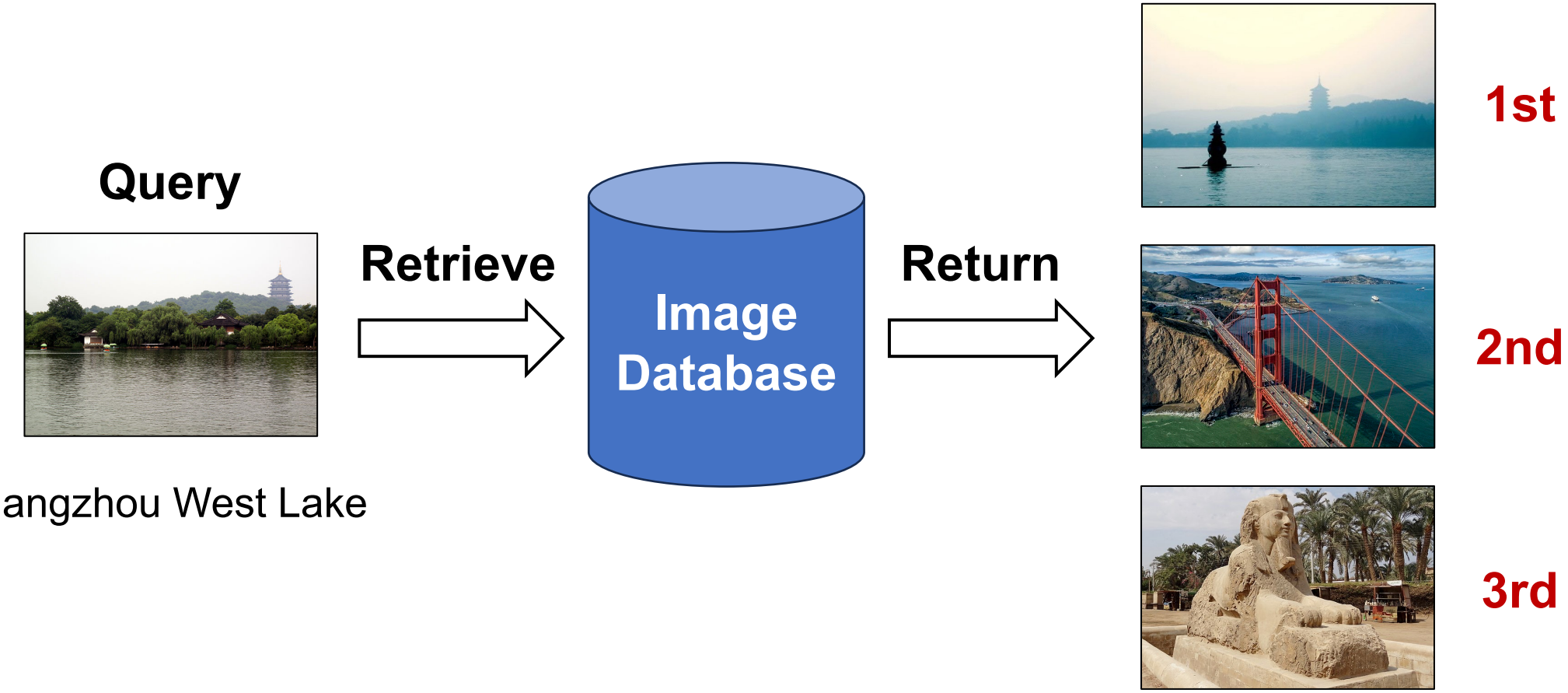
Let All Be Whitened: Multi-teacher Distillation for Efficient Visual Retrieval

Zhe Ma, Jianfeng Dong, Shouling Ji, Zhenguang Liu, Xuhong Zhang, Zonghui Wang, Sifeng He, Feng Qian, Xiaobo Zhang, Lei Yang

2024

Background

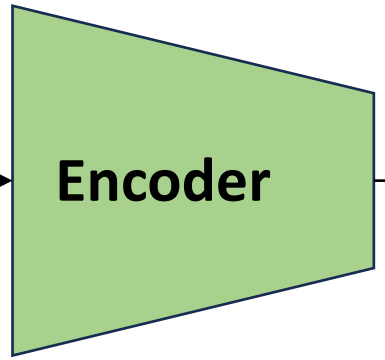
Ranking-based retrieval: **retrieve similar data with the higher similarity scores under some metric.**

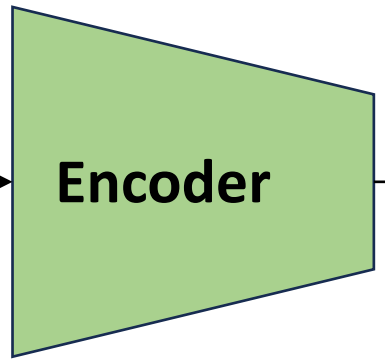


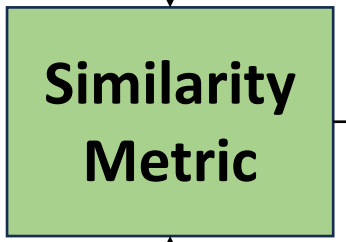
Background

Ranking-based retrieval: retrieve similar data with the highest similarity scores under some metric.

