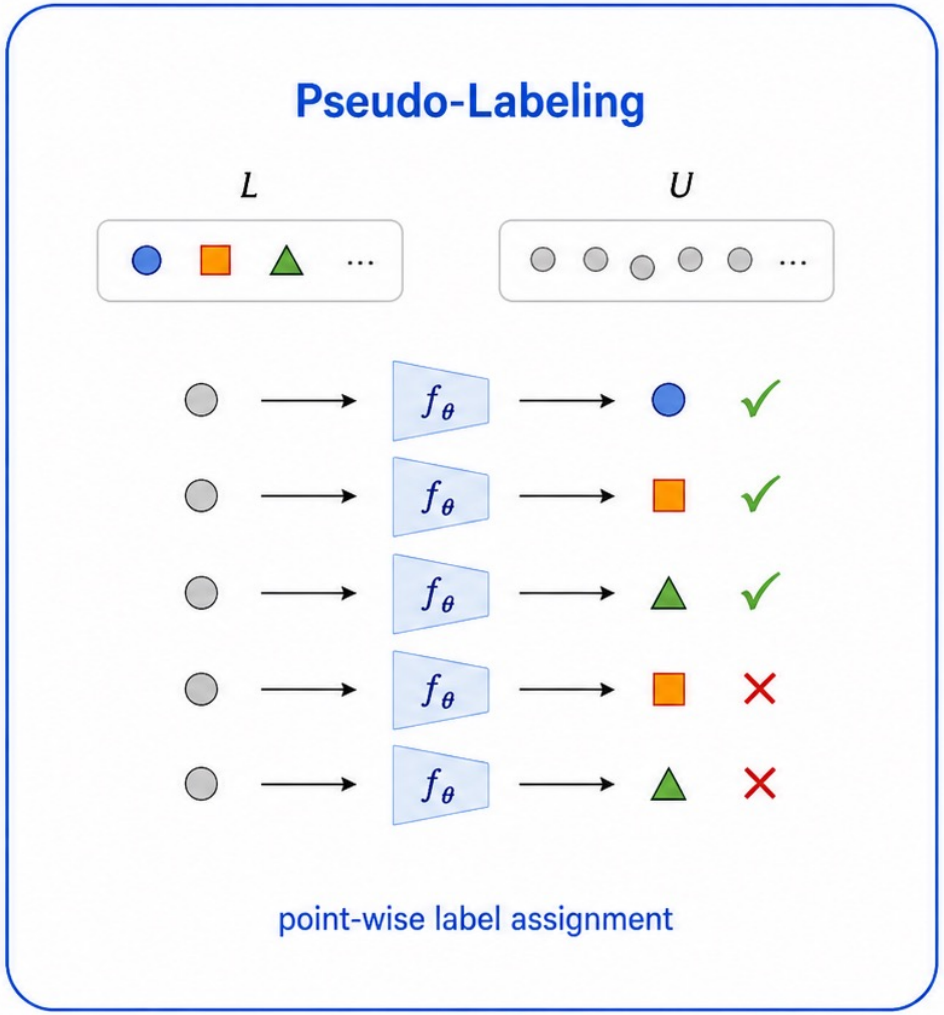


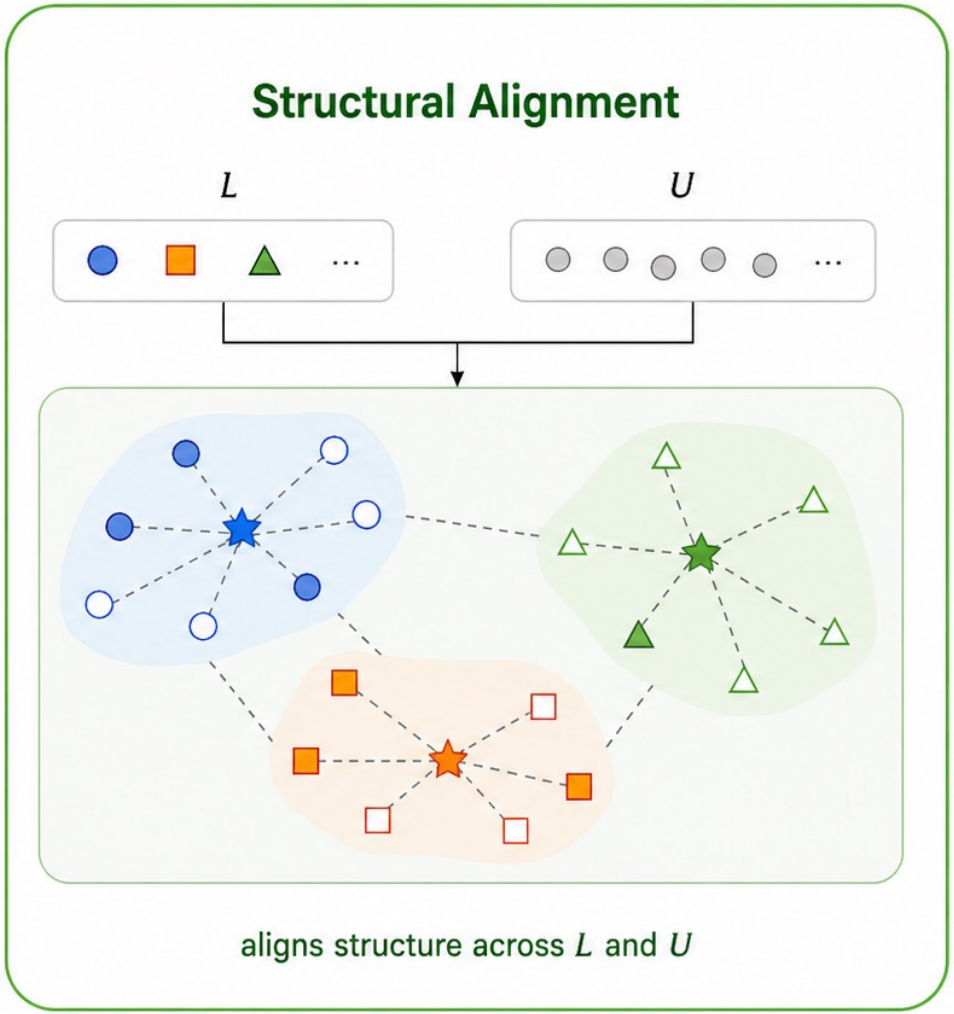
T_1: many cars are on two unparallel bridges over a river .



T_6: the desert with some dark rocks has orange stripe .

Geometry and Topology Can Improve VLM Semi-SL





+ Optimal Transport (S-CLIP)

+ Topology-Aware Alignment (ToMA)