ICML
International Conference
On Machine Learning


POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

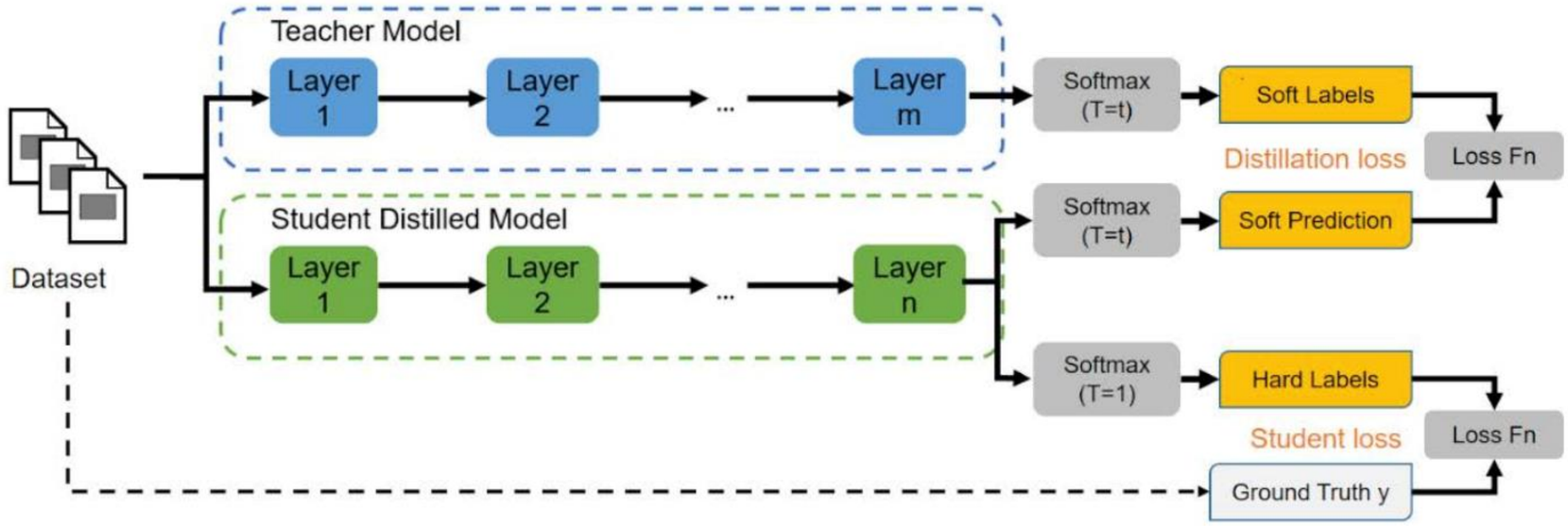
# Do Topological Characteristics Help in Knowledge Distillation?

*Jungeun Kim*, Junwon You*, Dongjin Lee*, Ha Young Kim, Jae-Hun Jung*
*Yonsei Univerisity & POSTECH*

# Motivation

## Knowledge distillation (KD)

- Transfer knowledge from larger (teacher) to smaller (student) networks.
- Model compression (large network to a small network).
- Determine which knowledge to transfer.

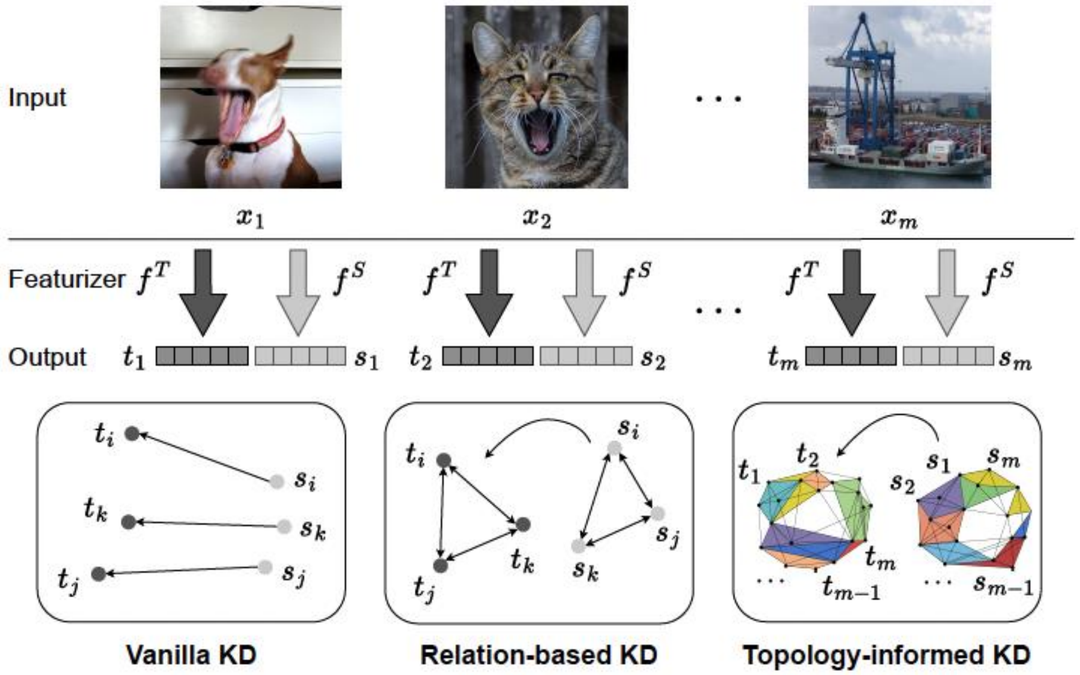



Hong, Yu-Wei & Leu, Jenq-Shiou & Faisal, Muhamad & Prakosa, Setya. (2022). Analysis of Model Compression Using Knowledge Distillation. IEEE Access.