Hangzhou West Lake







Similarity Score

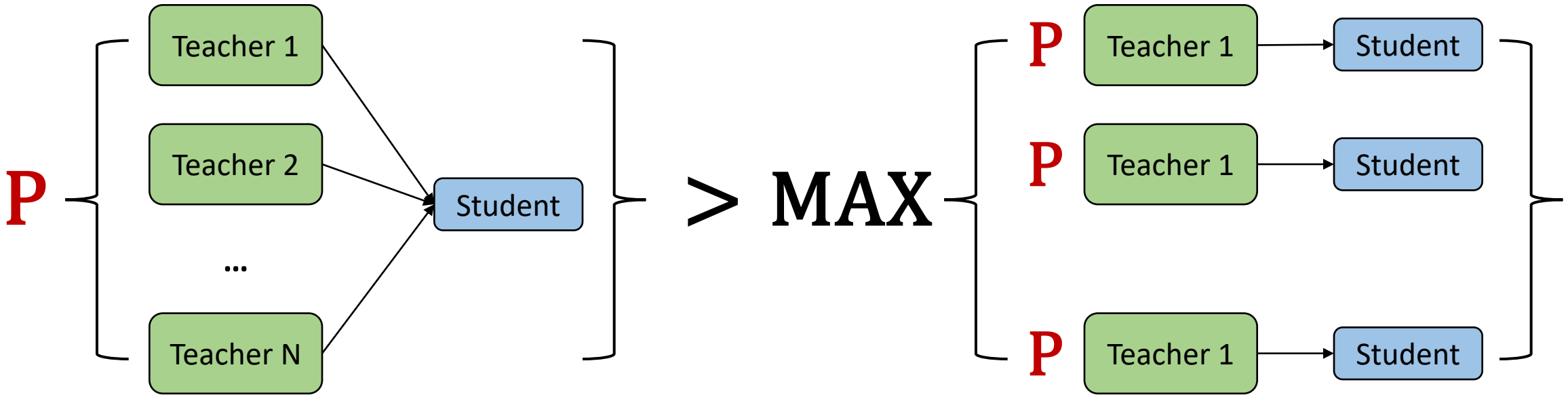
Background

Multi-teacher distillation for a balance between effectiveness and efficiency

Background

Multi-teacher distillation for a balance between effectiveness and efficiency

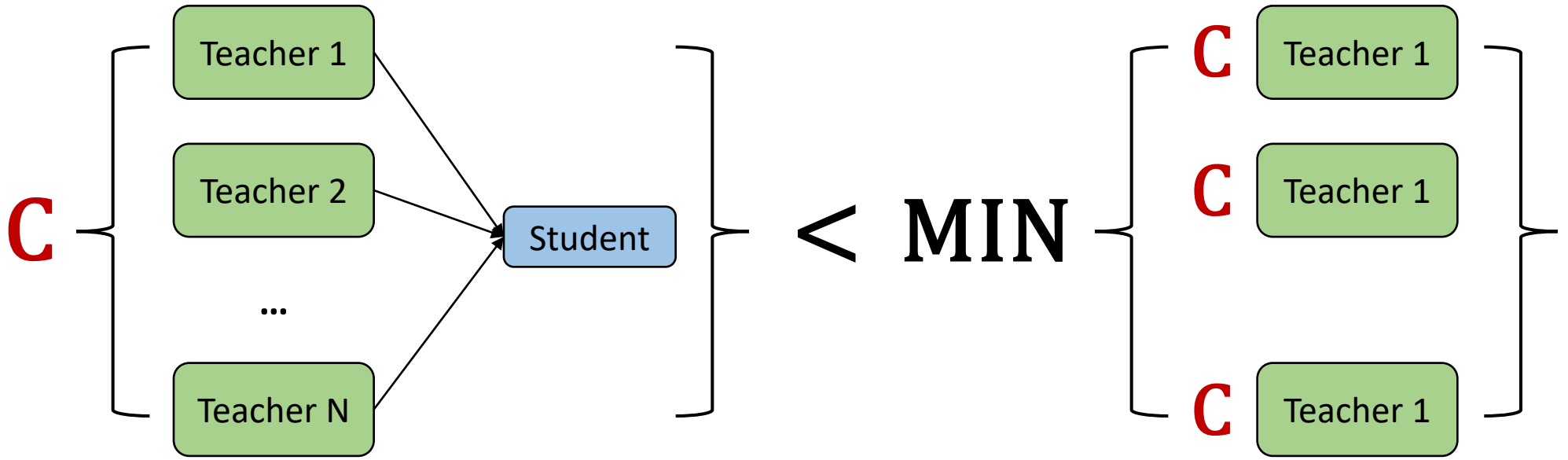
- Through multi-teacher distillation, we anticipate better **P**erformance than single model distillation.



Background

Multi-teacher distillation for a balance between effectiveness and efficiency

- Through multi-teacher distillation, we anticipate better performance than single model distillation.
- The student model is more **C**omputational efficient than initial teacher models.



Method

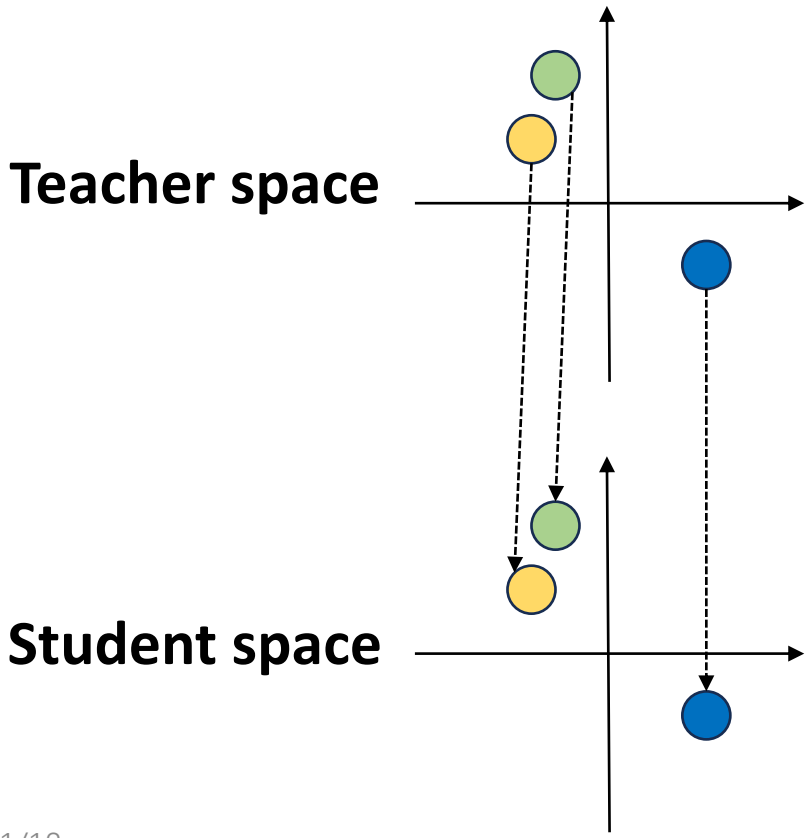
Similarity-based Knowledge Distillation

What type of knowledge to transfer?

