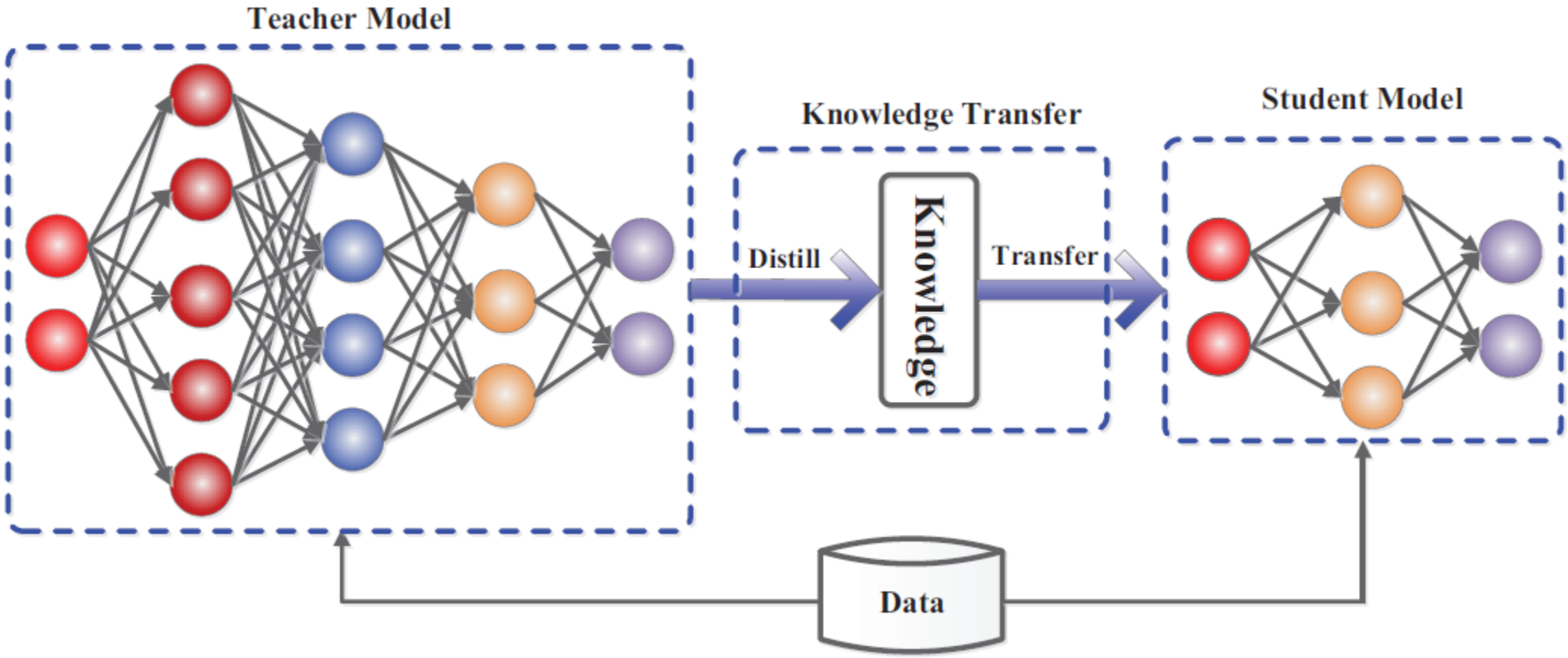
南京航空航天大學

Nanjing University of Aeronautics and Astronautics

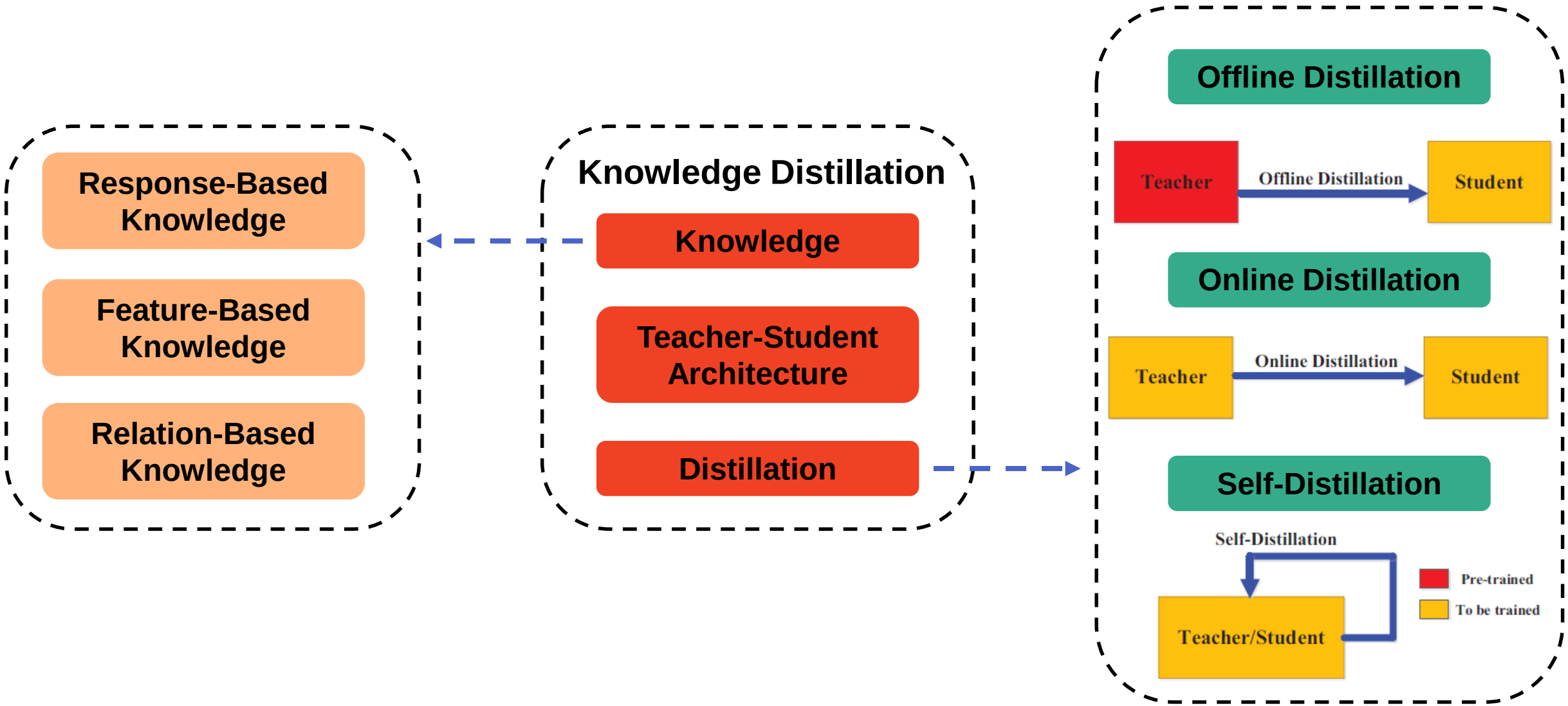
Two Articles about Knowledge Distillation and Active Learning

