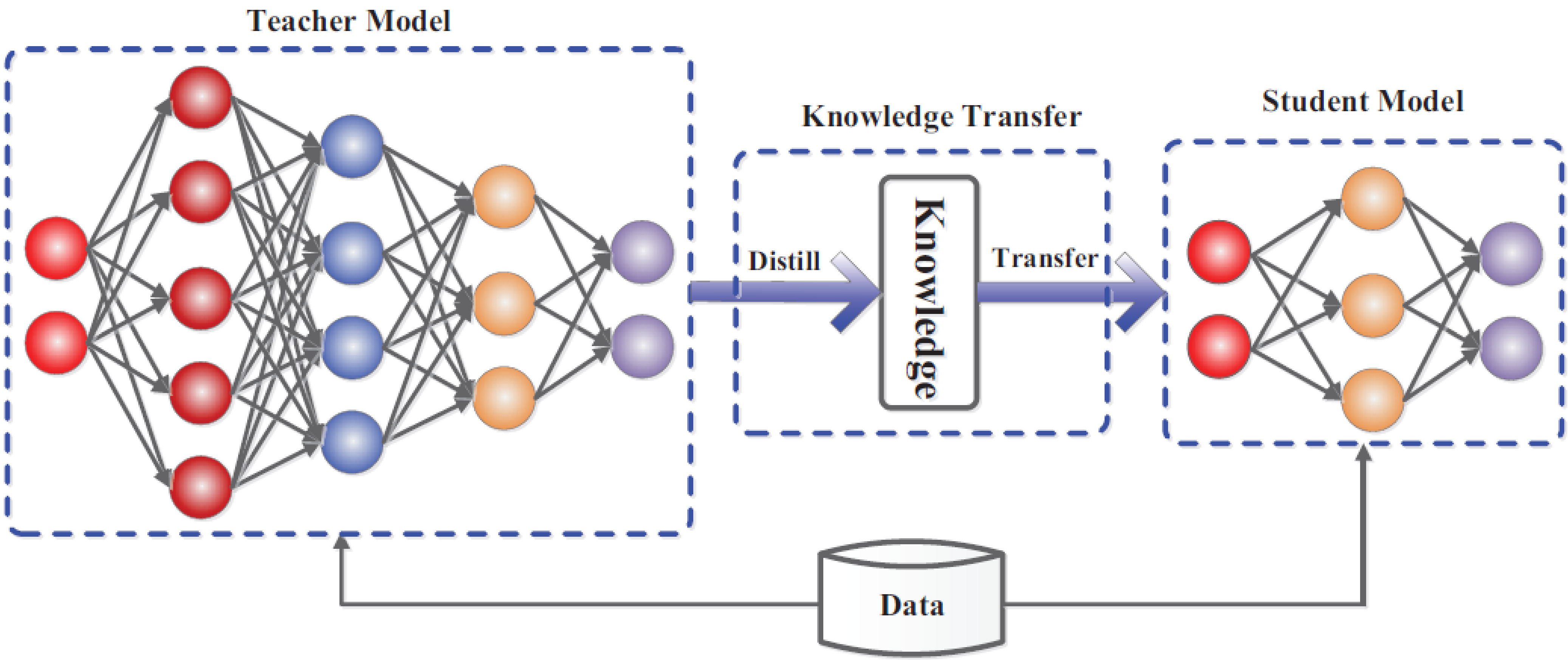
南京航空航天大學
Nanjing University of Aeronautics and Astronautics

# Mosaicking to Distill: Knowledge Distillation from Out-of-Domain Data

**Gongfan Fang**[1,4], **Yifan Bao**[1], **Jie Song**[1], **Xinchao Wang**[2], **Donglin Xie**[1]
**Chengchao Shen**[3], **Mingli Song**[1]*
[1]Zhejiang University, [2]National University of Singapore, [3]Central South University
[4]Alibaba-Zhejiang University Joint Institute of Frontier Technologies
{fgf,yifanbao,sjie,donglinxie,brooksong}@zju.edu.cn
xinchao@nus.edu.sg, scc.cs@csu.edu.cn

