南京航空航天大學
Nanjing University of Aeronautics and Astronautics

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

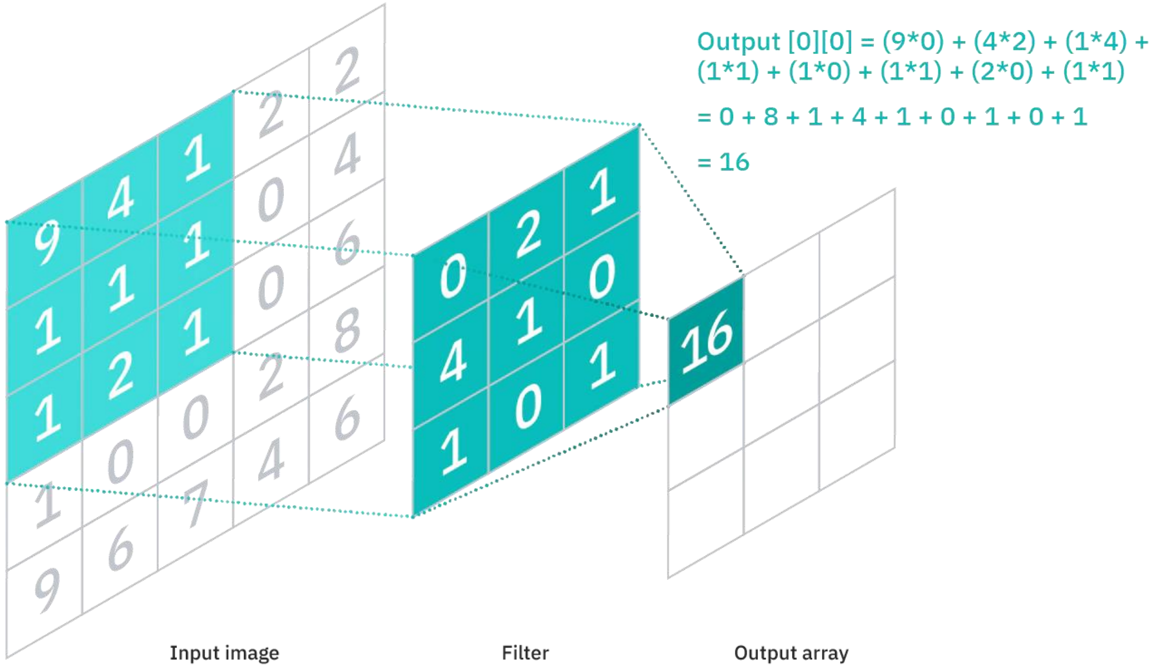
Distilling Holistic Knowledge with Graph Neural Networks

Sheng Zhou[1,2*], Yucheng Wang[1*], Defang Chen[1], Jiawei Chen[3], Xin Wang[4], Can Wang[1], Jiajun Bu[1†]

[1]Zhejiang Provincial Key Laboratory of Service Robot, Zhejiang University [2]School of Software Technology, Zhejiang University
[3]University of Science and Technology of China [4]Tsinghua University