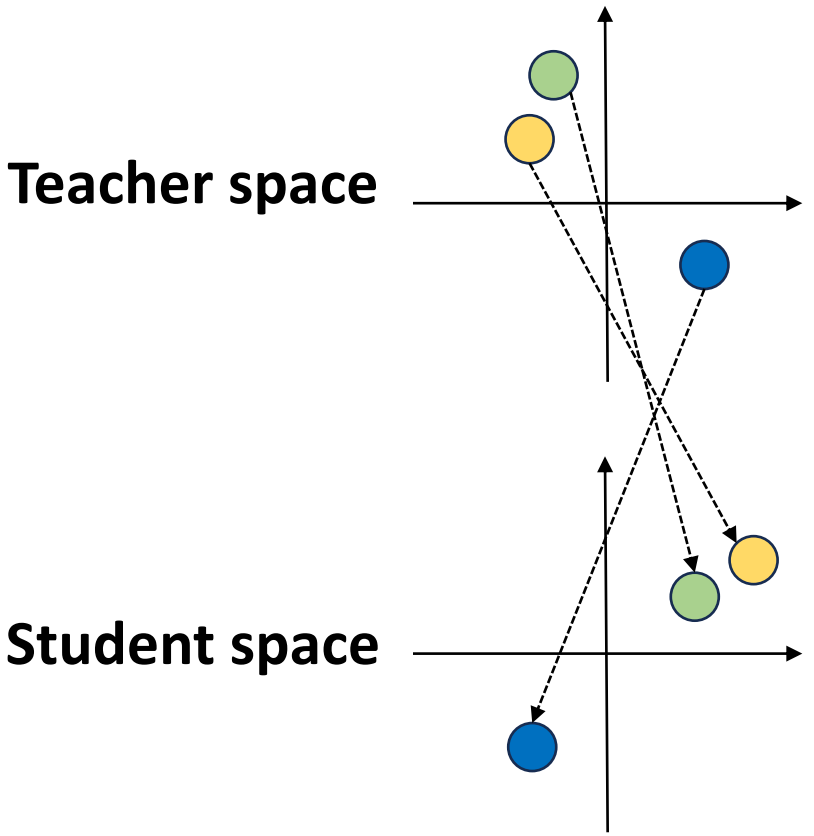
relevant samples

irrelevant samples

Feature-based distillation



Similarity-based distillation



Similarity-based Knowledge Distillation

Typical objectives to train a retrieval model

- Given some distance function $d(x_1, x_2)$ or similarity estimation function $s(x_1, x_2)$
- Anchor sample x, positive sample p (relevant to x), negative sample n (irrelevant to x)

Contrastive Loss

$$\mathcal{L} = \mathbb{E}[d(x,p) + \max(m - d(x,n), 0)]$$

Triplet Margin Loss

$$\mathcal{L} = \mathbb{E}[\max(m + d(x,p) - d(x,n), 0)]$$

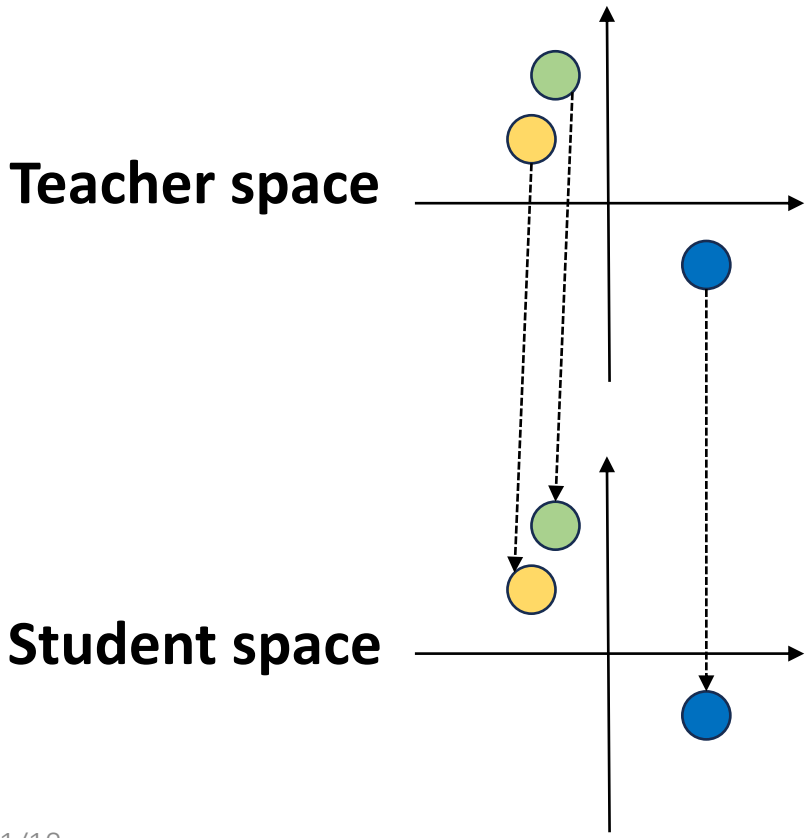
InfoNCE-like Loss

$$\mathcal{L} = \mathbb{E}[-\log \frac{e^{s(x,p)/\tau}}{e^{s(x,p)/\tau} + \sum_n e^{s(x,n)/\tau}}]$$