# Motivation

## Knowledge distillation (KD)

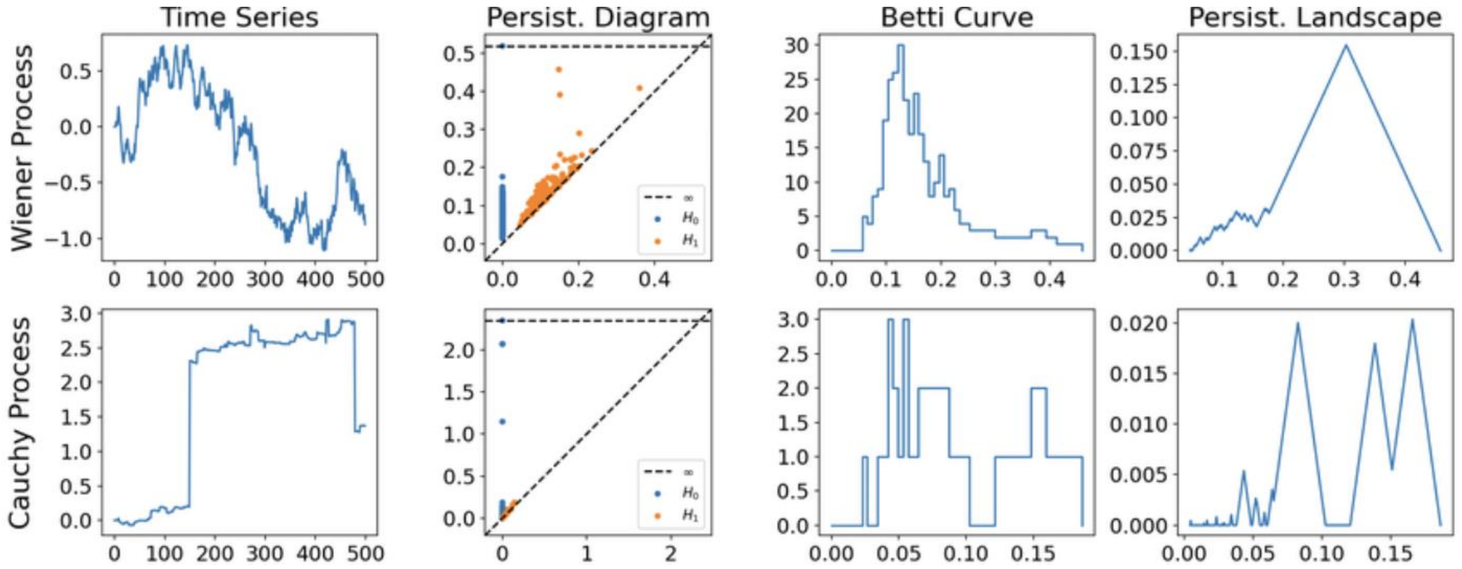
- Relation-based KD
  - The structure or similarity between two or three pairs of embedding features is distilled.
  - They teach the model the full relationship via interactions between few embedding features.
- Thus, **the broader context of relationships between all embedding features** should be defined.



# Motivation

## Persistent homology (PH)

- Primary method in topological data analysis (TDA), provides an efficient method of calculating the topological structure of point cloud data (PCD). Persistence of the topological features captures global topology of PCD.
- Comprehensive structural information (e.g. shape of distribution, multiscale structure, connectivity.)
- **Persistence diagram (PD)** : a method to visualize (summarize) birth and death of the topological features with the resolution.



# Motivation

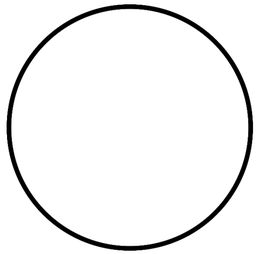
## Persistent homology in DL

- Numerous efforts have been made to integrate topological information into machine learning for geometric data analysis.
- It is not straightforward to feed PD into DNNs, so convert a PD into a fixed-size vector (Betti-sequence, persistence landscape, and persistence image).
- Limited experiments to domain-specific datasets or small-scale datasets.
- PH has only been used for extracting information to augment the input data.

# Background of persistent homology

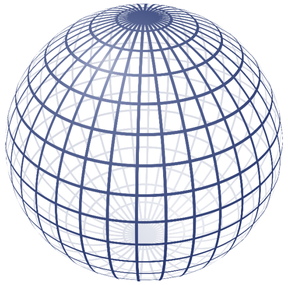
## Topological characteristics

- *Consistent properties of space that persist through continuous transformations*, providing insights into the structure, shape, connectivity, overall distribution within datasets.
  - *k-dimensional (k-dim) holes* : Connected components (0-dim), Loops (1-dim), Voids (2-dim).
  - The rank of *k-dim homology group* : the number of *k-dim holes*.
  - *PH* : a method that analyzes the creation and destruction of these topological characteristics by exploring the homology groups of spaces across different scales.
  - *PD* : visual tool that illustrates the results of PH.



1-dim hole

(Loops)



2-dim hole

(Void spaces)


Continuous transformations of a trous

Image from https://en.wikipedia.org/wiki/Homology_(mathematics) https://cerla.univ-lille.fr/projets-de-recherche/topological-photonics/

# Background of persistent homology

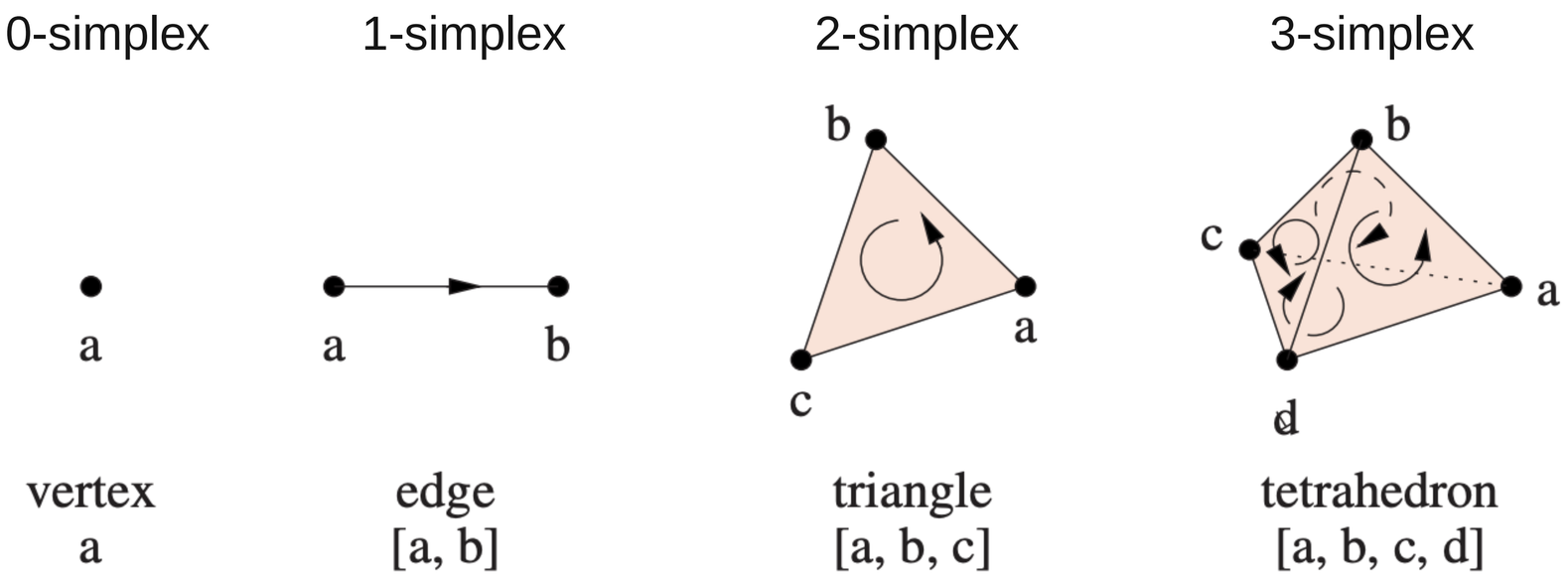
## Typical pipeline for using persistent homology as an input of neural network

- Convert a point cloud into a ***simplicial complexes*** *with different resolutions.*
- Construct a ***filtration***, a nested sequence of simplicial complexes.
- Compute the birth and death times of *k*-dim holes with respect to the filtration and summarize into the form of a ***PD.***
- ***Vectorize*** the PD and feed to a neural network.