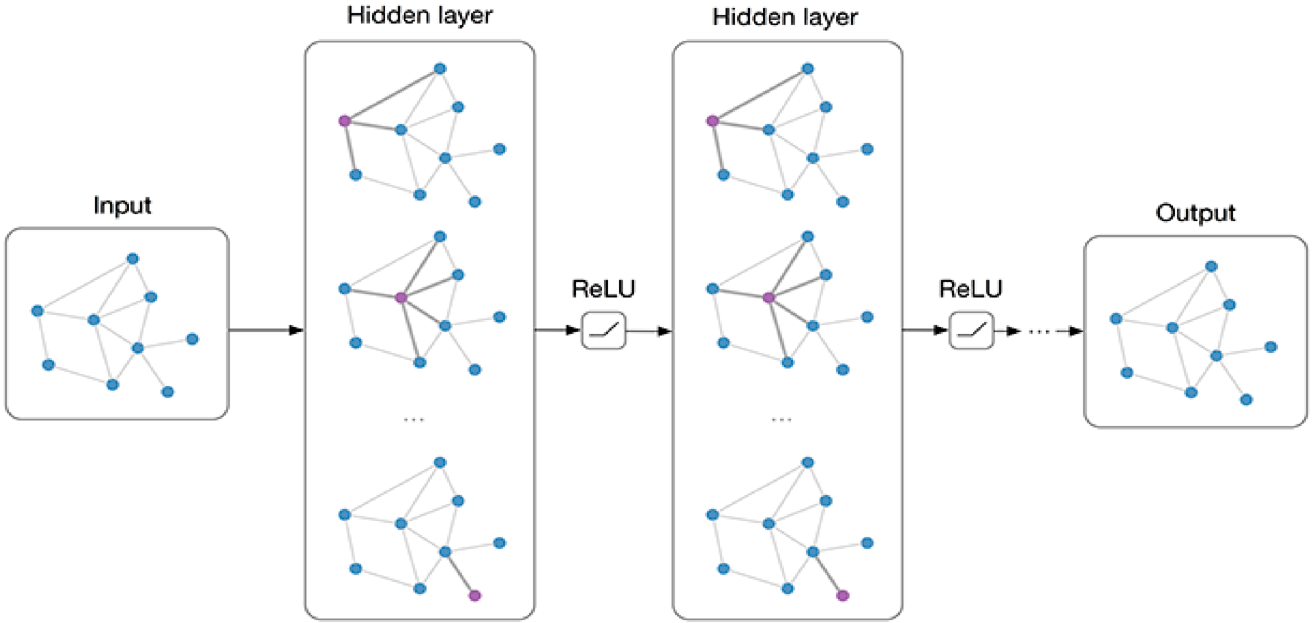
ICCV 2021

Graph Neural Networks



CNN

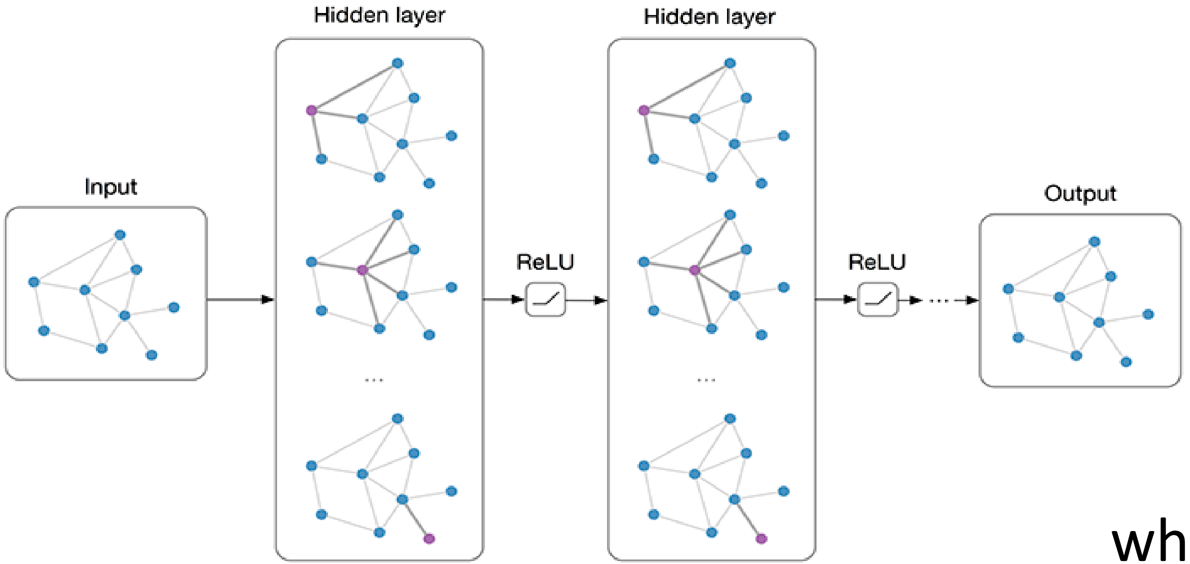
GCN



Graph Neural Networks

For these models, the goal is to learn a function of signals/features on a graph $G = (\mathcal{V}, \mathcal{E})$ which takes as input:

1. A feature description x_i for every node i; summarized in a $N \times K$ feature matrix X (N: number of nodes, K: number of input features)
2. A representative description of the graph structure in matrix form; typically in the form of an adjacency matrix A (or some function thereof)