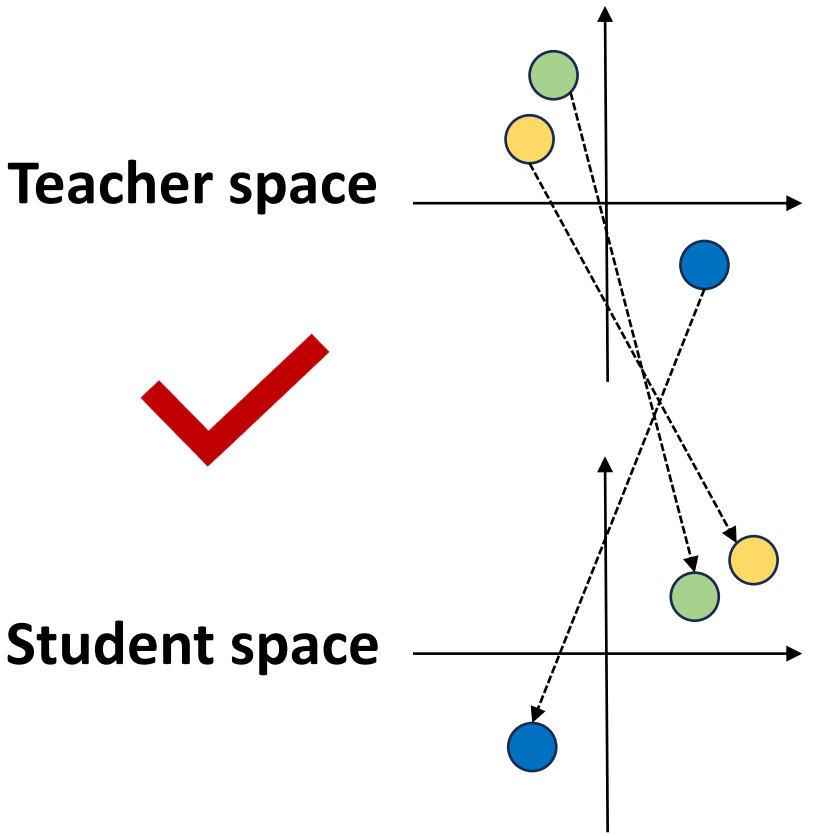
Similarity-based Knowledge Distillation

What type of knowledge to transfer? **Only relationship matters.**

Feature-based distillation

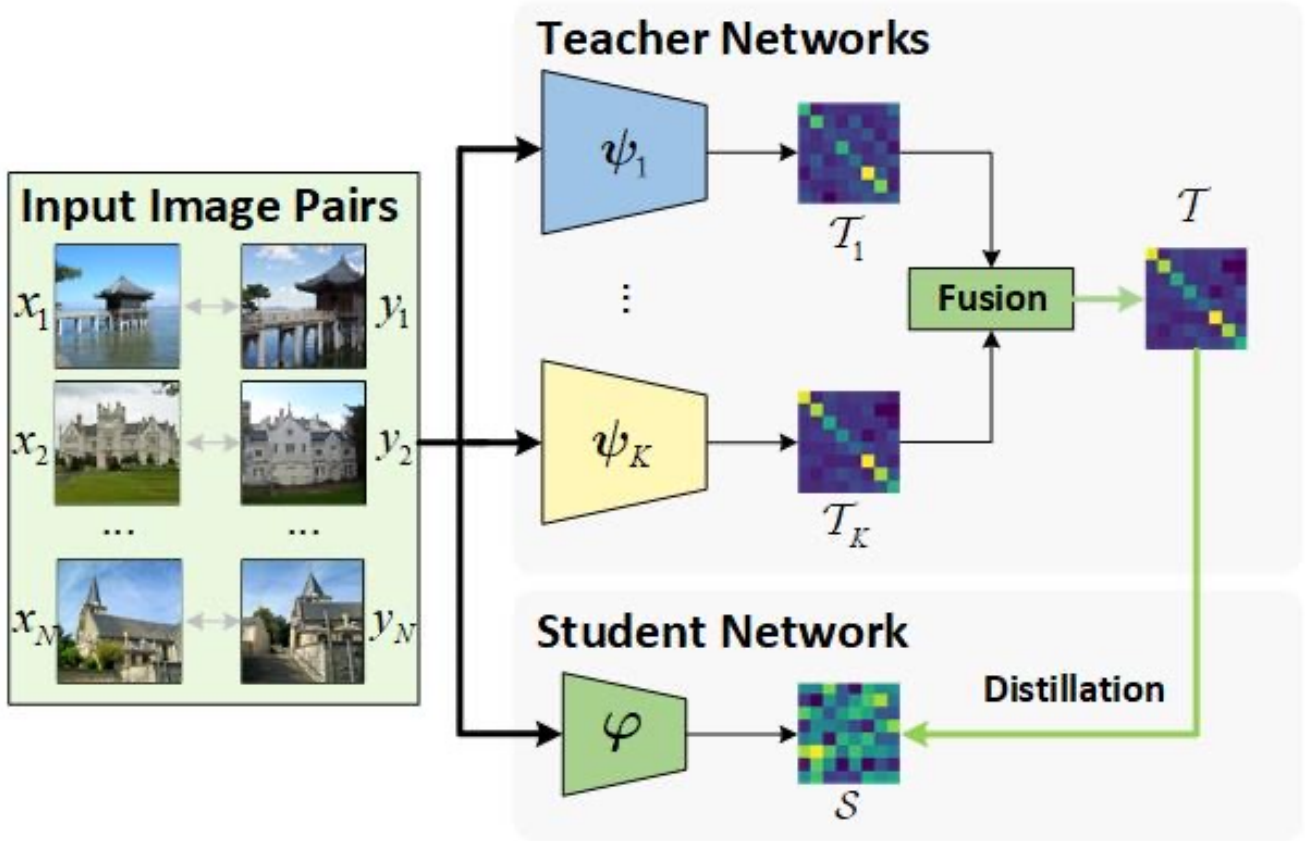


Similarity-based distillation



Similarity-based Knowledge Distillation

A simple multi-teacher distillation framework by aggregating similarities.