***NIPS 2021***

# Knowledge Distillation

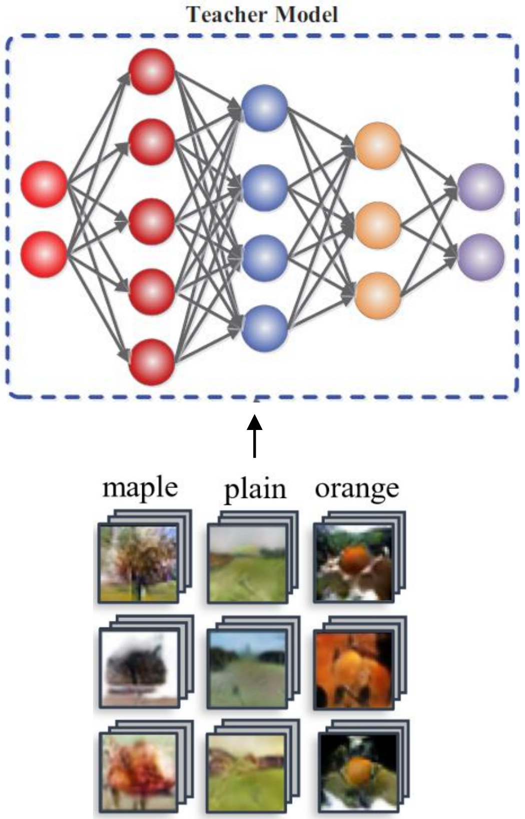


**Goal: transfer knowledge from a large model to a small model for model compression and acceleration.**

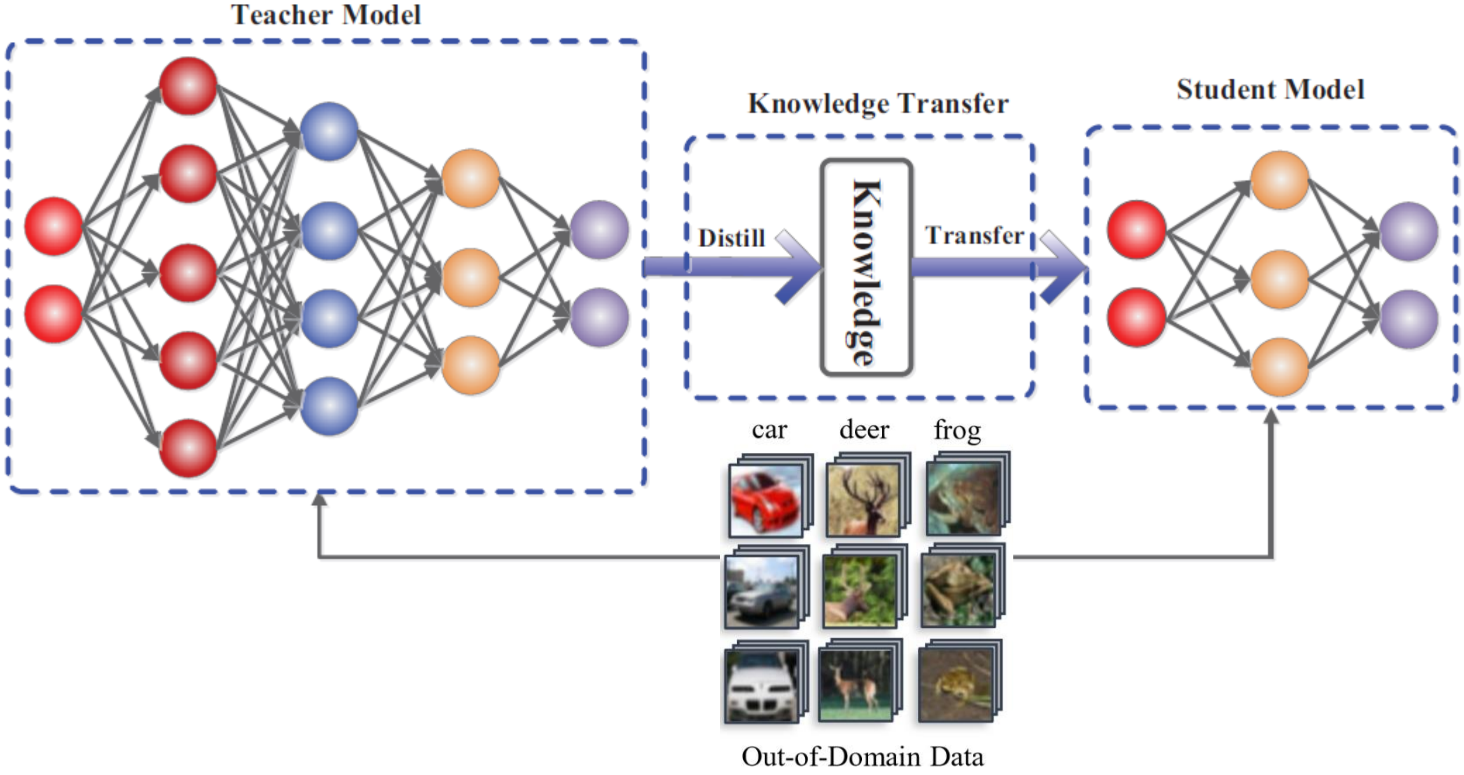
# Out-of-domain knowledge distillation (OOD-KD)

However, the original training data or even the data domain is often unreachable due to privacy or copyright reasons.