Knowledge Distillation

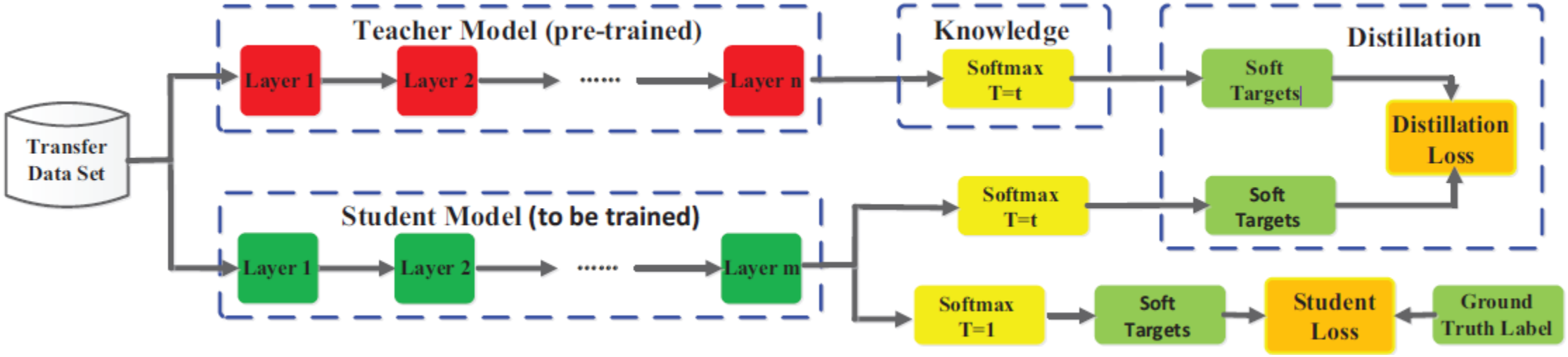


Goal: transfer knowledge from a large model to a small model for model compression and acceleration.