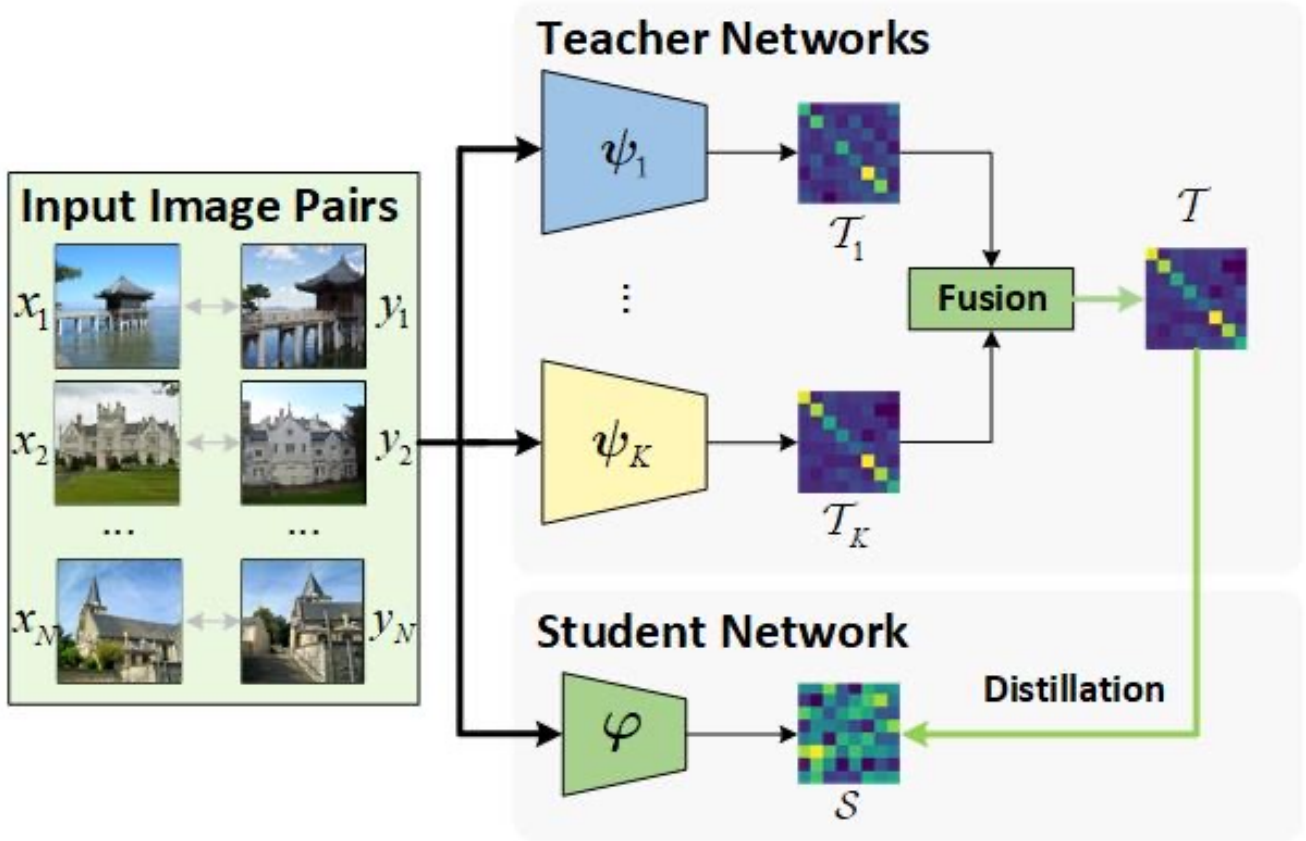


1. $\mathcal{T}_i$ is the similarity matrix predicted by the teacher model i.
2. $\mathcal{S}$ is the similarity matrix predicted by the student model.
3. $\mathcal{T}$ is a fusion of teachers' predictions.

$$\mathcal{L}_{distill} = \frac{1}{N}\sum_{i}^{N} KL(\mathcal{T}[i,:], \mathcal{S}[i,:])$$

Similarity-based Knowledge Distillation

A simple multi-teacher distillation framework by aggregating similarities.



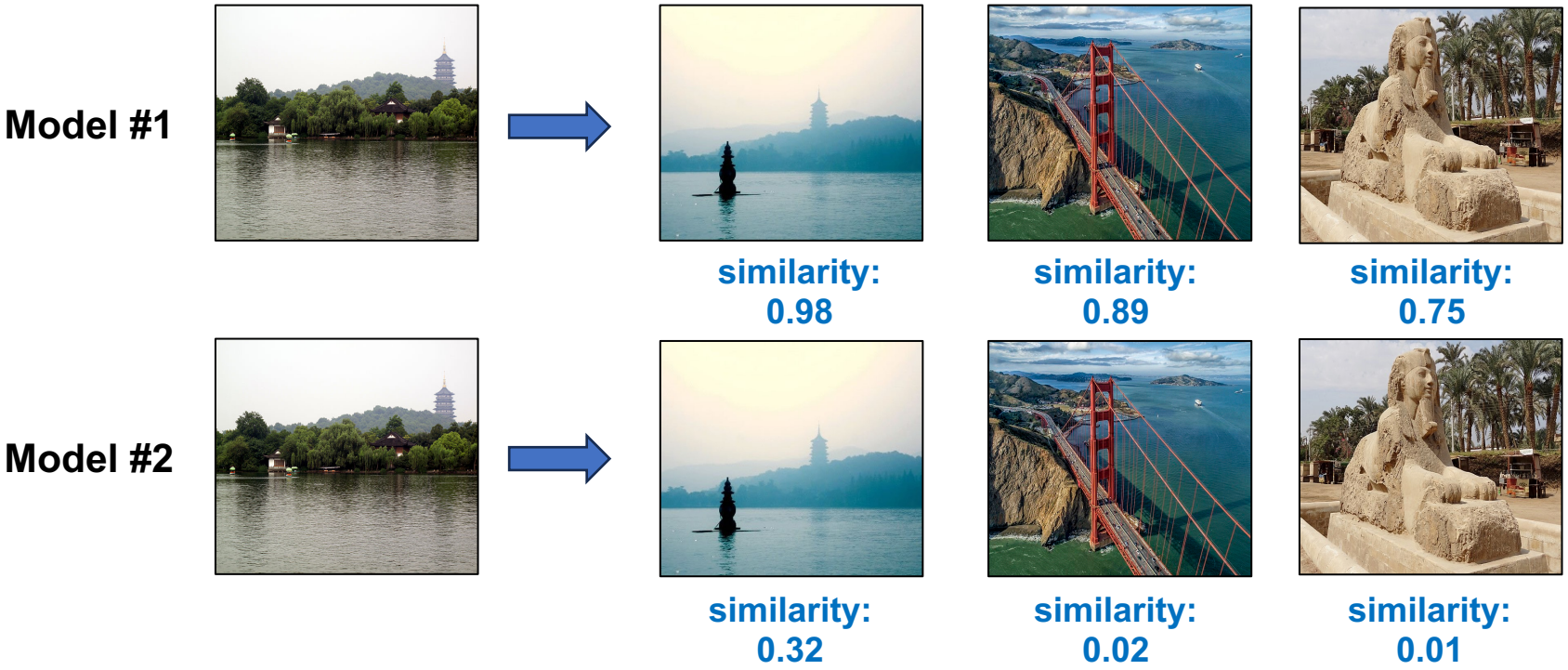
Heuristic Fusion Strategies:

- ***mean***: $\mathcal{T}[i,j] = \frac{1}{K}\sum_{k=1}^{K}\mathcal{T}_k[i,j]$
- ***rand***: $\mathcal{T}[i,j] = \mathcal{T}_r[i,j], r \in [1,K]$

- Always average diagonal similarity scores, for off-diagonal ones ($i \neq j$):
- ***max-min***: $\mathcal{T}[i,j] = \min_{k\in[1,K]}\mathcal{T}_k[i,j]$
- ***max-mean***: $\mathcal{T}[i,j] = \frac{1}{K}\sum_{k=1}^{K}\mathcal{T}_k[i,j]$
- ***max-rand***: $\mathcal{T}[i,j] = \mathcal{T}_r[i,j], r \in [1,K]$

Incommensurability of Retrieval Models

However, it does not work well……

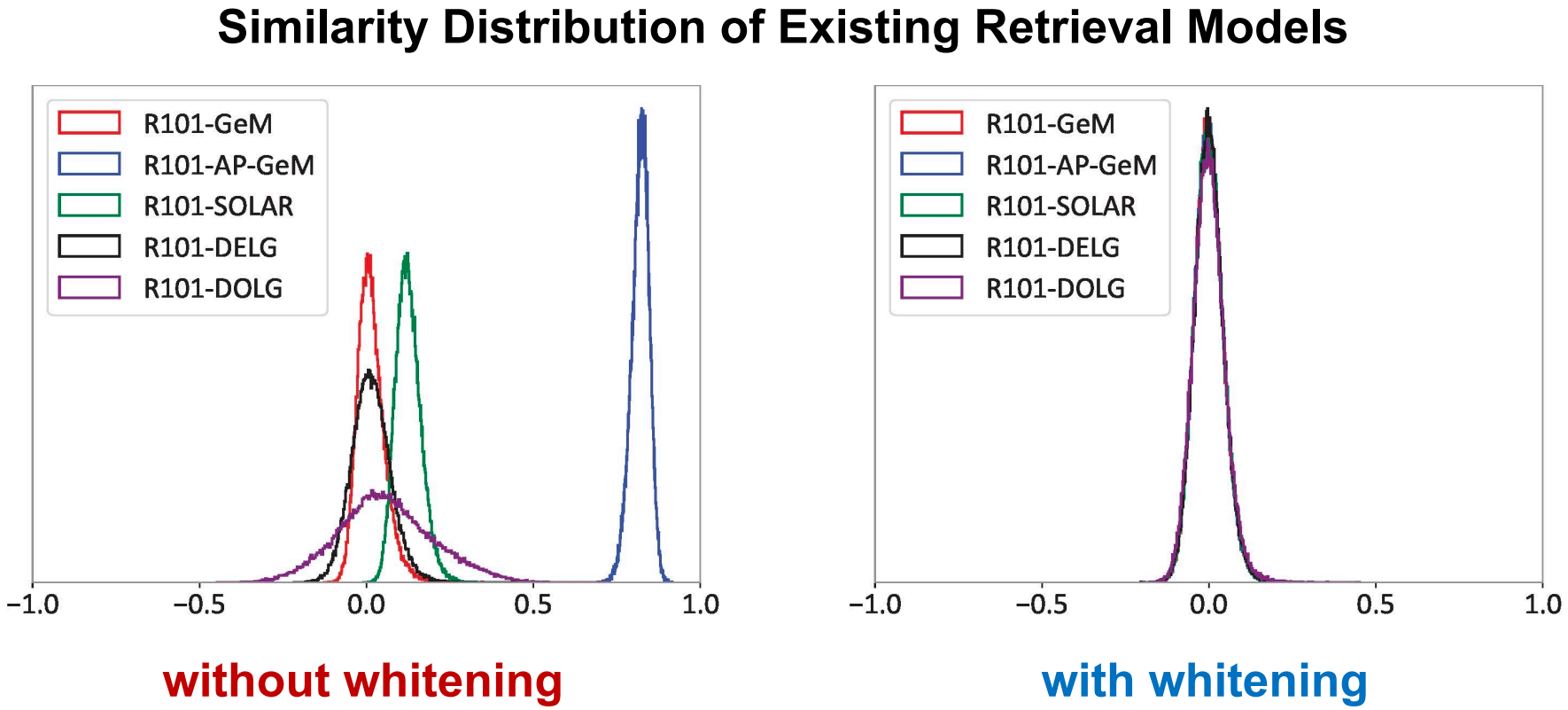


Question: Both model #1 and #2 correctly return an image of West Lake as the query image, which one is better?

Answer: Different models make decision based on their own measure, who cannot be compared directly.

Incommensurability of Retrieval Models

Whitening can transform representation of any model into spherical distribut ion, which leads to deterministically the same similarity distribution.