Teacher model training

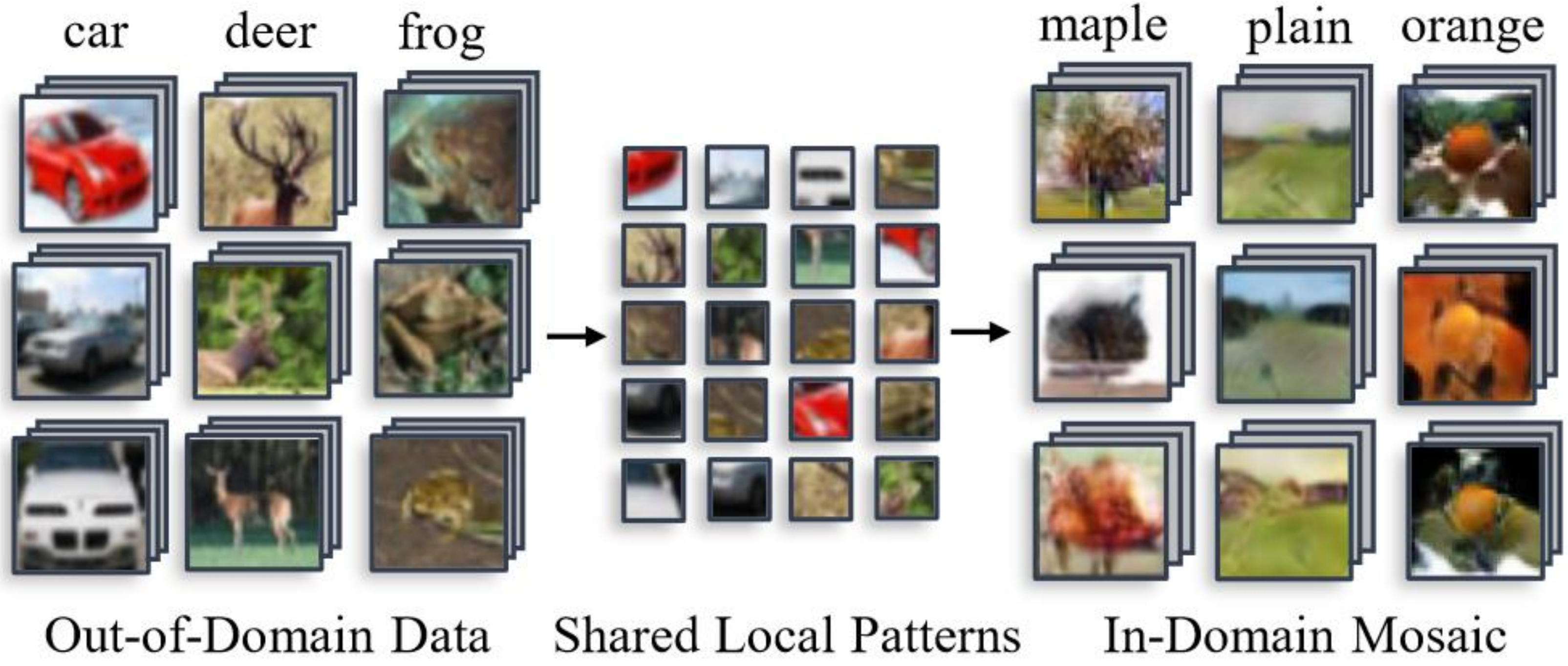


Student model training



# MosaicKD

Even though data from different domains exhibit divergent global distributions, their local distributions, such as patches in images, may however resemble each other.



The core idea of MosaicKD is to synthesize in-domain data.

# Formulation: vanilla KD

## Vanilla KD:

Dataset: $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}, P_{\mathcal{X}\times\mathcal{Y}}\}$ Teacher model: $T(x;\theta_t)$ Student model: $S(x;\theta_s)$

$$\theta_s^* = \arg\min_{\theta_s} \mathbb{E}_{(x,y)\sim P_{\mathcal{X}\times\mathcal{Y}}}\left[\ell_{\text{KL}}(T(x;\theta_t)\|S(x;\theta_s)) + \ell_{\text{CE}}(S(x;\theta_s),y)\right]$$

## OOD KD:

OOD dataset: $\mathcal{D}' = \{\mathcal{X}', \mathcal{Y}', P_{\mathcal{X}'\times\mathcal{Y}')}\}$, where $\mathcal{X}' \neq \mathcal{X}$ and $\mathcal{Y}' \neq \mathcal{Y}$

$$X' = \{x_1', x_2', ..., x_N'; x_i' \in \mathbb{R}^{H\times W\times 3}\}$$

Generator: $G(z;\theta_g)$ Discriminator: $D(x;\theta_d)$

$$\min_S \max_G \{\mathbb{E}_{x\sim P_G}\left[\ell_{\text{KL}}(T(G(z))\|S(G(z)))\right] : d(P_G, P_{X'}) \leq \epsilon\}$$