Knowledge Distillation



Knowledge Distillation

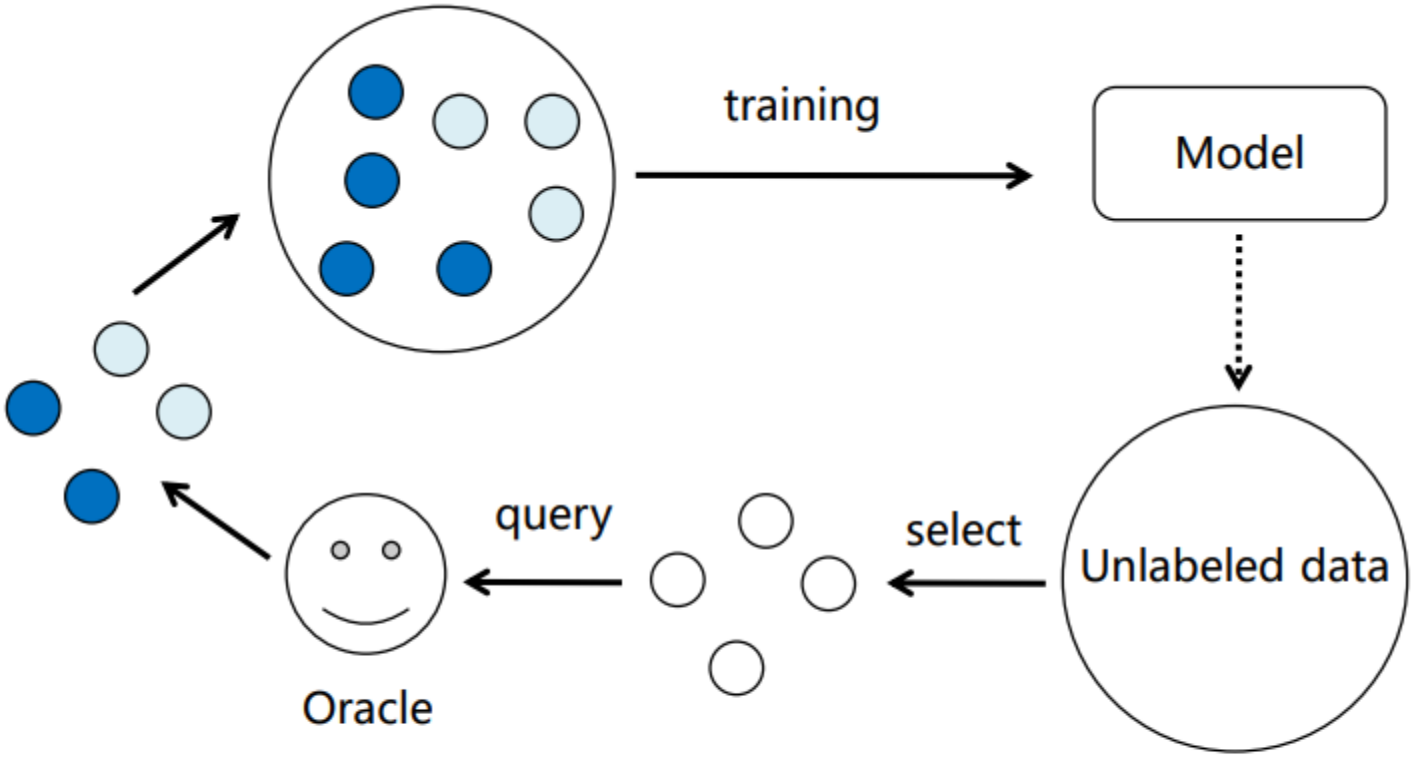


Soft targets are the probabilities that the input belongs to the classes and can be estimated by a softmax function as:

$$p(z_i, T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \ ,$$

where z_i is the logit for the i-th class, and a temperature factor T is introduced to control the importance of each soft target.

Active Learning



Goal: query less for more.

南京航空航天大學

Nanjing University of Aeronautics and Astronautics

Active Learning for Lane Detection: A Knowledge Distillation Approach

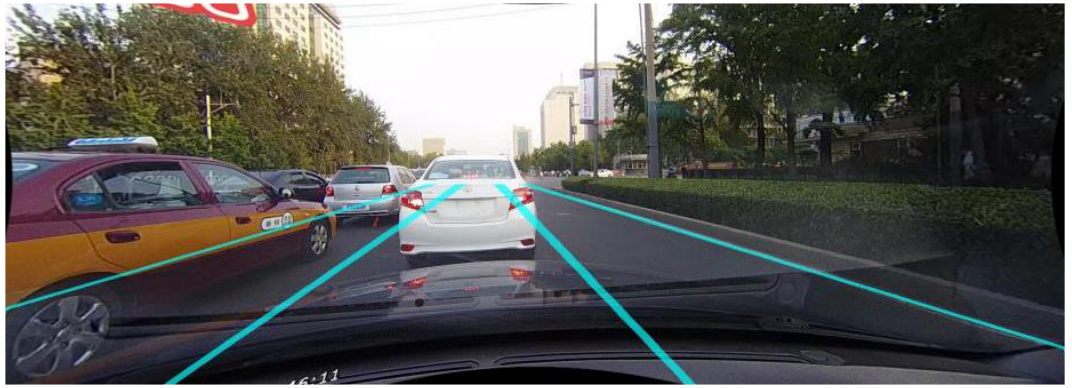
Fengchao Peng, Chao Wang, Jianzhuang Liu, Zhen Yang
Noah's Ark Lab, Huawei Technologies
{pengfengchao, wangchao165, liu.jianzhuang, yang.zhen}@huawei.com