# Background of persistent homology

## Simplex

- A generalization of the notion of a triangle or tetrahedron to arbitrary dimensions.




Zomorodian, Afra, and Gunnar Carlsson. Computing persistent homology. Discrete & Computational Geometry 33.2 (2005)

# Background of persistent homology

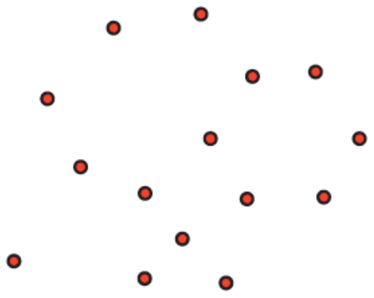
## Simplicial complex

A set of simplices that satisfies the following conditions:

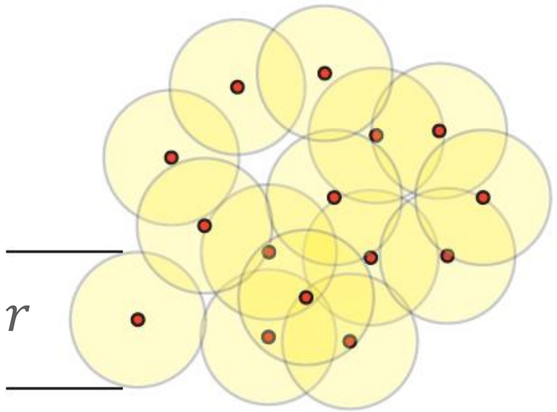
(i) Every face of a simplex from $\mathcal{K}$ is also in $\mathcal{K}$.

(ii) The nonempty intersection of any two simplices $\sigma_1, \sigma_2 \in \mathcal{K}$ is a face of both $\sigma_1$ and $\sigma_2$.

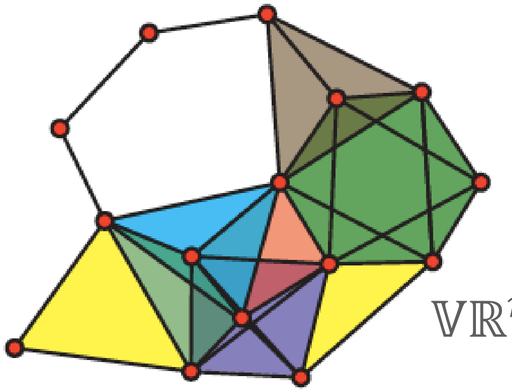
- $k$-dim homology group can be computed on a given simplicial complex.
- e.g., *Vietoris-Rips (Rips) complex:* $\mathbb{VR}^r(P) = \{\sigma = \{p_0, \dots, p_k\} \mid d_u(p_i, p_j) \leq 2r\}$.



Point cloud $P$




Balls with radius of $r$




Rips complex w.r.t $r$

https://mathtuition88.com/2017/01/20/introduction-to-persistent-homolgy-cech-and-vietoris-rips-complex/

# Background of persistent homology

## Filtration

- A nested sequence of simplicial complexes $\emptyset = \mathcal{X}_{\alpha_0=-\infty} \hookrightarrow \mathcal{X}_{\alpha_1} \hookrightarrow \cdots \hookrightarrow \mathcal{X}_{\alpha_n} = \mathcal{X}$.
- e.g., *Rips filtration*

$$\left\{\mathbb{VR}^{\alpha}(P) \hookrightarrow \mathbb{VR}^{\alpha'}(P)\right\}_{\alpha \leq \alpha'}.$$

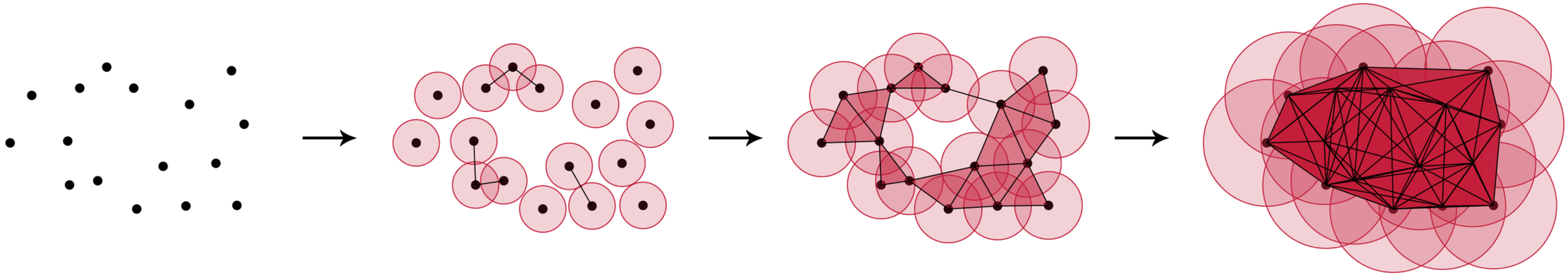


Image from https://christian.bock.ml/posts/persistent_homology/

# Background of persistent homology

## *P*ersistent homology

- A method that analyzes the creation and destruction of *k*-dim holes over filtrations.