$$f(H^{(l)}, A) = \sigma\left(AH^{(l)}W^{(l)}\right)$$

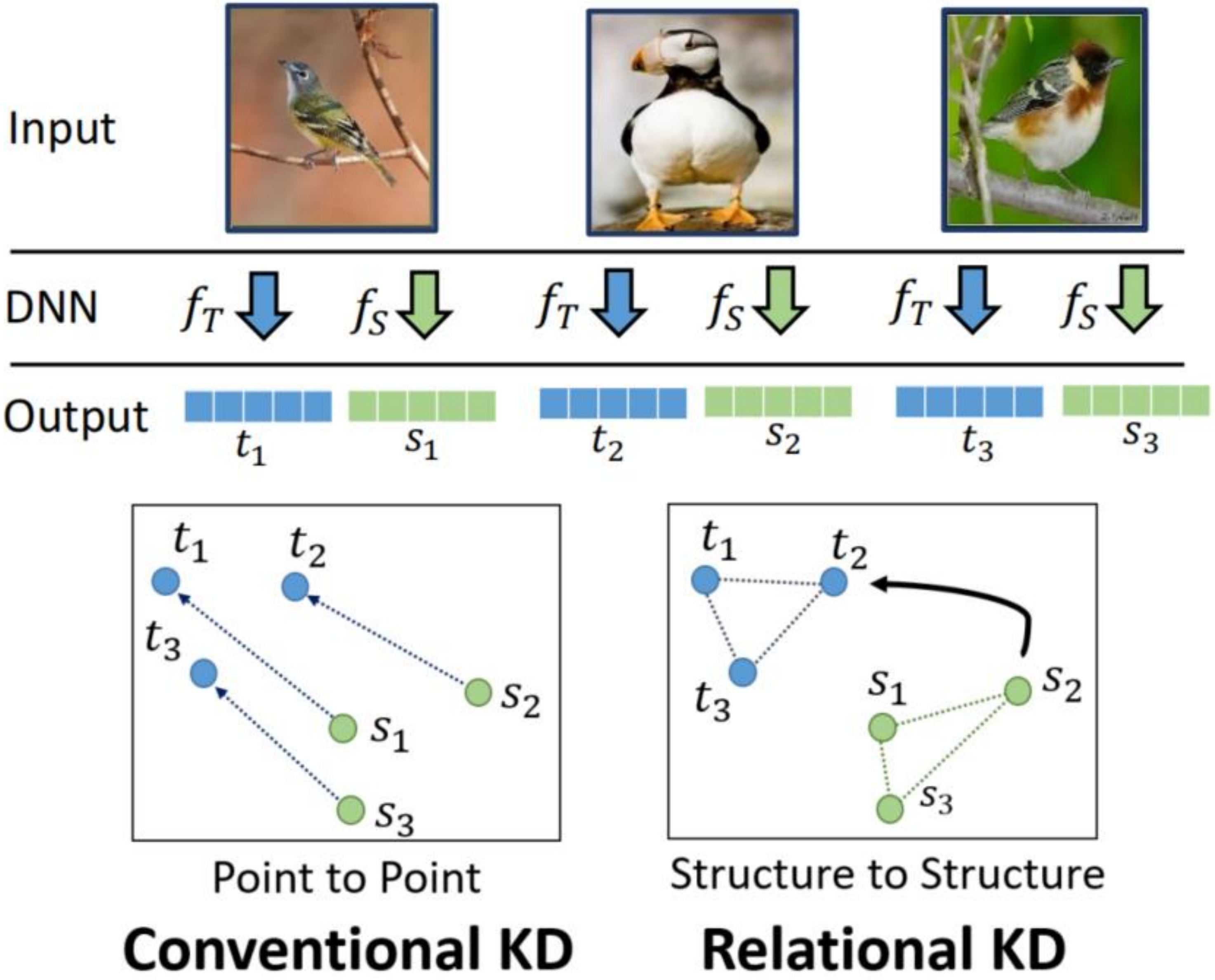
$$\Downarrow$$

$$f(H^{(l)}, A) = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$$

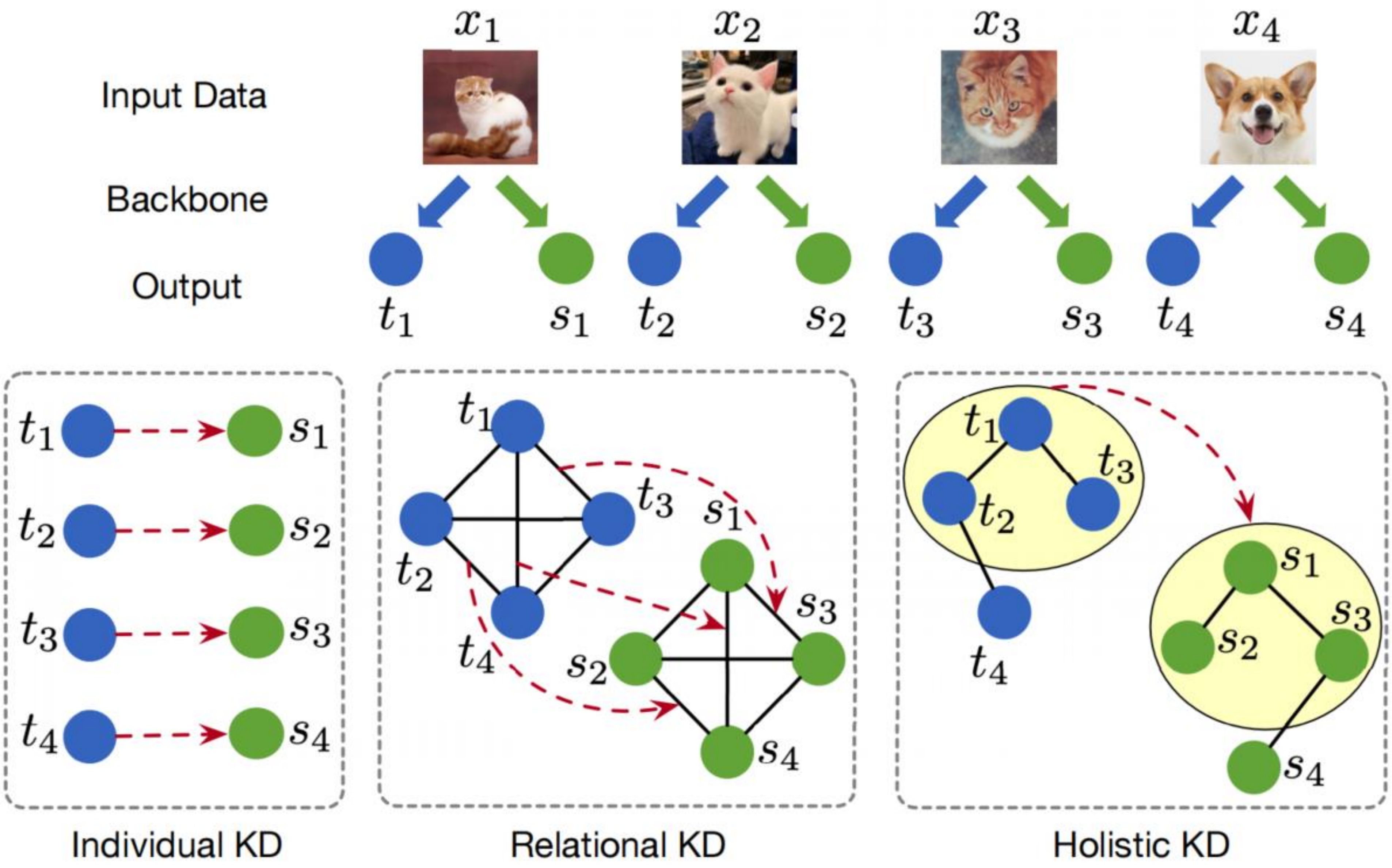
where $A\wedge = A + I$,
$D\wedge$ is the diagonal node degree matrix of $A\wedge$