# MosaicKD

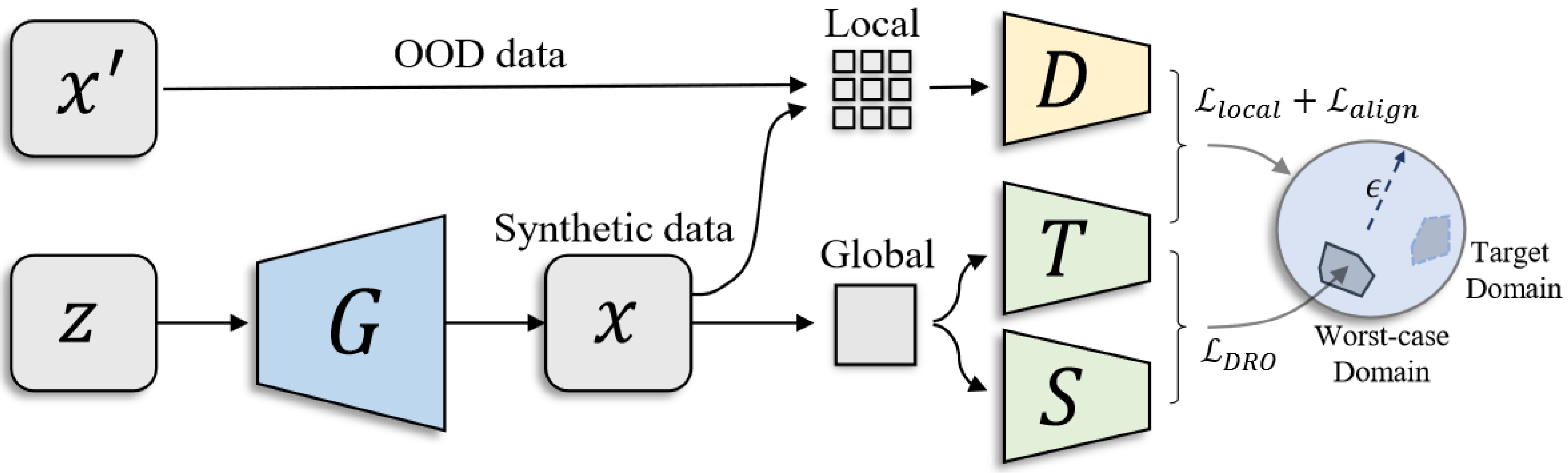


Figure 2: The framework of MosaicKD.

# Formulation: OOD KD

OOD dataset: $X' = \{x'_1, x'_2, \ldots, x'_N; x'_i \in \mathbb{R}^{H\times W\times 3}\}$

$L \times L$ cropping $C(\cdot)$

Patch dataset: $C = \{c_1, c_2, \ldots, c_M; c_i \in \mathbb{R}^{L\times L\times 3}\}$

**Patch Learning:**

$$\min_G \max_D \mathcal{L}_{local}(G, D) = \mathbb{E}_{x'\sim P_{X'}}\left[\log D(C(x'))\right] + \mathbb{E}_{z\sim P_z}\left[\log(1 - D(C(G(z))))\right]$$

**Label Space Aligning:**

$$\min_G \mathcal{L}_{align}(G, D, T) = \mathbb{E}_{z\sim P_z}\left[H\left[p(y|G(z), \theta_t)\right]\right]$$