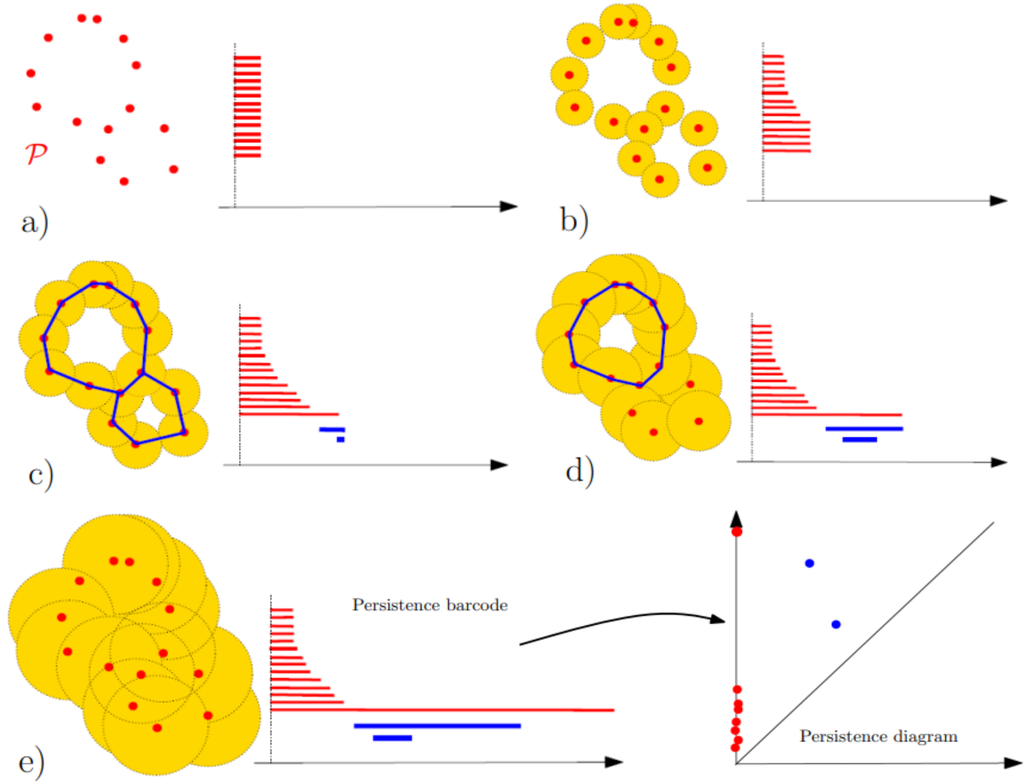
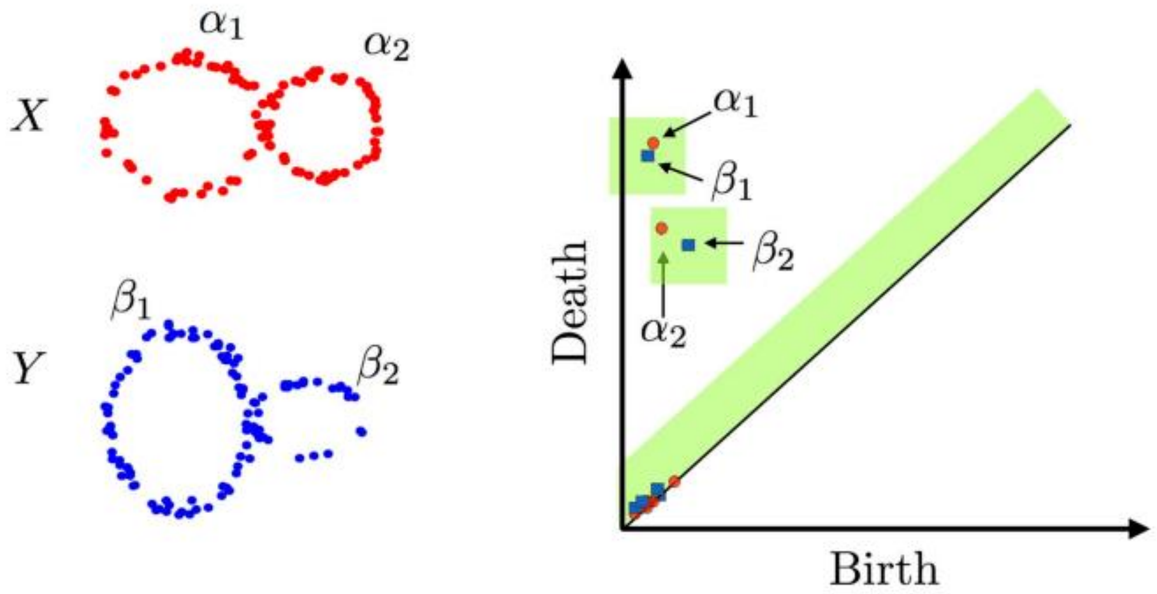



Image from https://www.math.ens.psl.eu/~eaamari/teaching/2022-2023/M2_Jussieu/Lesson%205.pdf

# Background of persistent homology

## *P*ersistence diagram

- A multiset $\{(\alpha_b, \alpha_b)\}$ where $\alpha_b$ and $\alpha_d$ is the birth and the death of a $k$-dim hole, respectively.

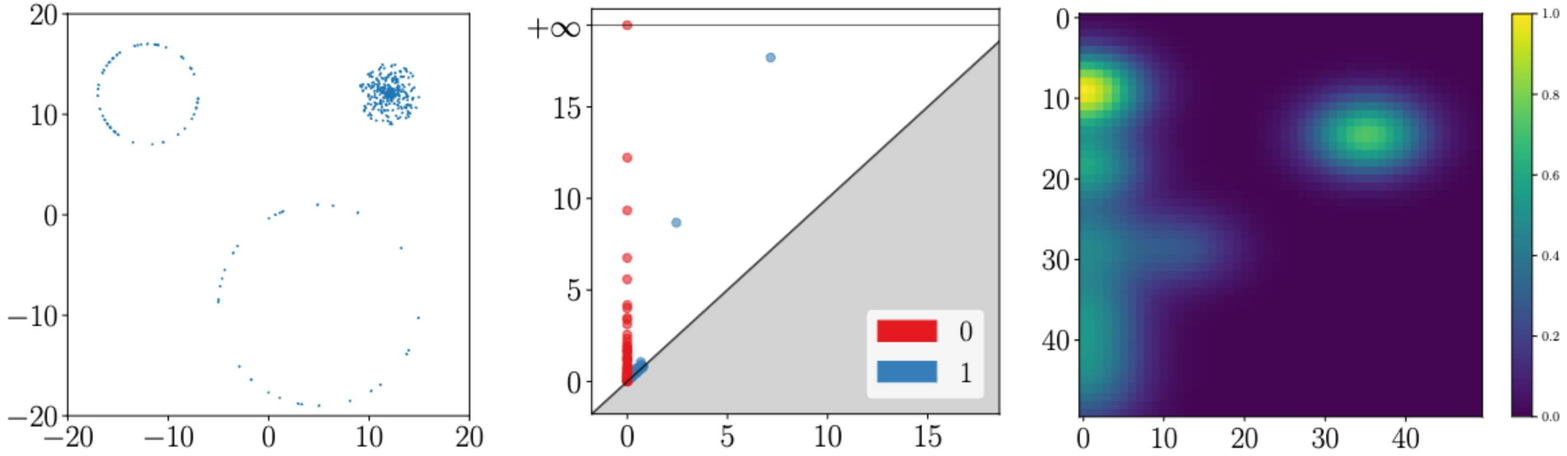



Kusano, G, & Fukumizu, K., Hiraoka., Y. Persistence weighted Gaussian kernel for topological data analysis. (2016)

# Background of persistent homology

## Persistence image

- Since a PD is defined as a multiset, it is not straightforward to feed it into a neural network.
- An alternative fixed-size representation of the persistence diagram reflecting the persistence and density of points in the PD.