https://tkipf.github.io/graph-convolutional-networks/

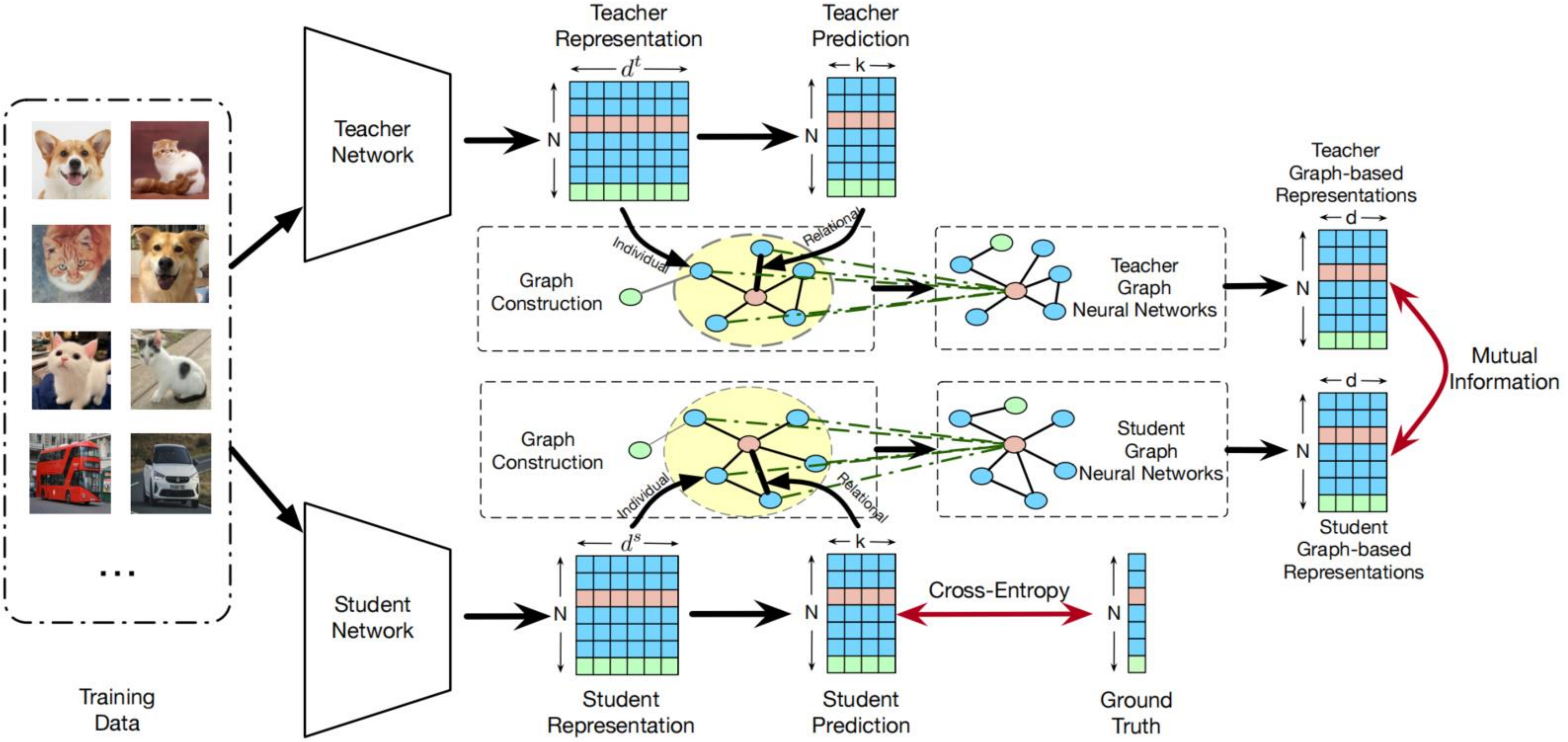
Relational Knowledge Distillation



Distilling Holistic Knowledge

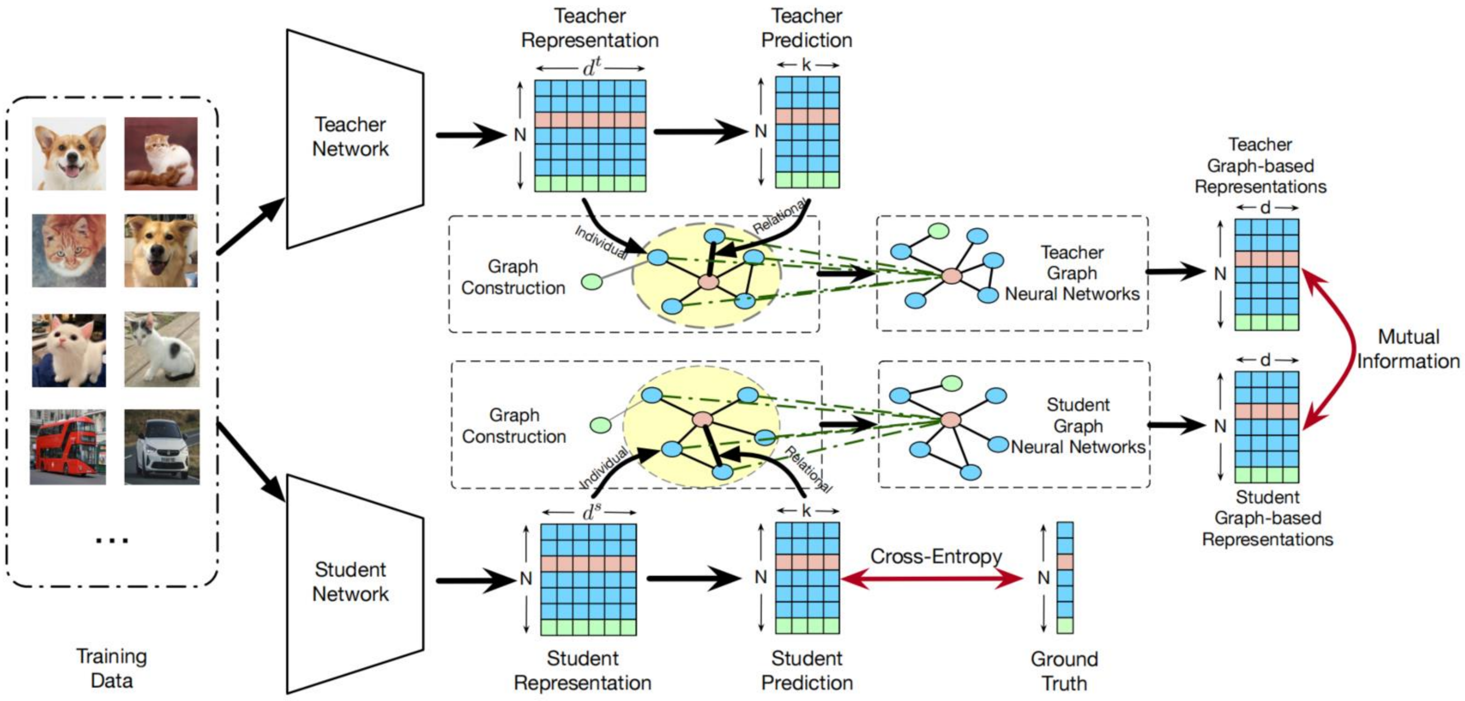


Distilling Holistic Knowledge



Graph Convolution Network

$$\mathbf{F}^t \in \mathbb{R}^{N \times d^t} \qquad \mathbf{A}^t = \phi(\mathbf{p}^t)$$



$$\mathbf{F}^s \in \mathbb{R}^{N \times d^s} \qquad \mathbf{A}^s = \phi(\mathbf{p}^s)$$

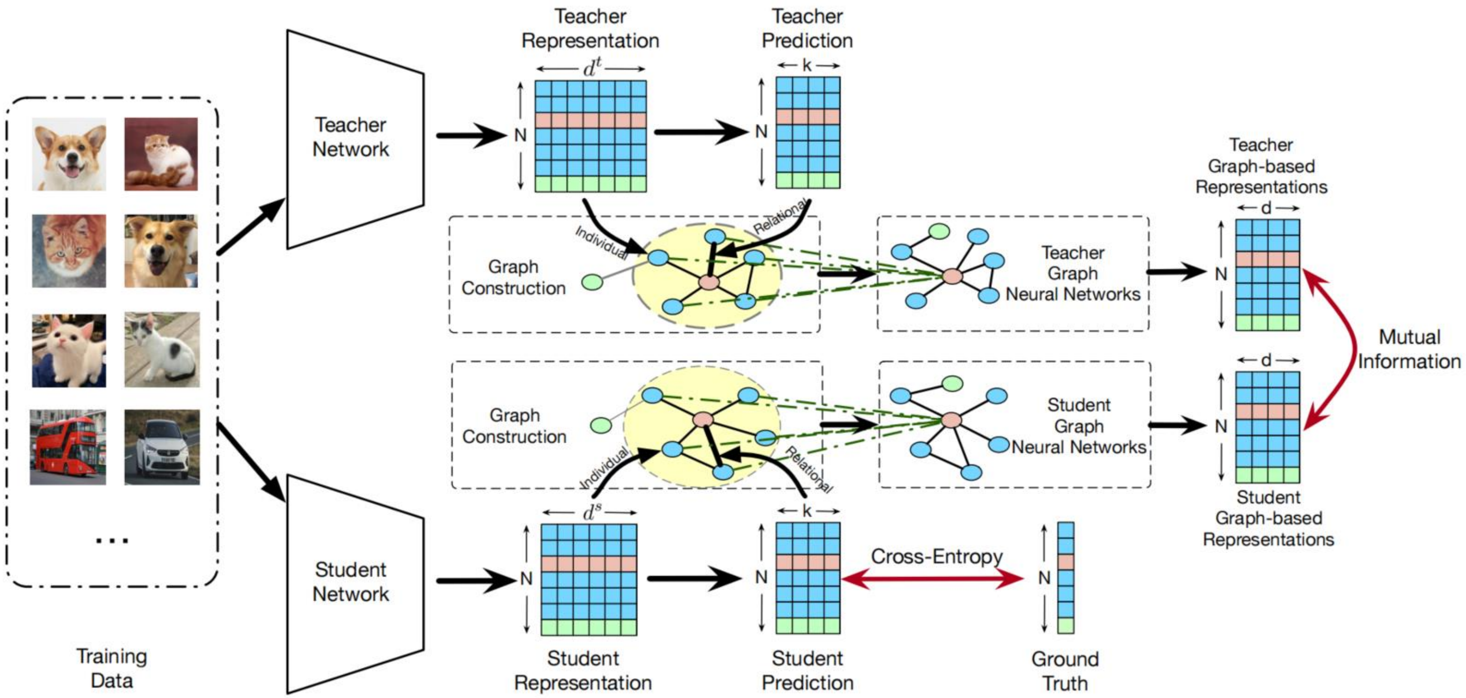
$\boldsymbol{D}$: diagonal degree matrix

$$\mathbf{H}^t = \sum_{l=0}^{L} \left(\mathbf{D}_t^{-1/2} \mathbf{A}^t \mathbf{D}_t^{-1/2} \right)^l \mathbf{F}^t \boldsymbol{\Theta}_l^t$$

$$\mathbf{H}^s = \sum_{l=0}^{L} \left(\mathbf{D}_s^{-1/2} \mathbf{A}^s \mathbf{D}_s^{-1/2} \right)^l \mathbf{F}^s \boldsymbol{\Theta}_l^s$$

$\phi(\cdot)$ is the KNN-based graph construction function

InfoNCE Estimator



$$\mathbf{I}(\mathbf{H}^t, \mathbf{H}^s) \geq \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} log \frac{e^{f(\mathbf{h}_i^t, \mathbf{h}_i^s)}}{\frac{1}{N}\sum_{j=1}^{N} e^{f(\mathbf{h}_i^t, \mathbf{h}_j^s)}}\right]$$

$$\widetilde{\mathcal{L}}_{HOL} = \sum_{i=1}^{N} log \frac{e^{f(\mathbf{h}_i^t, \mathbf{h}_i^s)}}{e^{f(\mathbf{h}_i^t, \mathbf{h}_i^s)} + \sum_{j=1, j\neq i}^{N} e^{f(\mathbf{h}_i^t, \mathbf{f}_j^s)}} + log \frac{e^{f(\mathbf{h}_i^s, \mathbf{h}_i^t)}}{e^{f(\mathbf{h}_i^s, \mathbf{h}_i^t)} + \sum_{j=1, j\neq i}^{N} e^{f(\mathbf{h}_i^s, \mathbf{f}_j^t)}}$$

The overall framework of the HKD method

Algorithm 1 Holistic Knowledge Distillation.