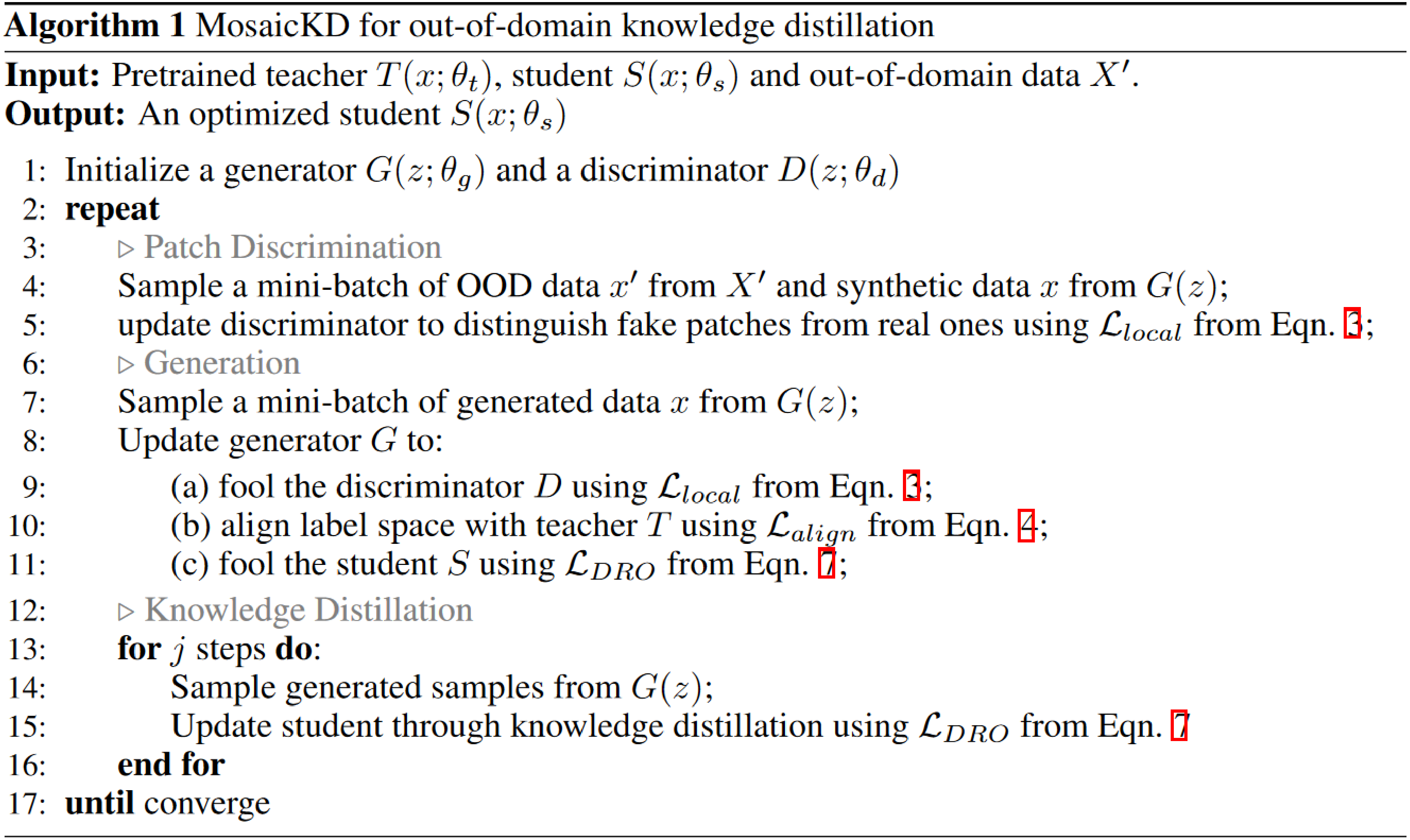
**MosaicKD:**

$$\min_S \max_G \mathcal{L}_{DRO}(G, D, S, T) = \{\mathbb{E}_{z\sim p_z(z)}\left[\ell_{\mathrm{KL}}(T(G(z))\|S(G(z)))\right] : \mathcal{R}(G, D, T)) \leq \epsilon\}$$

⇩

$$\min_S \max_G \mathcal{L}_{DRO}(G, D, S, T) = \mathbb{E}_{x\sim P_z}\left[\ell_{\mathrm{KL}}(T(G(z))\|S(G(z)))\right] - \lambda\mathcal{R}(G, D, T))$$