(a) Point cloud data (b) Persistence diagram (c) Persistence image

# Method

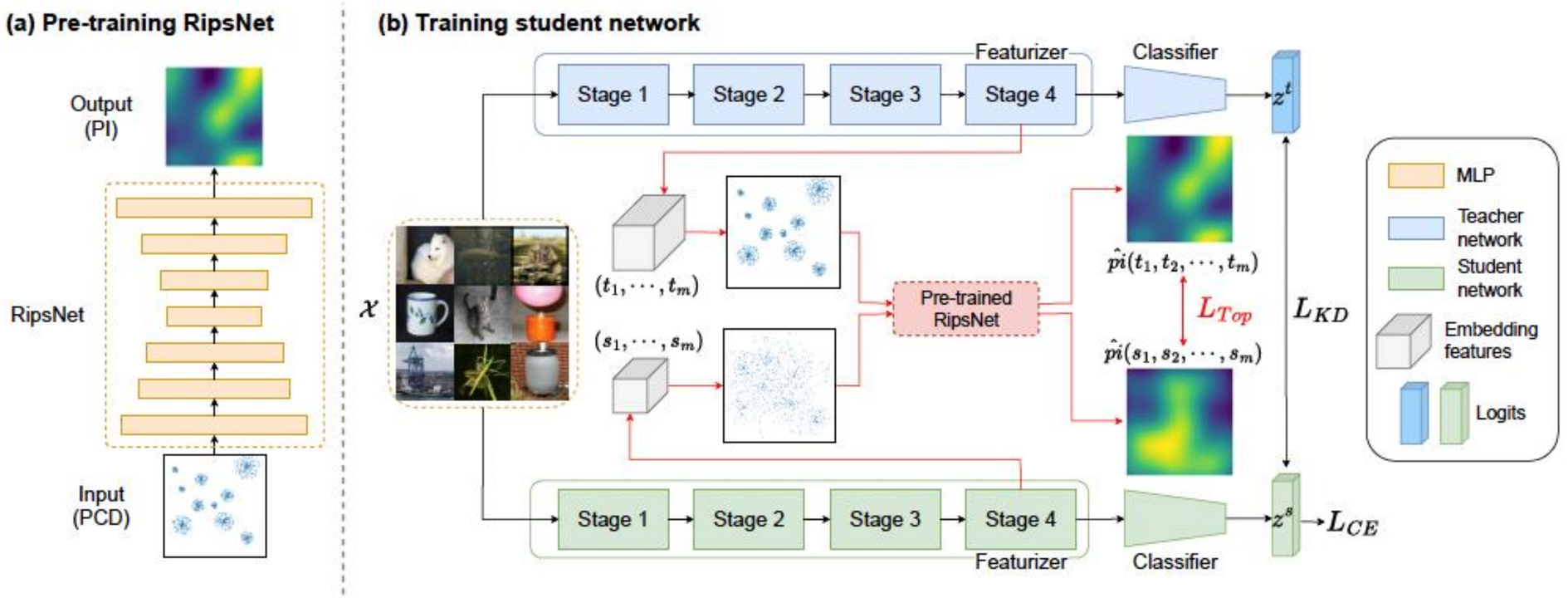
## Preliminaries

- *Notation :*
  - $T$ : pretrained teacher, $S$ : student
  - $x_i$ : training sample, $f^T(x_i)$, $f^S(x_i)$ : the output of any layer of the featurizer, $z^T(x_i), z^S(x_i)$ : logits.
  - $L_{CE} = \sum_{x_i \in X} CE(\sigma(z^S(x_i)), y_i)$ : the student model for image classification is trained by minimizing the cross-entropy loss.
- *Conventional KD :*
  - $L_{KD} = \sum_{x_i \in X} KL(\sigma(z^T(x_i)/\tau), \sigma(z^S(x_i)/\tau))$ : vanilla KD loss function.
  - Conventional KD is trained with the final loss $L = \alpha L_{CE} + \beta L_{KD}$.

# Method

## TopKD

- The proposed method, *TopKD*, consists of two stages, pre-training RipsNet and training student network.
- We defined the PD of the embedding features as the *global topology knowledge*, and the *topology distillation loss*.
- For the effective integration of PDs, we use the PI as a vectorization

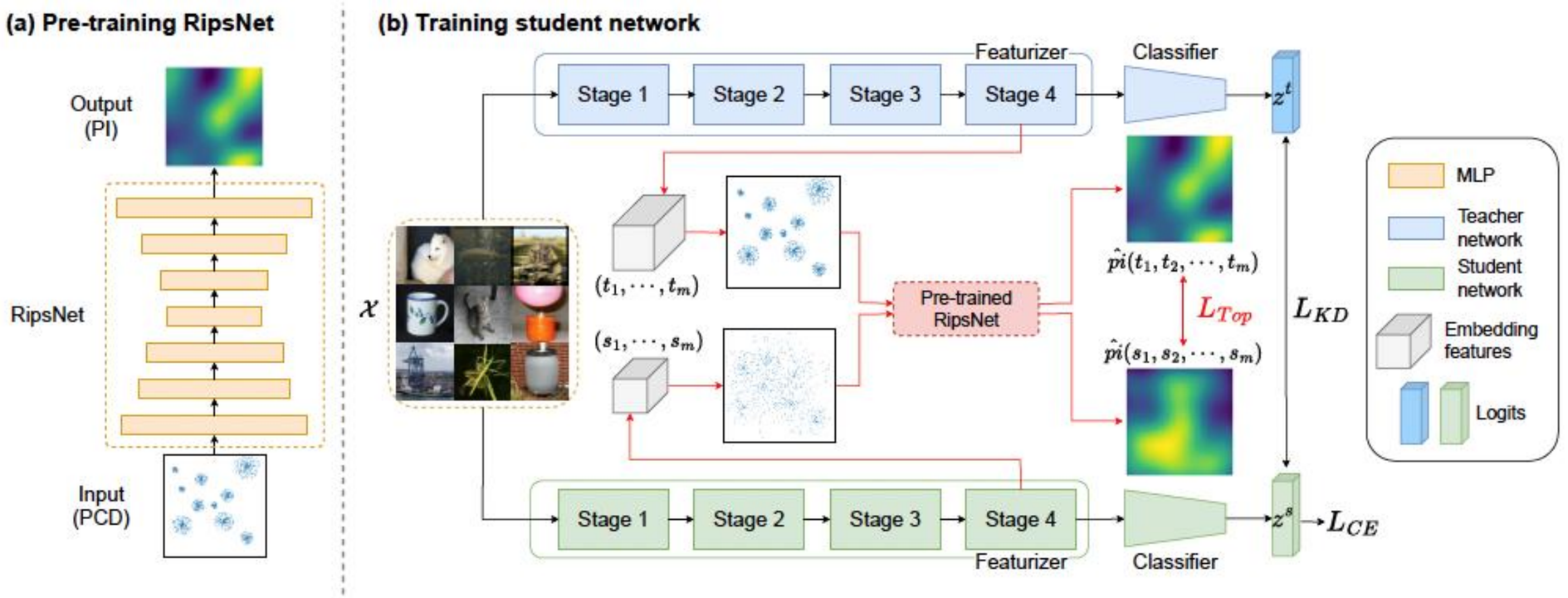


# Method

## Approximating PIs using RipsNet

1. Create PCDs from the training data using a pretrained teacher model.
2. Calculate PDs of PCDs using Gudhi library.
3. Train RipsNet by using PCDs and PDs as input and output, respectively