ICCV 2021

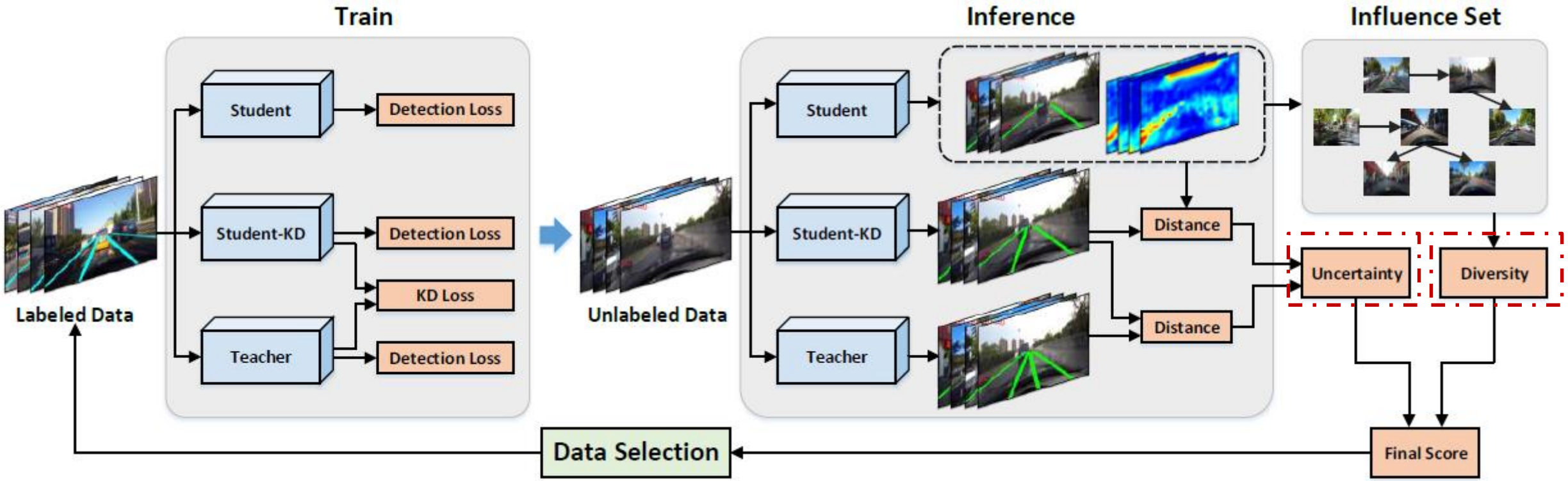
Motivation

Lane detection is a key task for autonomous driving vehicles.

- ✓ Rely on a huge amount of annotated images.
- ✓ Existing active learning methods perform poorly for lane detection.
  - Entropy-based active learning encourage to select images with very few lanes or even no lane at all.
  - Existing methods are not aware of the noise of lane annotations, which is caused by heavy occlusion and unclear lane marks.



Method


Train
Inference
Influence Set
Student
Detection Loss
Student-KD
Detection Loss
KD Loss
Teacher
Detection Loss
Labeled Data
Unlabeled Data
Student
Student-KD
Teacher
Distance
Distance
Uncertainty
Diversity
Data Selection
Final Score

Method-Uncertainty

For unsuitable entropy metric:

✓ Using prediction gap as the basic estimation of uncertainty.

The prediction gap between two models:

- Given an image p and two models $\mathrm{M}_1$and M_2, denote the sets of their predicted lanes as $\mathrm{M}_1(\mathrm{p})$ and $\mathrm{M}_2(\mathrm{p})$, respectively. For each lane $\mathrm{l}_1 \in \mathrm{M}_1(\mathrm{p})$, we find its closest lane in $\mathrm{M}_2(\mathrm{p})$ with:

$$l_2 = \arg\min_{l \in M_2(p)} Dist(l_1, l).$$

- The distance $\mathrm{Dist}()$ between two lanes is calculated as the segment-wise Euclidean distance. Then the prediction gap between M_1 and M_2 is defined as:

$$D_{12}(p) = \max_{l_1 \in M_1(p)} Dist(l_1, l_2).$$

Method-Uncertainty

For label noise:

- ✓ Useful knowledge can be transferred from the teacher to the student, but label noise is difficult to transfer. (a large prediction gap on label noise images)
- ✓ However, a large prediction gap does not necessarily indicate high label noise. There can be knowledge, i.e., label with no noise, which is naturally difficult for the student to learn. (train another student model without teacher)

Method-Uncertainty

We now have three trained models, the student M_S, the distilled student M_{S-KD}, and the teacher M_T. To model the uncertainty, we calculate the gap D_{SS} between M_S and M_{S-KD}, and the gap D_{ST} between M_{S-KD} and M_T.

- Small D_{SS} and small D_{ST} ($\times$)
- Small D_{SS} and large D_{ST} ($\surd$)
- Large D_{SS} and large D_{ST} ($\times$)
- Large D_{SS} and small D_{ST} ($\surd$)

Combining the above four cases, we propose a simple yet effective uncertainty metric for image p:

$$Uncr(p) = (D_{SS} + D_{ST}) \cdot \max\{\frac{D_{ST}}{D_{SS}}, \frac{D_{SS}}{D_{ST}}\}.$$

Method-Diversity

Reverse nearest neighbors(RNNs):

- Given a sample p, a dataset S, a distance function $d()$, and an integer k, the reverse k nearest neighbors of p is defined as:

$$RNN_k(p) = \{q \in S - \{p\} | p \in NN_k(q)\},$$

- where $NN_k(q)$ denotes the k nearest neighbors of q.