Whitening

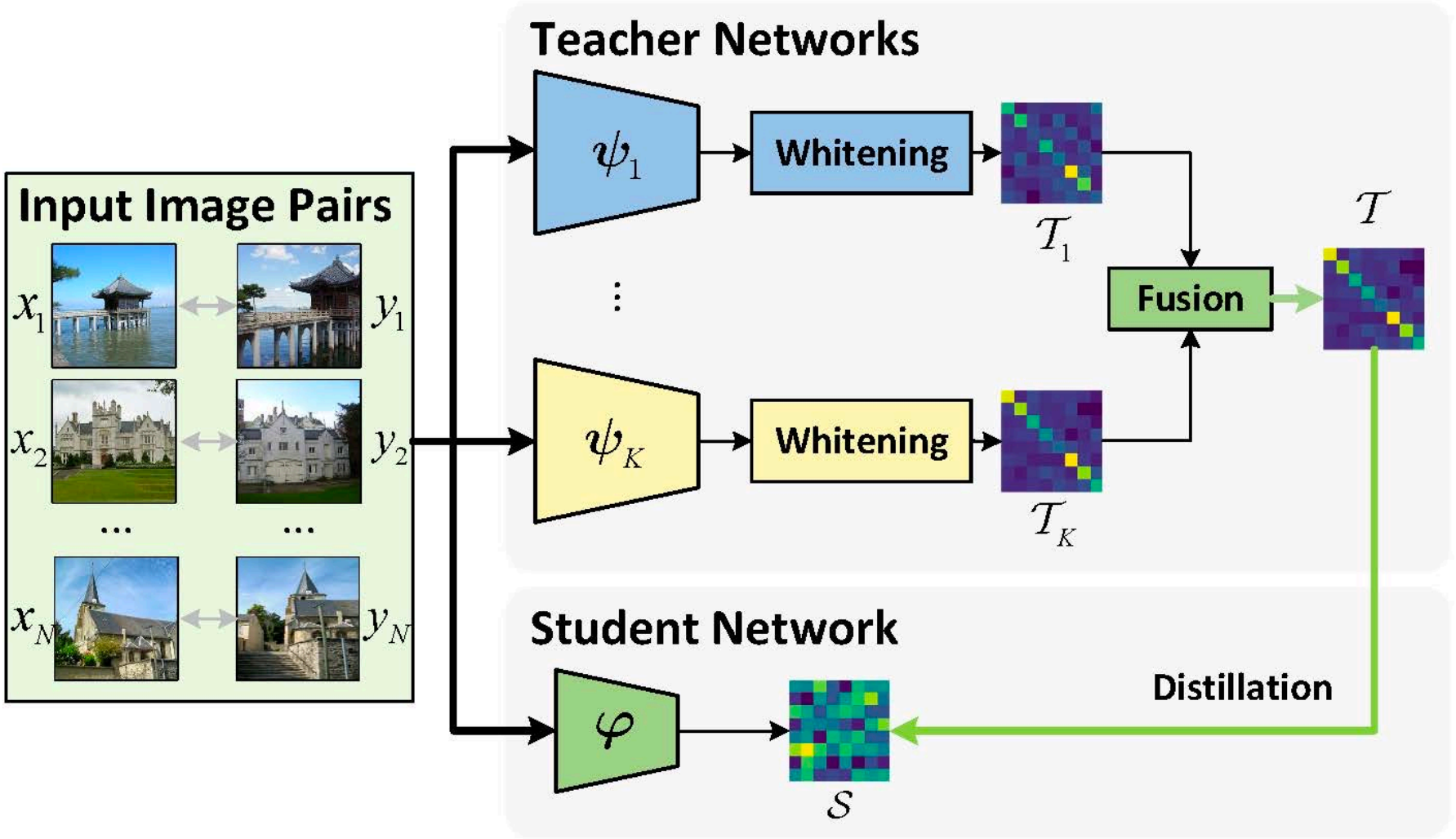
1. Teacher model outputs l_2-normalized representation $\boldsymbol{\psi}$.
2. Whitening transform $\boldsymbol{\psi}$ into a new representation $\boldsymbol{\psi_w}$:

$$\boldsymbol{\psi_w} = \boldsymbol{W}(\boldsymbol{\psi} - \mathbb{E}[\boldsymbol{\psi}]), \boldsymbol{W^T W} = \boldsymbol{\Sigma^{-1}}$$

3. Further l_2 normalize $\boldsymbol{\psi_w}$ into $\bar{\boldsymbol{\psi}}_w$.
4. $\bar{\boldsymbol{\psi}}_w$ distribute uniformly on the surface of unit sphere, i.e., the whitened re presentation space has equal density everywhere.
5. Cai et al.[1] proved that the angle between two independent random vector s distributed uniformly on the unit sphere surface converges to a distributio n with the probability density function $f(\theta) = \frac{1}{\sqrt{\pi}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \cdot (\sin\theta)^{n-2}, \theta \in [0, \pi]$.

[1] Cai, T. T.; Fan, J.; and Jiang, T. 2013. Distributions of angles in random packing on spheres. J. Mach. Learn. Res., 14(1): 1837–1864.

Modified Multi-teacher Distillation Framework



Evaluations

Experiments

- **Landmark image retrieval**
  - Influence of whitening on teacher models
  - Effectiveness of whitening
  - Comparison to optional multi-model aggregation approaches
  - Comparison to state-of-the-art

- **Near-duplicate video retrieval**
  - Comparison to state-of-the-art

Experiment Settings

➢ Landmark image retrieval

- ➢ Dataset: train on Google Landmark V2 (1.6M), evaluate on Rparis6k(+1M) and ROxford5k(+1M).
- ➢ Teacher models: R(esNet)101-GeM, R101-AP-GeM, R101-SOLAR, R101-DELG, R101-DOLG
- ➢ Student architecture: ResNet-18/34

➢ Near-duplicate video retrieval

- ➢ Dataset: Short Video Dataset(0.56M).
- ➢ Teacher models: R50-MoCoV3, R50-BarlowTwins
- ➢ Student architecture: ResNet-18/34

Influence on Teacher Models

- **Whitening has little influence on the distillation performance of teacher models.**

Method	Whitening?	$\mathcal{R}$Oxf		$\mathcal{R}$Par	
		M	H	M	H
R101-GeM	✗	69.88	45.00	82.69	65.13
	✓	68.97	45.08	82.15	65.02
R101-AP-GeM	✗	69.17	44.08	80.44	62.08
	✓	70.37	45.93	81.02	63.29
R101-SOLAR	✗	71.14	46.63	83.04	66.24
	✓	68.52	43.65	81.82	64.39
R101-DELG	✗	84.43	66.69	91.97	82.87
	✓	85.28	68.52	92.08	83.26
R101-DOLG	✗	80.09	61.64	88.68	77.01
	✓	83.27	64.95	89.78	78.58

Fusion Strategies

- **Whitening consistently improves and stabilize multi-teacher distillation.**
- **max-min performs the best among five heuristic strategies.**

Strategy	Whitening?	$\mathcal{R}$Oxf		$\mathcal{R}$Par	
		M	H	M	H
mean	✗	69.62	44.01	81.62	63.64
rand	✗	65.68	38.89	75.78	55.09
max-min	✗	67.63	40.78	82.00	65.01
max-mean	✗	67.09	40.07	83.66	66.93
max-rand	✗	71.53	46.71	80.26	62.15
mean	✓	71.11	46.38	82.55	65.62
rand	✓	71.01	46.39	82.06	64.01
max-min	✓	**74.67**	**50.69**	**84.48**	**68.34**
max-mean	✓	73.90	49.53	83.42	66.55
max-rand	✓	72.53	48.10	82.60	65.46