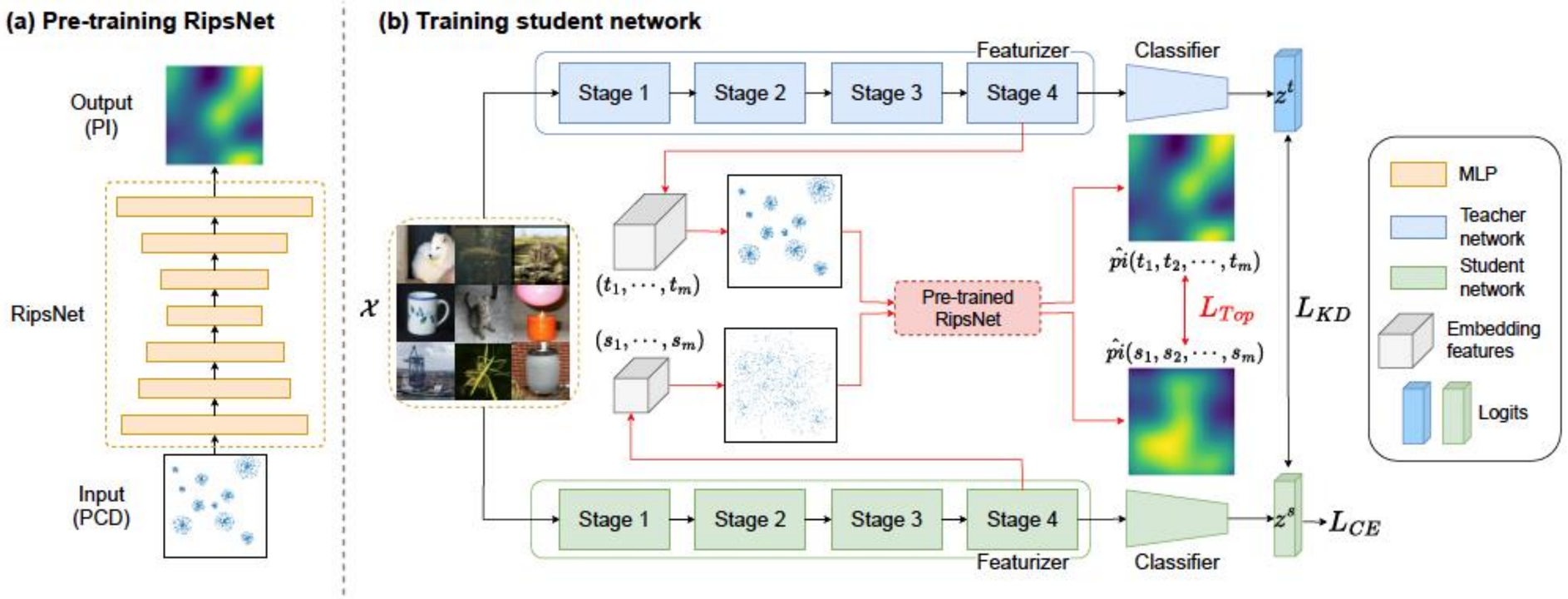
RipsNet is frozen during the training of the student network to approximate PIs of the embedding features of the teacher and student networks.



# Method

## Topology distillation loss

- $L_{Top} = \sum_{(x_1,\cdots,x_m)\in X^m} L_2(\hat{pi}(t_1,\cdots,t_m), \hat{pi}(s_1,\cdots,s_m)))$.
- If the channel dimension of $t_i$ and $s_i$ differ, a $1 \times 1$ convolution is applied to $s_i$.
- The final loss function of TopKD : $L_{Total} = \alpha L_{CE} + \beta L_{KD} + \gamma L_{Top}$.



# Experiment

## CIFAR-100

*Table 1.* Top-1 accuracy (%) comparison on CIFAR-100 with other KD approaches. Teacher and student networks have the same architectural style. Blue inverted triangles indicate lower performance than KD; red triangles signify better performance than KD. "-" indicates the absence of any available results. Relation denotes the relationships between a specific number of points (e.g., pairwise or triple-wise) to extract knowledge.

| Distillation Mechanism | Knowledge | Relation | Teacher<br>Acc. | ResNet56<br>72.34 | ResNet110<br>74.31 | ResNet110<br>74.31 | ResNet32×4<br>79.42 | WRN-40-2<br>75.61 | WRN-40-2<br>75.61 | VGG13<br>74.64 |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Student<br>Acc. | ResNet20<br>69.06 | ResNet20<br>69.06 | ResNet32<br>71.14 | ResNet8×4<br>72.50 | WRN-16-2<br>73.26 | WRN-40-1<br>71.98 | VGG8<br>70.36 |
| Logit | Soft logits | - | KD | 70.66 | 70.67 | 73.08 | 73.33 | 74.92 | 73.54 | 72.98 |
| Feature | Feature value | - | FitNet | 69.21 ▼ | 68.99 ▼ | 71.06 ▼ | 73.50 ▼ | 73.58 ▼ | 72.24 ▼ | 71.02 ▼ |
| | Attention map | - | AT | 70.55 ▼ | 70.22 ▼ | 72.31 ▼ | 73.44 ▲ | 74.08 ▼ | 72.77 ▼ | 71.43 ▼ |
| | Variational distribution | - | VID | 70.38 ▼ | 70.16 ▼ | 72.61 ▼ | 73.09 ▼ | 74.11 ▼ | 73.30 ▼ | 71.23 ▼ |
| | Preactivation feature | - | OFD | 70.98 ▲ | - | 73.23 ▲ | 74.95 ▲ | 75.24 ▲ | 74.33 ▲ | 73.95 ▲ |
| Relation | Correlation coefficient | Pair | CC | 69.63 ▼ | 69.48 ▼ | 71.48 ▼ | 72.97 ▼ | 73.56 ▲ | 72.21 ▼ | 70.71 ▼ |
| | Similarity matrix | Pair | SP | 69.67 ▼ | 70.04 ▼ | 72.69 ▲ | 72.94 ▼ | 73.83 ▼ | 72.43 ▼ | 72.68 ▼ |
| | Direction | Pair | FSP | 69.95 ▼ | 70.11 ▼ | 71.89 ▼ | 72.62 ▼ | 72.91 ▼ | - | 70.23 ▼ |
| | Distance&angle | Pair/triple | RKD | 69.61 ▼ | 69.25 ▼ | 71.82 ▼ | 71.90 ▼ | 73.35 ▼ | 72.22 ▼ | 71.48 ▼ |
| | Probability of features | Pair | PKT | 70.34 ▼ | 70.25 ▼ | 72.61 ▼ | 73.64 ▲ | 74.54 ▼ | 73.45 ▼ | 72.88 ▼ |
| | Contrastive learning | Pair | CRD | 71.16 ▲ | 71.46 ▲ | 73.48 ▲ | 75.51 ▲ | 75.48 ▲ | 74.14 ▲ | 73.94 ▲ |
| | Contrastive learning | Pair | CRCD | 73.21 ▲ | 72.33 ▲ | 74.98 ▲ | 76.42 ▲ | 76.67 ▲ | 75.95 ▲ | 74.97 ▲ |
| Topology | Global topology | All | Ours | 71.58 ▲ | 71.47 ▲ | 73.77 ▲ | 75.40 ▲ | 75.75 ▲ | 74.43 ▲ | 74.01 ▲ |

# Experiment

## CIFAR-100

*Table 2.* Top-1 accuracy (%) comparison on CIFAR-100 with other KD approaches. These teacher and student networks have different architectural styles.