- Given the unlabeled dataset S_U, the current subset of selected samples $V \subset S_U$, and an image $p \in S_U$, we define the diversity of p as the number of its reverse k nearest neighbors in S_U:

$$Div(p|V, S_U) = |RNN_k(p) - V|.$$

Method

Algorithm 1 Active Learning with Knowledge Distillation

Input: Labeled dataset S_L, unlabeled dataset S_U, number of rounds r, budget b per round;
Output: Selected dataset $V \subset S_U$, with annotations;

$M_S, M_T \leftarrow Train(S_L)$;
$M_{ST} \leftarrow Train_{KD}(S_L)$;
$V \leftarrow \emptyset$;
while $|V| < r \cdot b$ **do**
  for $p \in S_U$ **do**
    $P_S, P_{ST}, P_T \leftarrow Predict(M_S, M_{ST}, M_T, p)$;
    Compute D_{SS} and D_{ST} with P_S, P_{ST}, P_T;
    $uncr \leftarrow (D_{SS} + D_{ST}) \cdot \max\{\frac{D_{ST}}{D_{SS}}, \frac{D_{SS}}{D_{ST}}\}$;
    $div \leftarrow Div(p|V, S_U)$;
    $S_{score}(p) \leftarrow uncr + \beta \cdot div$;
  end for
  $Q \leftarrow Greedy(S_U, S_{score}, b)$;
  $S_U \leftarrow S_U - Q$;
  $Q \leftarrow Annotation(Q)$;
  $V \leftarrow V \cup Q$;
  $M_S, M_T \leftarrow Train(S_L \cup V)$;
  $M_{ST} \leftarrow Train_{KD}(S_L \cup V)$;
end while

$$\max_{V \subset S_U} \quad \sum_{p \in V} (Uncr(p) + \beta Div(p|V, S_U)),$$
$$\text{s.t.} \quad |V| = b,$$

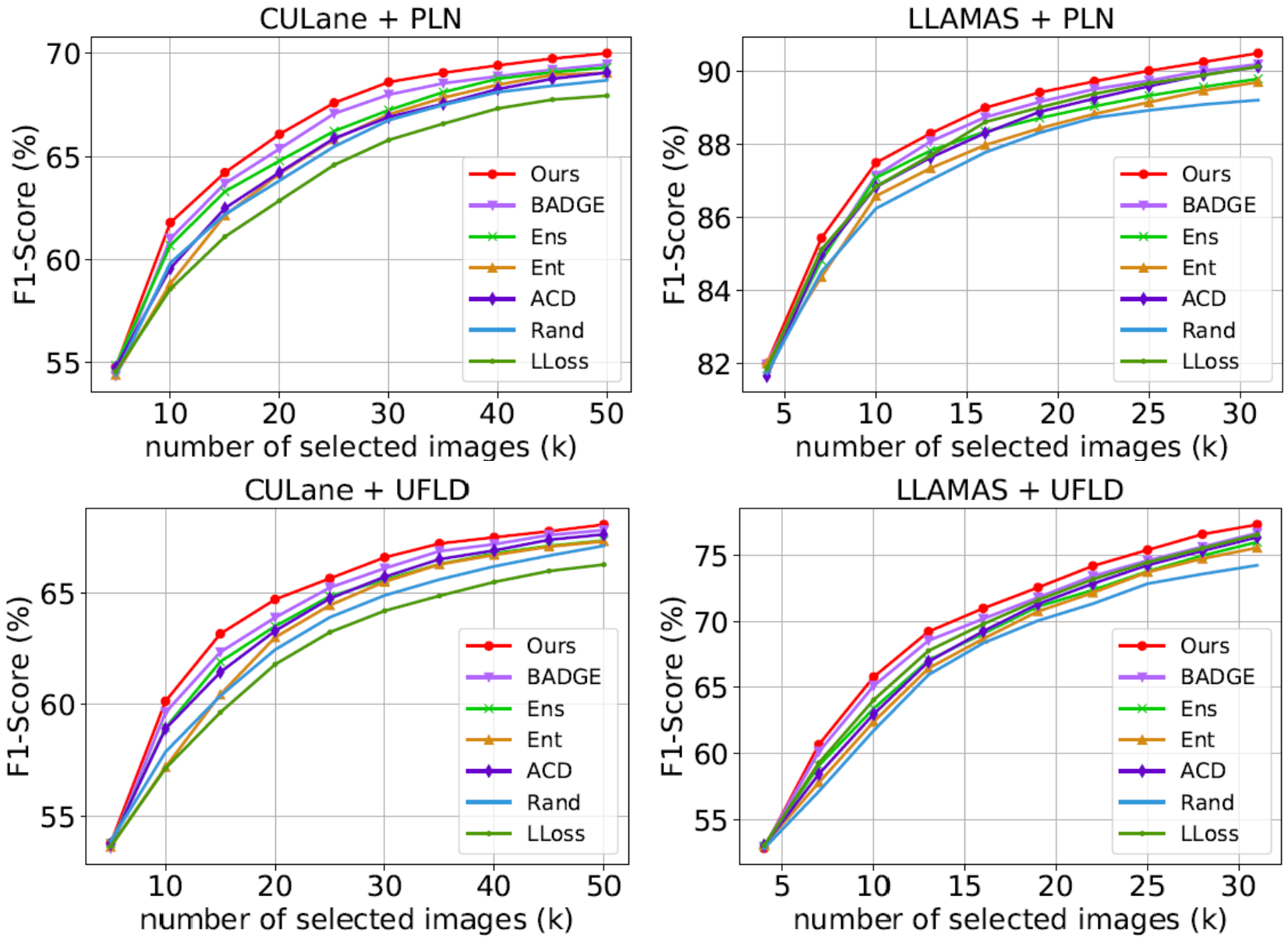
where β is a weighting factor and b is the annotation budget (number of selected samples).