Single/Double/Triple Teacher Distillation

➢ **Whitening brings incremental performance gain in multi-teacher distillation.**

Method	Teacher models	Params (M)	GFLOPs	$\mathcal{R}$Oxf		$\mathcal{R}$Par		$\mathcal{R}$Oxf+1M		$\mathcal{R}$Par+1M	
				M	H	M	H	M	H	M	H
R101-GeM	-	46.70	124	68.97	45.08	82.15	65.02	55.68	29.24	61.73	35.07
R101-AP-GeM	-	46.70	124	70.37	45.93	81.02	63.29	57.48	32.01	61.26	35.59
R101-SOLAR	-	56.15	139	68.52	43.65	81.82	64.39	56.24	29.65	61.53	35.12
Single-teacher distillation	R101-GeM	11.44	28.62	70.95	46.42	81.68	64.03	55.82	28.77	61.04	33.90
	R101-AP-GeM			71.41	46.46	80.64	62.82	57.61	31.17	59.67	33.57
	R101-SOLAR			69.78	45.12	81.78	63.76	55.87	28.84	60.63	34.04
Double-teacher distillation	R101-GeM, -AP-GeM	11.44	28.62	73.90	49.81	84.09	67.91	61.12	33.72	64.44	38.89
	R101-GeM, -SOLAR			70.79	46.95	82.48	64.93	56.20	29.41	62.57	36.13
	R101-AP-GeM, -SOLAR			**74.71**	50.21	83.59	67.01	60.93	33.01	64.28	39.53
Triple-teacher distillation	R101-GeM, -AP-GeM, -SOLAR	11.44	28.62	74.67	**50.69**	**84.48**	**68.34**	**61.74**	**34.49**	**64.82**	**39.63**

Optional Multi-Model Aggregation Approaches

- **Whitening-based multi-teacher distillation shows the best trade-off between performance and efficiency.**

Method	Params (M)	GFLOPs	$\mathcal{R}$Oxf		$\mathcal{R}$Par	
			M	H	M	H
EM	149.55	387	71.14	46.67	83.38	66.41
ED	11.44	28.62	68.19	42.29	80.47	61.73
CL	11.44	28.62	65.72	40.25	81.78	63.69
Whiten-MTD	11.44	28.62	**74.67**	**50.69**	**84.48**	**68.34**

EM: Ensemble Mean

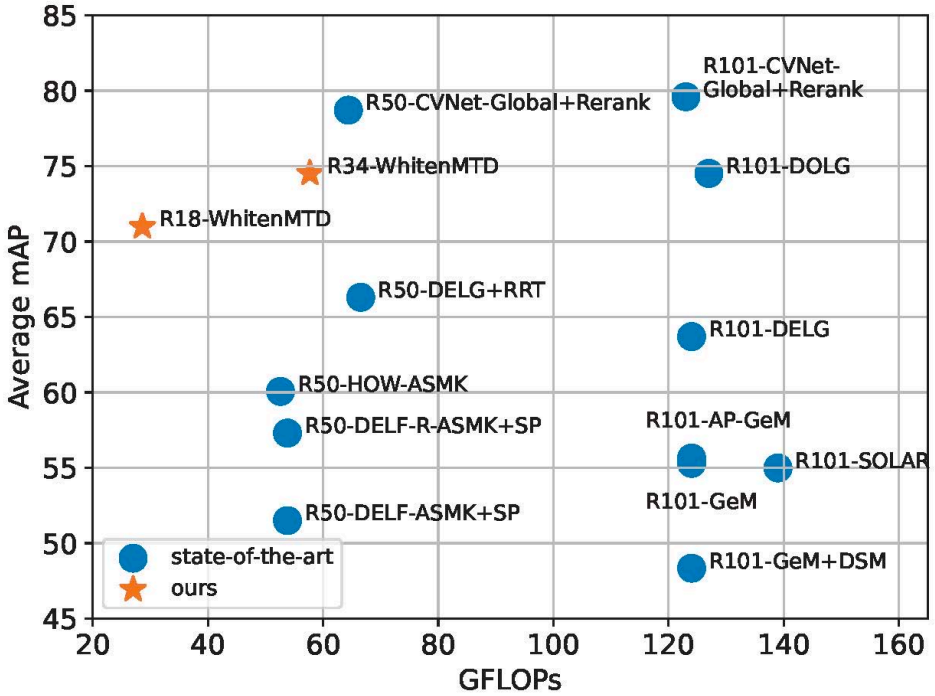
ED: Embedding Distillation

CL: Contrastive Learning

Comparison to SOTA

➢ **Whitening-based multi-teacher distillation has a good trade-off between performance and efficiency.**

Landmark image retrieval



Near-duplicate video retrieval

Method	Params (M)	GFLOPs	mAP@100	mAP
VGG16-CNNL (2017a)	134	15.47	61.04	55.55
VGG16-CNNV (2017a)	134	15.47	25.10	19.09
VGG16-CTE (2013)	134	15.47	-	50.97
VGG16-DML (2017b)	139	15.47	81.27	78.47
R50-VRL (2022)	23.5	4.14	86.00	-
R50-DnS (2022)	27.55	4.13	-	**90.20**
R50-MoCoV3 (2021)	23.5	4.14	87.31	85.47
R50-BarlowTwins (2021)	23.5	4.14	87.22	84.80
R18-Whiten-MTD (ours)	11.2	1.83	88.62	86.82
R34-Whiten-MTD (ours)	21.3	3.68	**88.84**	86.78

Discussion

Discussion

- **Multi-teacher distillation for retrieval models**
  - A simple similarity-based multi-teacher distillation framework.
  - Five heuristic fusion strategies.
  - Whitening-based elimination of teacher models' distribution discrepancy.

- **Limitation and future work**
  - Additional training cost, which could be mitigated by offline caching teachers' outputs.
  - Heuristic fusion strategies might not be optimal.
  - More effective normalization techniques to tackle the problem of distribution discrepancy.

mz.rs@zju.edu.cn