| Distillation Mechanism | Knowledge | Relation | Teacher<br>Acc.<br>Student<br>Acc. | VGG13<br>74.64<br>MobileNetV2<br>64.60 | ResNet50<br>79.34<br>MobileNetV2<br>64.60 | ResNet50<br>79.34<br>VGG8<br>70.36 | ResNet32×4<br>79.42<br>ShuffleNetV1<br>70.50 | ResNet32×4<br>79.42<br>ShuffleNetV2<br>71.82 | WRN-40-2<br>75.61<br>ShuffleNetV1<br>70.50 |
|---|---|---|---|---|---|---|---|---|---|
| Logit | Soft logits | - | KD | 67.37 | 67.35 | 73.81 | 74.07 | 74.45 | 74.83 |
| Feature | Feature value | - | FitNet | 64.14 ▼ | 63.16 ▼ | 70.69 ▼ | 73.59 ▼ | 73.54 ▼ | 73.73 ▼ |
| | Attention map | - | AT | 59.40 ▼ | 58.58 ▼ | 71.84 ▼ | 71.73 ▼ | 72.73 ▼ | 73.32 ▼ |
| | Variational distribution | - | VID | 65.56 ▼ | 67.57 ▲ | 70.30 ▼ | 73.38 ▼ | 73.40 ▼ | 73.61 ▼ |
| | Preactivation feature | - | OFD | 69.48 ▲ | 69.04 ▲ | - | 75.98 ▲ | 76.82 ▲ | 75.85 ▲ |
| Relation | Correlation coefficient | Pair | CC | 64.86 ▼ | 65.43 ▼ | 70.25 ▼ | 71.14 ▼ | 71.29 ▼ | 71.38 ▼ |
| | Similarity matrix | Pair | SP | 66.30 ▼ | 68.08 ▲ | 73.34 ▼ | 73.48 ▼ | 74.56 ▲ | 74.52 ▼ |
| | Distance&angle | Pair/triple | RKD | 64.52 ▼ | 64.43 ▼ | 71.50 ▼ | 72.28 ▼ | 73.21 ▼ | 72.21 ▼ |
| | Probability of features | Pair | PKT | 67.13 ▼ | 66.52 ▼ | 73.01 ▼ | 74.10 ▼ | 74.69 ▲ | 73.89 ▼ |
| | Contrastive learning | Pair | CRD | 69.73 ▲ | 69.11 ▲ | 74.30 ▲ | 75.11 ▲ | 75.65 ▲ | 76.05 ▲ |
| Topology | Global topology | All | Ours | 68.83 ▲ | 69.12 ▲ | 74.25 ▲ | 75.04 ▲ | 76.33 ▲ | 76.18 ▲ |

# Experiment

## ImageNet-1K

*Table 3.* Top-1 and top-5 accuracy (%) (Acc.) comparison on the ImageNet-1K validation dataset with ResNet34 as the teacher and ResNet18 as the student network. The best accuracy values are bolded, and "-" indicates the absence of any available results.

| Acc. | Teacher | Student | AT | KD | SemCKD | OFD | CRD | CAT-KD | RKD | ReviewKD | DKD | SRRL | MGD | DistPro | NORM | Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Top-1 | 73.31 | 70.00 | 70.59 | 70.68 | 70.87 | 71.08 | 71.17 | 71.26 | 71.34 | 71.61 | 71.70 | 71.73 | 71.80 | 71.89 | 72.14 | **73.60** |
| Top-5 | 91.42 | 89.60 | 89.73 | 90.16 | - | - | 90.13 | 90.45 | 90.37 | **90.51** | 90.41 | - | 90.40 | - | - | 90.50 |

*Table 4.* Top-1 and top-5 accuracy (%) on the ImageNet-1K validation dataset with ResNet50 as the teacher and MobileNetV2 as the student network.

| Acc. | Teacher | Student | AT | KD | OFD | CRD | CAT-KD | RKD | ReviewKD | DKD | SRRL | MGD | DistPro | NORM | Ours |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Top-1 | 76.16 | 66.20 | 69.56 | 68.58 | 71.25 | 71.37 | 72.24 | 71.32 | 72.56 | 72.05 | 72.49 | 72.59 | 73.26 | 74.26 | **76.80** |
| Top-5 | 92.86 | 85.80 | 89.33 | 88.98 | 90.34 | 90.41 | 91.13 | - | 91.00 | 91.05 | - | 90.94 | - | - | **92.80** |

# Analysis

## Analysis regarding approximated PIs

How matching the approximated PIs for the embedding features of the teacher and student networks affects the actual distance between their exact PIs.

Triangle inequality : $||pi_T - pi_S||_2 \leq ||pi_T - \widehat{pi}_T||_2 + ||\widehat{pi}_T - \widehat{pi}_S||_2 + ||pi_S - \widehat{pi}_S||_2$.

- $||\widehat{pi}_T - \widehat{pi}_S||_2$ : topological distillation loss.
- $||pi_T - \widehat{pi}_T||_2, ||pi_S - \widehat{pi}_S||_2$ : the approximation errors for the teacher and student.
- Approximation capability is crucial for matching the exact PI of the student to that of the teacher network.
- However, $||pi_S - \widehat{pi}_S||_2$ is not directly minimized through the training process of the student.