Experiment

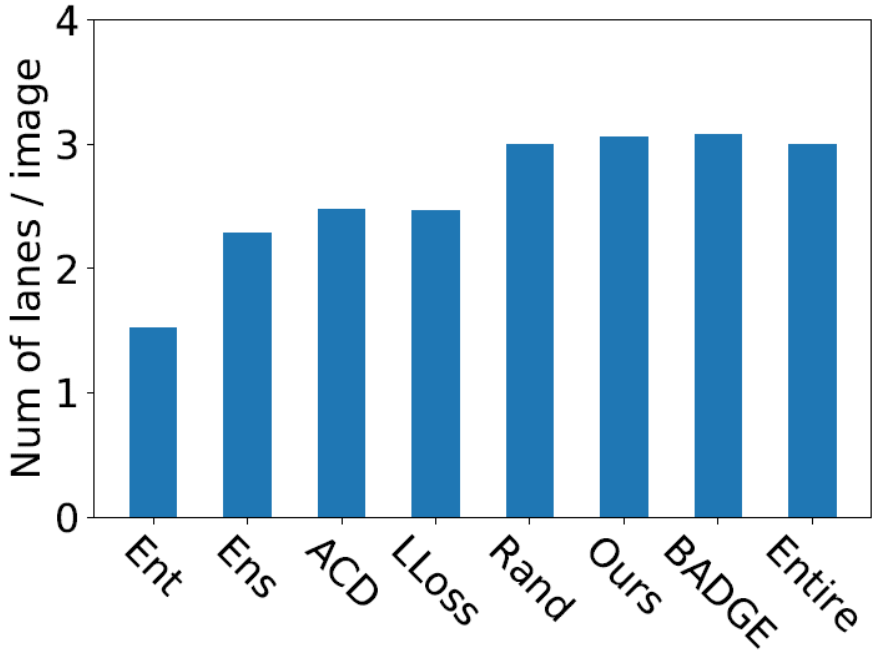
Compared methods:

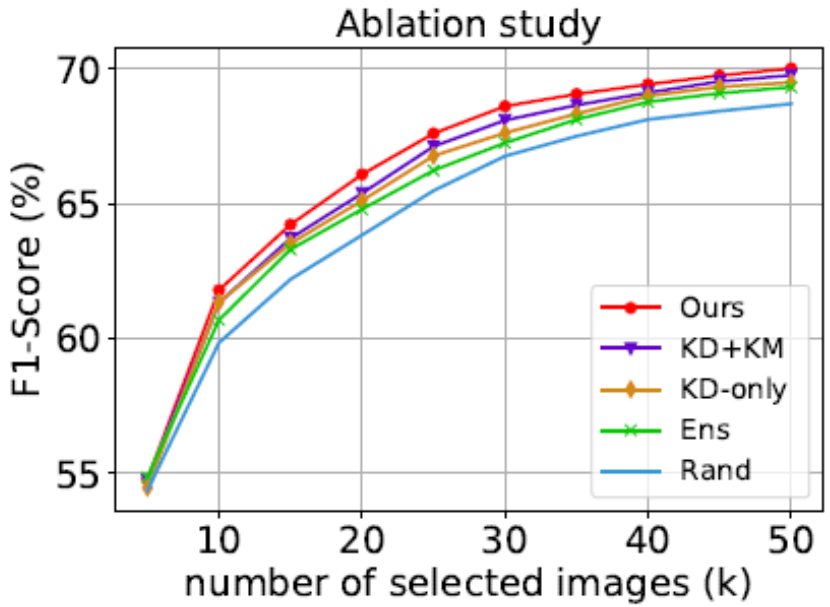
- Random (Rand).
- Entropy (Ent).
- Ensemble (Ens).
- ACD: This method is designed specifically for object detection. It incorporates the spatial information to estimate the entropy.
- LLoss: This method adds a header to the network to estimate the loss of each sample. Samples with largest predicted losses are selected.
- BADGE: This method combines an uncertainty metric (gradient norm) and a diversity metric (KMeans++) to select samples.

Experiment


CULane + PLN
F1-Score (%)
number of selected images (k)
LLAMAS + PLN
F1-Score (%)
number of selected images (k)
CULane + UFLD
F1-Score (%)
number of selected images (k)
LLAMAS + UFLD
F1-Score (%)
number of selected images (k)
Ours
BADGE
Ens
Ent
ACD
Rand
LLoss

Experiment


Num of lanes / image
0
1
2
3
4
Ent
Ens
ACD
LLoss
Rand
Ours
BADGE
Entire


Ablation study
F1-Score (%)
55
60
65
70
10
20
30
40
50
number of selected images (k)
Ours
KD+KM
KD-only
Ens
Rand

南京航空航天大學

Nanjing University of Aeronautics and Astronautics

Neural Networks Are More Productive Teachers Than Human Raters: Active Mixup for Data-Efficient Knowledge Distillation from a Blackbox Model

Dongdong Wang [1*] Yandong Li [1*] Liqiang Wang[1] Boqing Gong[2]

[1]University of Central Florida [2]Google

{daniel.wang, liyandong}@Knights.ucf.edu lwang@cs.ucf.edu bgong@google.com

CVPR 2020

A new problem scenario

How to distill knowledge from a black-box teacher model in a data-efficient manner?