# Pseudo-code

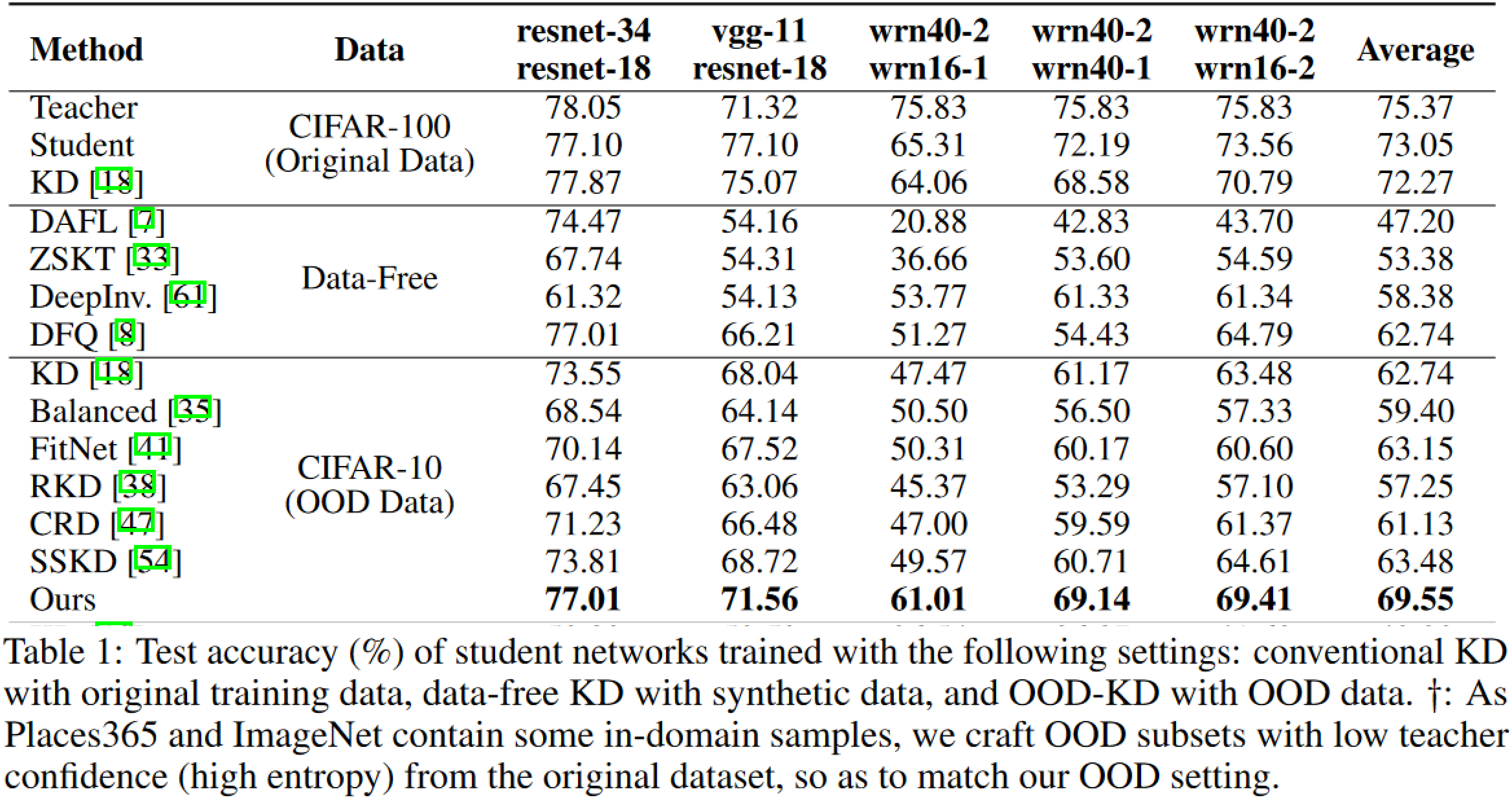


**Algorithm 1** MosaicKD for out-of-domain knowledge distillation

**Input:** Pretrained teacher $T(x;\theta_t)$, student $S(x;\theta_s)$ and out-of-domain data $X'$.
**Output:** An optimized student $S(x;\theta_s)$

```
Initialize a generator $G(z;\theta_g)$ and a discriminator $D(z;\theta_d)$
repeat
    ▷ Patch Discrimination
    Sample a mini-batch of OOD data $x'$ from $X'$ and synthetic data $x$ from $G(z)$;
    update discriminator to distinguish fake patches from real ones using $\mathcal{L}_{local}$ from Eqn. 3;
    ▷ Generation
    Sample a mini-batch of generated data $x$ from $G(z)$;
    Update generator $G$ to:
        (a) fool the discriminator $D$ using $\mathcal{L}_{local}$ from Eqn. 3;
        (b) align label space with teacher $T$ using $\mathcal{L}_{align}$ from Eqn. 4;
        (c) fool the student $S$ using $\mathcal{L}_{DRO}$ from Eqn. 7;
    ▷ Knowledge Distillation
    for $j$ steps do:
        Sample generated samples from $G(z)$;
        Update student through knowledge distillation using $\mathcal{L}_{DRO}$ from Eqn. 7
    end for
until converge
```