Input: Training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$; A pre-trained teacher model with parameter $\mathbf{W}^t$; A student model with random initialized parameters $\mathbf{W}^s$;

Output: A well-trained student model;

1: **while** $\mathbf{W}^s$ is not converged **do**

2: Sample a mini-batch $\mathcal{B}$ with size b from $\mathcal{D}$.

3: Forward propagation $\mathcal{B}$ into $\mathbf{W}^t$ and $\mathbf{W}^s$ to obtain feature representation $\mathbf{f}^t, \mathbf{f}^s$ and prediction $\mathbf{p}^t, \mathbf{p}^s$.

4: Construct attributed context graph $\mathbf{G}^t$ and $\mathbf{G}^s$.

5: Extract holistic knowledge with graph neural networks by Equation (5),(6).

6: Calculate the Mutual information between graph-based representation as Equation (10).

7: Update parameters $\mathbf{W}^s$ by backward propagation the gradients of the loss in Equation (9).

8: **end while**

Experiments

Table 1. Test accuracy (%) of the student networks on the CIFAR100 dataset of combining distillation methods with KD.

Teacher Student	ResNet32×4 ResNet8×4	ResNet32×4 ShuffleNetV2	VGG13 MobileNetV2	ResNet50 VGG8	ResNet50 MobileNetV2	ARI (%)
Teacher	79.42	79.42	74.64	79.34	79.34	/
Student	72.79 ± 0.26	72.63 ± 0.71	65.33 ± 0.63	70.56 ± 0.32	65.33 ± 0.63	
KD	73.55 ± 0.20	75.38 ± 0.52	68.08 ± 0.24	73.76 ± 0.09	67.83 ± 0.46	126.48 %
AT+KD	74.80 ± 0.15	76.51 ± 0.16	66.37 ± 0.13	73.91 ± 0.24	66.81 ± 0.11	152.84 %
PKT+KD	74.68 ± 0.07	76.16 ± 0.16	68.08 ± 0.94	74.19 ± 0.27	68.42 ± 0.39	55.63 %
SP+KD	73.99 ± 0.05	76.02 ± 0.34	68.46 ± 0.37	73.50 ± 0.20	68.18 ± 0.57	80.89 %
CC+KD	74.44 ± 0.14	75.81 ± 0.20	68.54 ± 0.21	73.48 ± 0.16	68.92 ± 0.16	58.96 %
RKD+KD	74.18 ± 0.09	75.64 ± 0.24	68.24 ± 0.46	73.81 ± 0.11	68.52 ± 0.14	72.15 %
CRD+KD	75.64 ± 0.25	76.41 ± 0.36	69.82 ± 0.22	74.41 ± 0.31	69.86 ± 0.04	15.32 %
SSKD+KD	75.80 ± 0.58	76.36 ± 0.38	69.12 ± 0.54	74.68 ± 0.22	69.53 ± 0.43	18.86 %
HKD	75.63 ± 0.22	76.31 ± 0.30	69.97 ± 0.42	74.86 ± 0.17	69.83 ± 0.15	12.94 %
HKD+KD	$\mathbf{76.13 \pm 0.05}$	$\mathbf{76.92 \pm 0.22}$	$\mathbf{70.48 \pm 0.25}$	$\mathbf{74.88 \pm 0.30}$	$\mathbf{70.72 \pm 0.32}$	/

Experiments

Table 2. Test accuracy (%) of the student networks on the TinyImageNet dataset of combining distillation methods with KD.