- ✓ The distilled student network should perform well as the teacher model as possible at the inference time.
- ✓ The number of queries to the black-box teacher model should be minimized to save costs.
- ✓ Using as a small number of examples as possible to save data collection efforts.

⇩

Mix-up + Active Learning

Method

Mix-Up:

- Given two natural images x_i and x_j , mix-up generates multiple synthetic images by a convex combination of the two with different coefficients,

$$\hat{x}_{ij}(\lambda) = \lambda x_i + (1-\lambda)x_j,$$

- where the coefficient $\lambda \in [0, 1]$. Note that this notation also includes the original unlabeled data x_i and x_j when $\lambda = 1$ and $\lambda = 0$, respectively.

Method

Active Learning:

- Let $\{\hat{x}_{ij}(\lambda), \lambda \in [0,1], i \neq j\}$ denote the augmented pool of images. Using active learning strategies to query the teacher model to obtain the (soft) labels for these synthetic images.

- We define the student neural network's confidence over an input x as:

$$C_1(x) := \max_y P_S(y|x),$$

- We define the student network's confidence over an image pair x_i and x_j as the following:

$$C_2(x_i, x_j) := \min_\lambda C_1(\hat{x}_{ij}(\lambda)), \quad \lambda \in [0,1],$$

- The synthetic $\hat{x}_{ij}(\lambda^*)$ Image is selected into the query set if the confidence score $C_2(x_i, x_j)$ is among the lowest k ones.

Method

Algorithm 1 Data-efficient blackbox knowledge distillation

INPUT: Pre-trained teacher model $\mathcal{M}^T$
INPUT: A small set of unlabeled images $X = \{x_i\}_{i=1}^n$
INPUT: Hyper-parameters (learning rate, subset size, etc.)
OUTPUT: Student network $\mathcal{M}^S$

Query $\mathcal{M}^T$ and acquire labels Y_0 for all images in X
Train an initial student network $\mathcal{M}_0^S$ with (X, Y_0)
Construct a synthetic image pool $\mathcal{P} = \{\hat{x}_{ij}(\lambda)\}$ by using the unlabeled images X with eq. (1)
Initialize $\mathcal{P}_1^s = X, \mathcal{Y}_1 = \mathcal{Y}_0$.
for $t = 1, 2..., T$ **do**
    Select a subset $\Delta\mathcal{P}_t^s$ from $\mathcal{P}$ with lowest confidence scores $\{C_2(x_i, x_j)\}$ returned by student $\mathcal{M}_{t-1}^S$
    Query $\mathcal{M}^T$, acquire labels $\Delta\mathcal{Y}_t$ for all images $\Delta\mathcal{P}_t^s$
    $\mathcal{P}_t^s \leftarrow \mathcal{P}_t^s \cup \Delta\mathcal{P}_t^s, \mathcal{Y}_t \leftarrow \mathcal{Y}_t \cup \Delta\mathcal{Y}_t$
    Train a new student network $\mathcal{M}_t^S$ with $(\mathcal{P}_t^s, \mathcal{Y}_t)$
    Update $\mathcal{P} \leftarrow \mathcal{P}$ - $\Delta\mathcal{P}_t^s$
end for

Experiment

Evaluation Metric:

- Classification accuracy
- Success rate:
  - The ratio between the student network's classification accuracy and the teacher's accuracy.

Compared methods:

- Zero-shot knowledge distillation (ZSKD) ➔ White-box teacher model.
- Few-shot knowledge distillation (FSKD) ➔ White-box / Black-box teacher model.
- Vanilla knowledge distillation ➔ Can access whole dataset, as an upper bound.

Experiment

Task (Model)	Teacher	KD Accuracy	Success	Black/White	Queries	Unlabeled Data
Places365-Standard (ZSKD) [32]	–	–	–	–	–	0
Places365-Standard (FSKD [21])	53.69	38.18	71.11	White	480,000	80,000
Places365-Standard (KD)	53.69	49.01	90.35	Black	1,800,000	1,800,000
Places365-Standard (Ours)	53.69	**45.71**	**85.14**	Black	480,000	80,000
CIFAR-10 (ZSKD) [32]	83.03*	69.56*	83.78	White	>2,000,000	0
CIFAR-10 (FSKD [21])	83.07	40.58	48.85	White	40,000	2,000
CIFAR-10 (KD)	83.07	80.01	96.31	Black	50,000	50,000
CIFAR-10 (Ours)	83.07	**74.60**	**89.87**	Black	40,000	2,000
MNIST (ZSKD) [32]	99.34*	**98.77***	99.42	White	>1,200,000	0
MNIST (FSKD [21])	99.29	80.43	81.01	White	24,000	2,000
MNIST (KD)	99.29	99.05	99.76	Black	60,000	60,000
MNIST (Ours)	99.29	98.74	**99.45**	Black	24,000	2,000
Fashion-MNIST(ZSKD) [32]	90.84*	79.62*	87.65	White	>2,400,000	0
Fashion-MNIST (FSKD [21])	90.80	68.64	75.60	White	48,000	2,000
Fashion-MNIST (KD)	90.80	87.79	96.69	Black	60,000	60,000
Fashion-MNIST(Ours)	90.80	**80.90**	**89.10**	Black	48,000	2,000

Experiment

Table 2. Classification accuracy on CIFAR-10 with different numbers of real images and selected synthetic images.