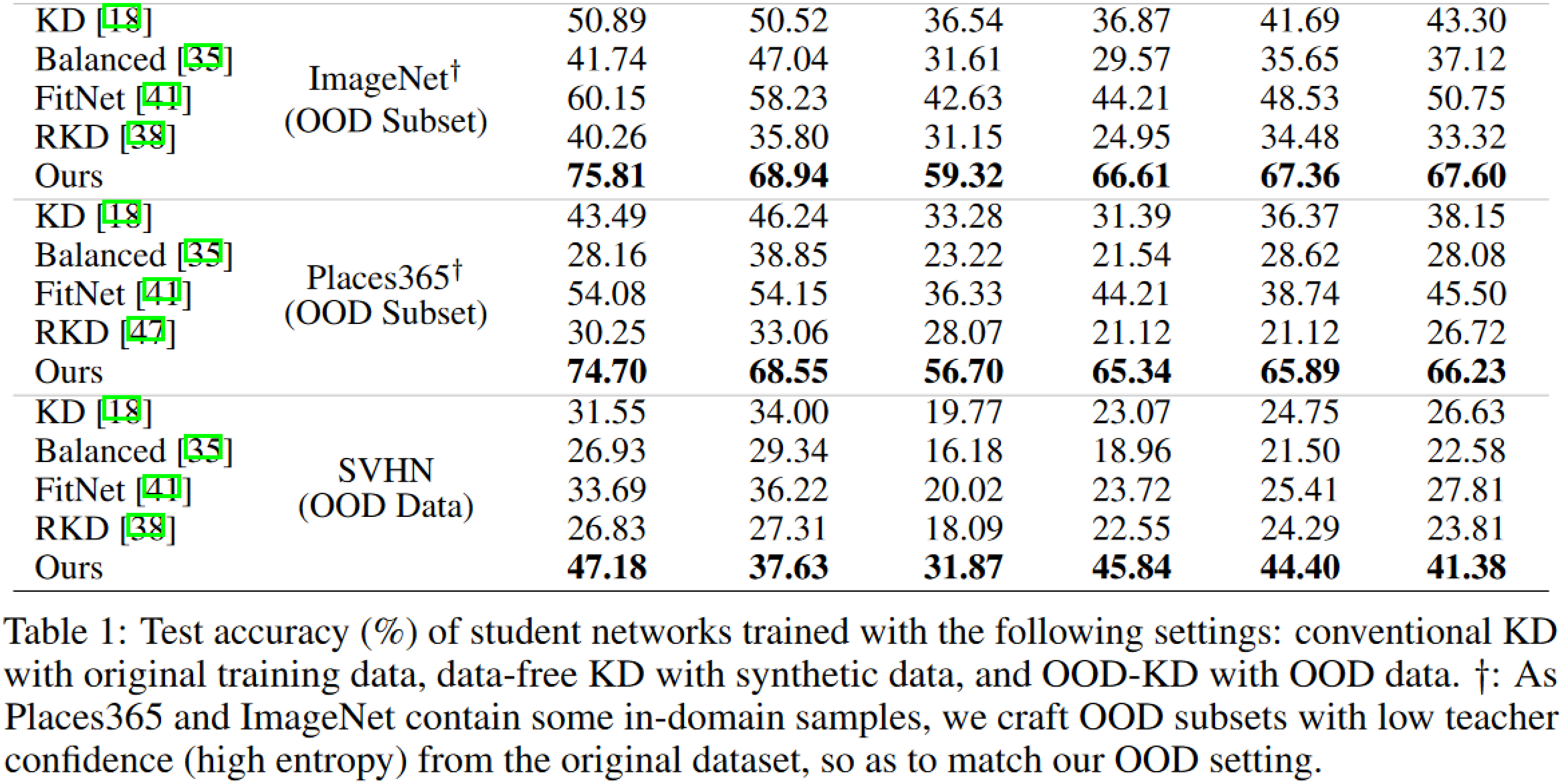
# Experiment



| Method | Data | resnet-34<br>resnet-18 | vgg-11<br>resnet-18 | wrn40-2<br>wrn16-1 | wrn40-2<br>wrn40-1 | wrn40-2<br>wrn16-2 | Average |
|---|---|---|---|---|---|---|---|
| Teacher | CIFAR-100<br>(Original Data) | 78.05 | 71.32 | 75.83 | 75.83 | 75.83 | 75.37 |
| Student | | 77.10 | 77.10 | 65.31 | 72.19 | 73.56 | 73.05 |
| KD [18] | | 77.87 | 75.07 | 64.06 | 68.58 | 70.79 | 72.27 |
| DAFL [7] | Data-Free | 74.47 | 54.16 | 20.88 | 42.83 | 43.70 | 47.20 |
| ZSKT [33] | | 67.74 | 54.31 | 36.66 | 53.60 | 54.59 | 53.38 |
| DeepInv. [61] | | 61.32 | 54.13 | 53.77 | 61.33 | 61.34 | 58.38 |
| DFQ [8] | | 77.01 | 66.21 | 51.27 | 54.43 | 64.79 | 62.74 |
| KD [18] | CIFAR-10<br>(OOD Data) | 73.55 | 68.04 | 47.47 | 61.17 | 63.48 | 62.74 |
| Balanced [35] | | 68.54 | 64.14 | 50.50 | 56.50 | 57.33 | 59.40 |
| FitNet [41] | | 70.14 | 67.52 | 50.31 | 60.17 | 60.60 | 63.15 |
| RKD [38] | | 67.45 | 63.06 | 45.37 | 53.29 | 57.10 | 57.25 |
| CRD [47] | | 71.23 | 66.48 | 47.00 | 59.59 | 61.37 | 61.13 |
| SSKD [54] | | 73.81 | 68.72 | 49.57 | 60.71 | 64.61 | 63.48 |
| Ours | | **77.01** | **71.56** | **61.01** | **69.14** | **69.41** | **69.55** |

Table 1: Test accuracy (%) of student networks trained with the following settings: conventional KD with original training data, data-free KD with synthetic data, and OOD-KD with OOD data. †: As Places365 and ImageNet contain some in-domain samples, we craft OOD subsets with low teacher confidence (high entropy) from the original dataset, so as to match our OOD setting.