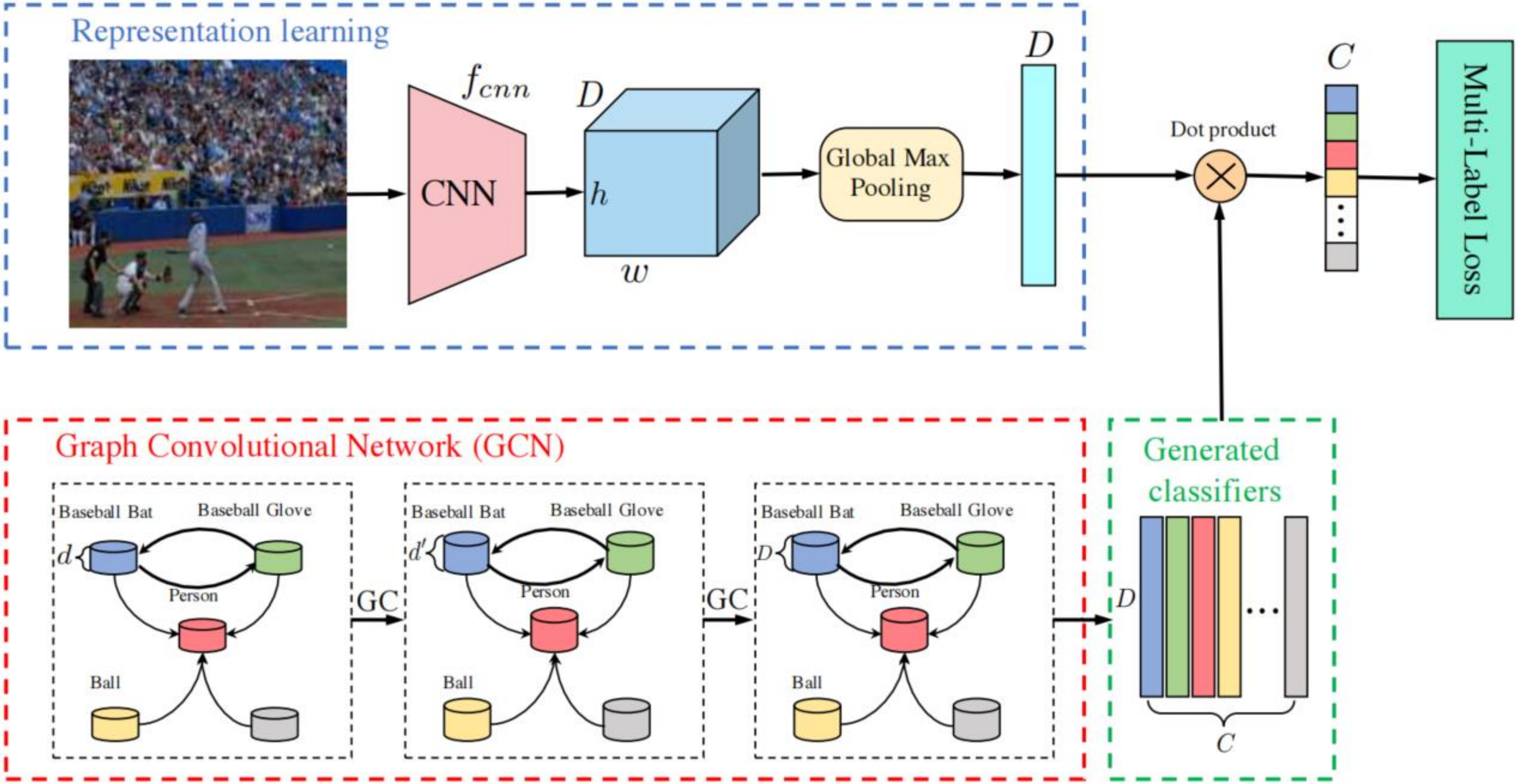
Teacher Student	ResNet32×4 ResNet8×4	ResNet32×4 ShuffleNetV2	VGG13 MobileNetV2	ResNet50 VGG8	VGG13 VGG8	ARI (%)
Teacher Student	57.92 49.91 ± 0.16	57.92 50.60 ± 0.23	52.02 44.20 ± 0.22	55.44 47.00 ± 0.17	52.02 47.00 ± 0.17	/
KD	52.28 ± 0.07	57.27 ± 0.03	45.39 ± 0.59	51.50 ± 0.36	51.34 ± 0.08	123.18 %
AT+KD	54.79 ± 0.23	57.56 ± 0.38	45.13 ± 0.60	51.42 ± 0.42	51.03 ± 0.28	122.61 %
PKT+KD	54.11 ± 0.18	58.33 ± 0.36	47.73 ± 0.31	51.45 ± 0.28	51.61 ± 0.28	35.51 %
SP+KD	54.22 ± 0.41	58.66 ± 0.25	48.10 ± 0.59	51.70 ± 0.12	51.51 ± 0.32	29.98 %
CC+KD	54.08 ± 0.32	58.20 ± 0.06	47.67 ± 1.14	50.87 ± 0.20	51.07 ± 0.33	44.12 %
RKD+KD	53.78 ± 0.15	57.85 ± 0.24	48.10 ± 0.26	51.01 ± 0.23	50.59 ± 0.32	46.70 %
CRD+KD	55.53 ± 0.41	58.95 ± 0.05	49.12 ± 0.04	52.87 ± 0.30	52.25 ± 0.26	7.88 %
SSKD+KD	55.10 ± 2.05	57.48 ± 0.04	47.02 ± 0.90	52.36 ± 0.36	51.60 ± 0.16	35.51 %
HKD	55.53 ± 0.07	58.83 ± 0.09	49.53 ± 0.32	52.20 ± 0.20	51.97 ± 0.33	10.48 %
HKD+KD	$\mathbf{56.18 \pm 0.12}$	$\mathbf{59.31 \pm 0.01}$	$\mathbf{49.57 \pm 0.54}$	$\mathbf{53.30 \pm 0.33}$	$\mathbf{52.62 \pm 0.03}$	/

Experiments

Table 4. Representation transferability experiments of the student network. The student network is trained on the CIFAR100 dataset and transferred to the TinyImageNet and the STL10 dataset. A linear classifier is evaluated on the frozen representations of the student network.