Real images / Selected Syn.	0.5K	1K	2K	4K	8K	16K
0	44.72	56.87	68.09	76.59	83.61	86.89
5K	66.97	71.67	77.76	81.76	85.76	87.05
10K	73.60	77.27	81.27	83.27	86.56	88.79
20K	77.44	81.18	84.19	86.29	88.07	89.01
40K	82.28	84.25	86.06	87.71	89.00	90.49
80K	85.18	86.53	87.89	88.71	89.61	90.96
160K	86.56	88.94	89.42	90.26	90.87	91.51

Experiment

Table 3. Classification accuracy on Places365-Standard with different numbers of real images and selected synthetic images.

Selected Syn. \ Real images	20K	40K	80K
100K	40.72	41.95	43.52
200K	41.15	42.86	44.77
400K	41.94	43.42	45.71

Experiment

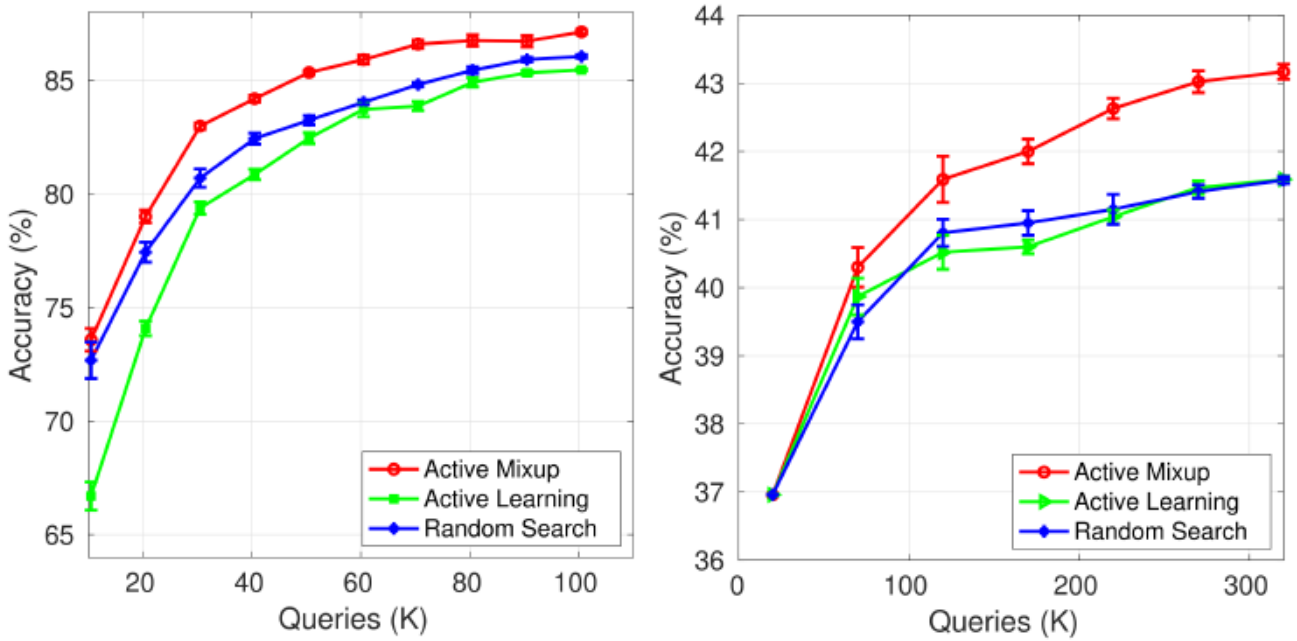


Figure 4. Test accuracy of student networks vs. number of queries into the blackbox teacher model on CIFAR-10 (left) and Places365-Standard (right). We use 500 and 20K natural images for the two datasets, respectively. The plot for CIFAR-10 starts from first active learning stage ($t = 1$ in Algorithm 1) and the one for Places365 starts from the initial student network training by natural images. The initial student network for CIFAR-10 trained by using natural images only yields 43.67% accuracy.

Experiment

Table 4. CIFAR-10 classification accuracy by the student neural networks which are distilled by using out-of-domain data.

Selected Syn.	10K	20K	40K	80K
Accuracy (%)	64.10	71.39	77.89	83.03

Table 5. CIFAR-10 classification accuracy by the student neural networks which are distilled by using out-of-domain data. We set the number of selected synthetic images to 40K and vary the numbers of real images.

Real images	500	1000	1500	2000
Accuracy (%)	70.21	74.60	75.54	77.89

THANKS