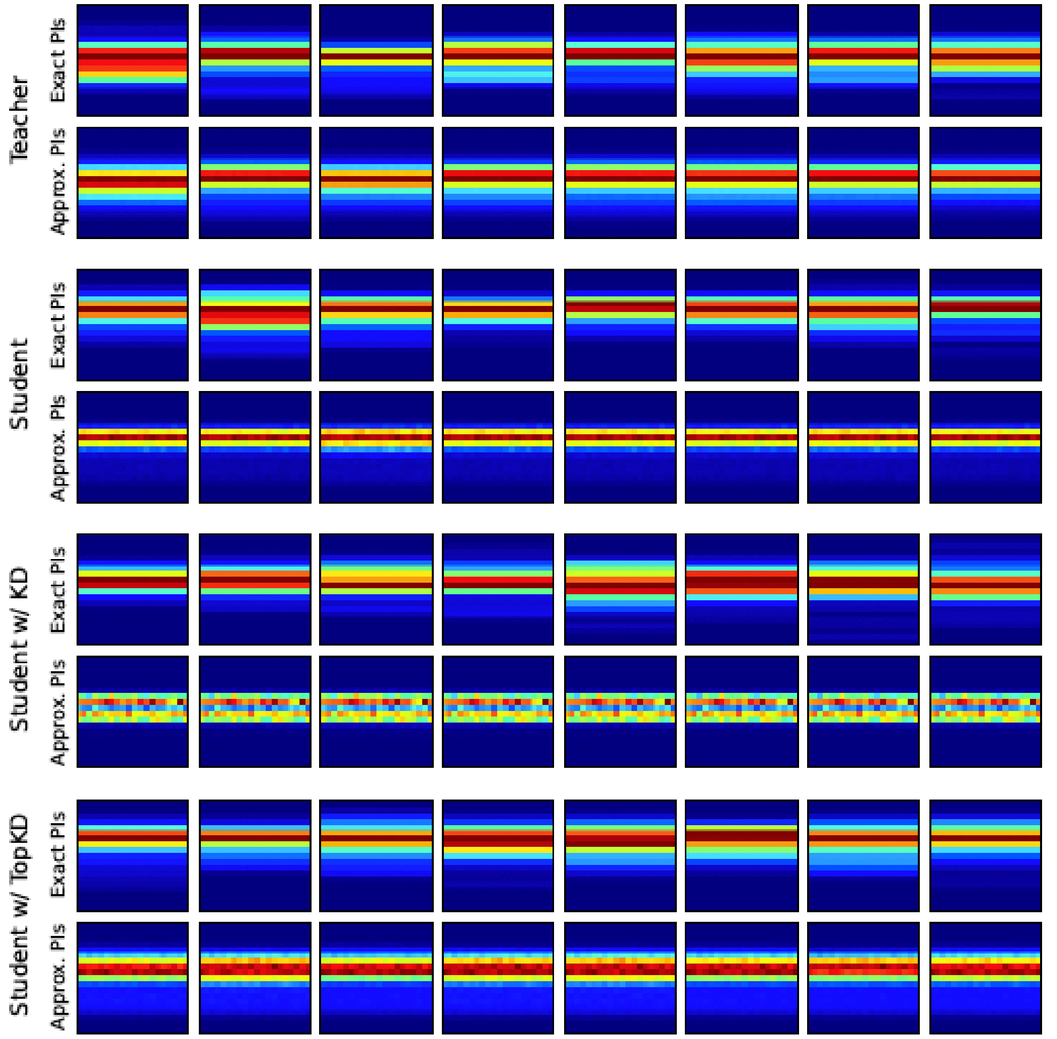
# Analysis

## Analysis regarding approximated PIs

- Error analysis.
  - Evaluate the approximation errors on the embedding features of the student and teacher networks.
  - TopKD has smaller errors compared to the students trained from scratch or with KD.

Table 8: Approximation errors on CIFAR-100. $\mathcal{L}_{RN}$ denotes the training error for the teacher as in Eq. (3). The values are averaged across minibatches of the training dataset. The bolded value indicates the smallest error.

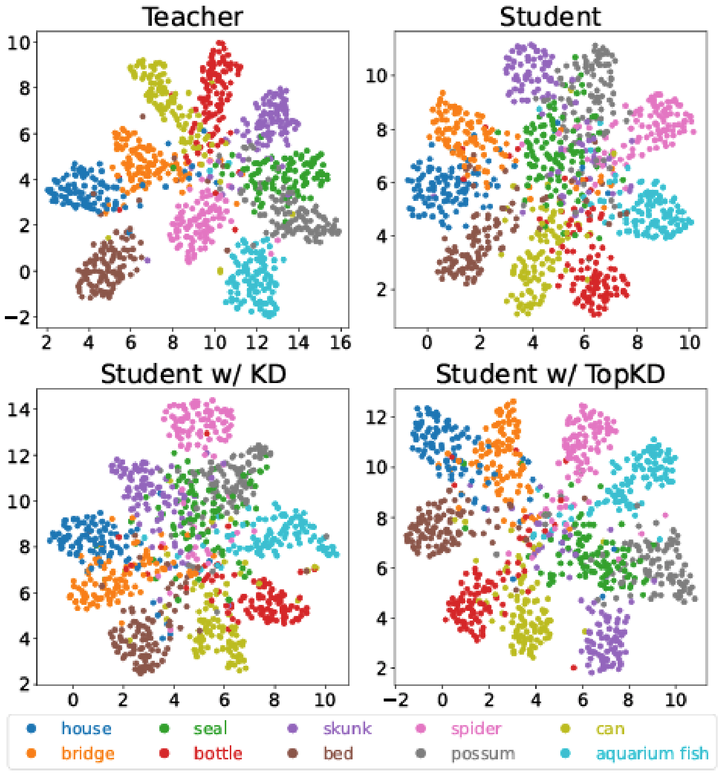
| Teacher | Student | $\mathcal{L}_{RN}$ ($\|pi_T - \hat{pi}_T\|_2$) | $\|pi_S - \hat{pi}_S\|_2$ Student | Student w/ KD | Student w/ TopKD |
|---|---|---|---|---|---|
| VGG13 | MobileNetV2 | 0.00229 | 0.02335 | 0.03408 | **0.02103** |
| ResNet50 | MobileNetV2 | 0.00218 | 0.00997 | **0.00734** | 0.00740 |
| ResNet50 | VGG8 | 0.00218 | 0.11178 | 0.01961 | **0.01679** |
| ResNet32×4 | ShuffleNetV1 | 0.00248 | 0.05301 | 0.05114 | **0.04084** |
| ResNet32×4 | ShuffleNetV2 | 0.00248 | 0.06427 | 0.07016 | **0.00434** |
| WRN-40-2 | ShuffleNetV1 | 0.00164 | 0.06591 | 0.05419 | **0.05001** |



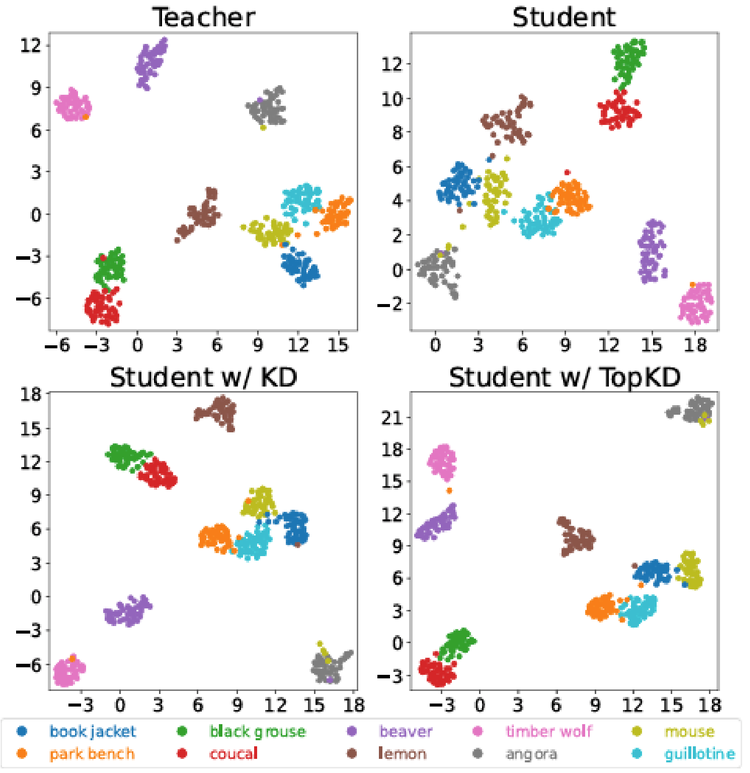
# Analysis

## Visualization of overall topology

- UMAP
  - TopKD more effectively gathers points by class than vanilla KD, making clearer distinctions between classes.




CIFAR-100




ImageNet-1K

# Thank you

Any questions

jekim5418@yonsei.ac.kr