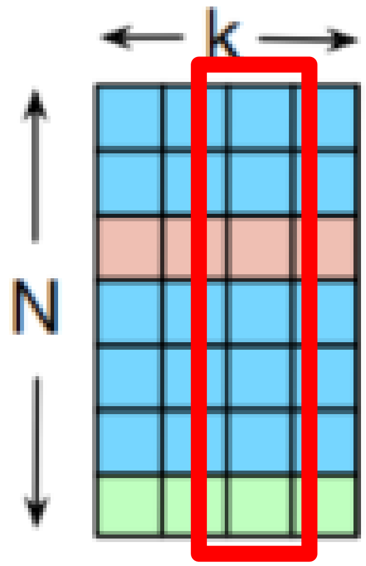
Dataset	TinyImageNet	STL-10
T:ResNet50	30.79 ± 0.01	70.16 ± 0.07
S:MobileNetV2	23.01 ± 0.05	61.42 ± 0.10
KD	22.92 ± 0.13	61.25 ± 0.09
AT+KD	25.02 ± 0.01	62.05 ± 0.06
PKT+KD	26.04 ± 0.11	63.71 ± 0.05
SP+KD	24.98 ± 0.08	62.25 ± 0.13
CC+KD	25.68 ± 0.03	62.52 ± 0.10
RKD + KD	26.10 ± 0.03	63.26 ± 0.03
CRD + KD	28.98 ± 0.05	65.87 ± 0.10
SSKD + KD	24.24 ± 0.02	61.78 ± 0.02
HKD + KD	$\mathbf{30.55 \pm 0.03}$	$\mathbf{67.28 \pm 0.08}$

ML-GCN


Representation learning
f_{cnn}
CNN
D
h
w
Global Max Pooling
D
Dot product
C
Multi-Label Loss
Graph Convolutional Network (GCN)
Baseball Bat
Baseball Glove
Person
Ball
d
GC
d'
GC
D
Generated classifiers
D
C

Multi-Label + Knowledge Distillation

We have tried these in the past few weeks:



1. Classwise Relational Knowledge Distillation
   → it works, but not enough

2. Correlation Matrix KL divergence/MSE

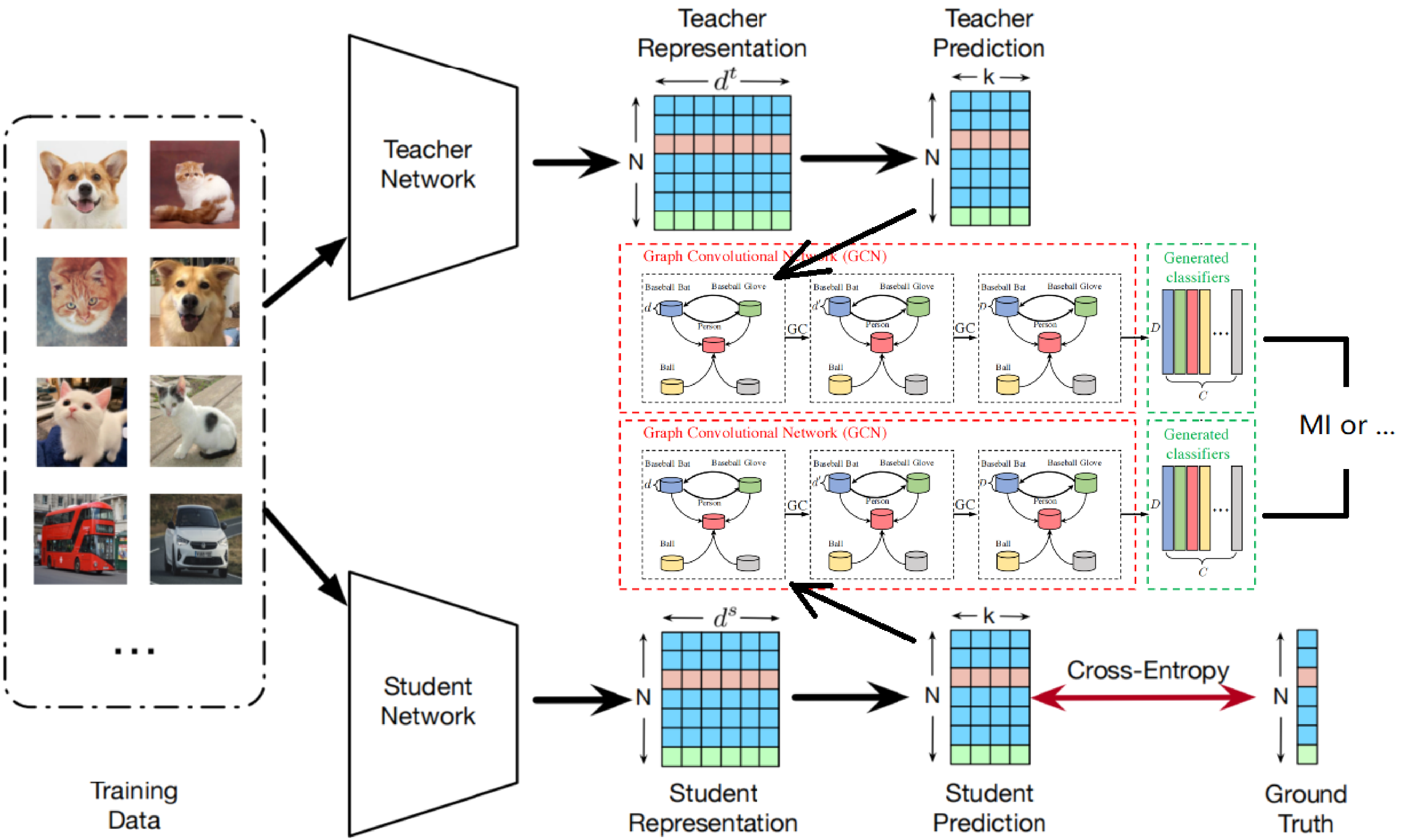
   "We model the label correlation dependency in the form of conditional probability, i.e., $P(L_j|L_i)$ which denotes the probability of occurrence of label L_j when label L_i appears. As shown in Fig.3, $P(L_j|L_i) \neq P(L_i|L_j)$. Thus, the correlation matrix is asymmetrical."

3. Cosine Similarity Weighted Distance

$$[P]_{k\times k} = \frac{t_i^\top t_j}{||t_i||\,||t_j||}$$

$$l = \sum_{i,j} P_{ij}\Delta(s_i, s_j)$$

Multi-Label KD + GNN?



Thanks for Listening