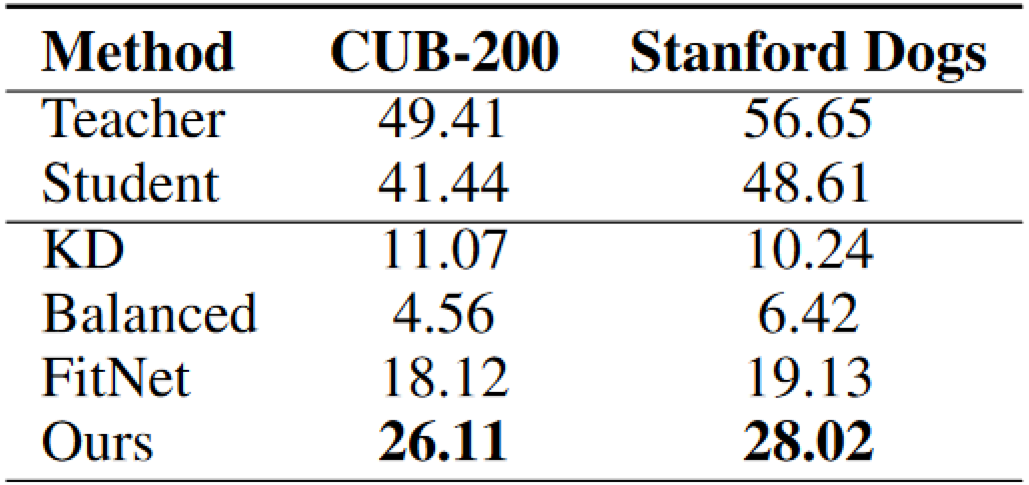
# Experiment



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| KD [18] | ImageNet$^\dagger$ (OOD Subset) | 50.89 | 50.52 | 36.54 | 36.87 | 41.69 | 43.30 |
| Balanced [35] | | 41.74 | 47.04 | 31.61 | 29.57 | 35.65 | 37.12 |
| FitNet [41] | | 60.15 | 58.23 | 42.63 | 44.21 | 48.53 | 50.75 |
| RKD [38] | | 40.26 | 35.80 | 31.15 | 24.95 | 34.48 | 33.32 |
| Ours | | **75.81** | **68.94** | **59.32** | **66.61** | **67.36** | **67.60** |
| KD [18] | Places365$^\dagger$ (OOD Subset) | 43.49 | 46.24 | 33.28 | 31.39 | 36.37 | 38.15 |
| Balanced [35] | | 28.16 | 38.85 | 23.22 | 21.54 | 28.62 | 28.08 |
| FitNet [41] | | 54.08 | 54.15 | 36.33 | 44.21 | 38.74 | 45.50 |
| RKD [47] | | 30.25 | 33.06 | 28.07 | 21.12 | 21.12 | 26.72 |
| Ours | | **74.70** | **68.55** | **56.70** | **65.34** | **65.89** | **66.23** |
| KD [18] | SVHN (OOD Data) | 31.55 | 34.00 | 19.77 | 23.07 | 24.75 | 26.63 |
| Balanced [35] | | 26.93 | 29.34 | 16.18 | 18.96 | 21.50 | 22.58 |
| FitNet [41] | | 33.69 | 36.22 | 20.02 | 23.72 | 25.41 | 27.81 |
| RKD [38] | | 26.83 | 27.31 | 18.09 | 22.55 | 24.29 | 23.81 |
| Ours | | **47.18** | **37.63** | **31.87** | **45.84** | **44.40** | **41.38** |

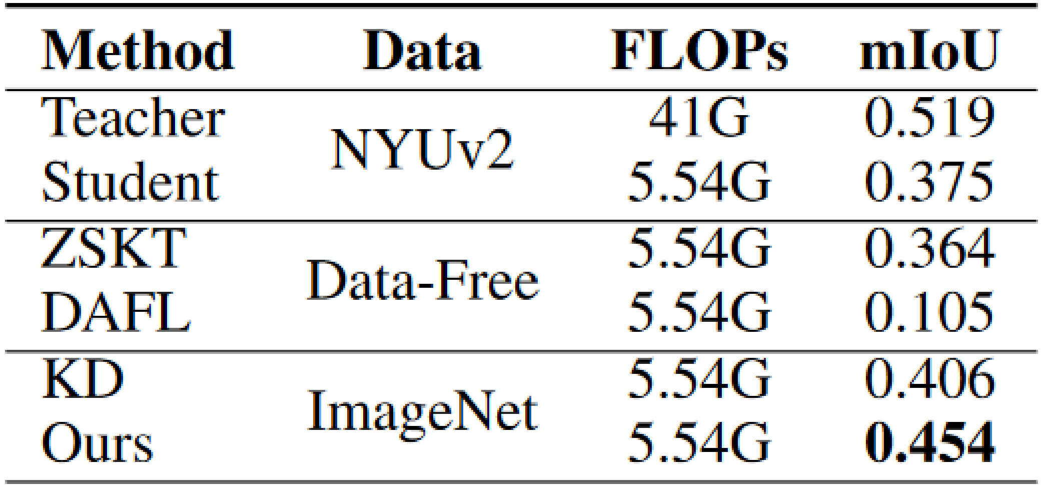
Table 1: Test accuracy (%) of student networks trained with the following settings: conventional KD with original training data, data-free KD with synthetic data, and OOD-KD with OOD data. †: As Places365 and ImageNet contain some in-domain samples, we craft OOD subsets with low teacher confidence (high entropy) from the original dataset, so as to match our OOD setting.

# Experiment



| Method | CUB-200 | Stanford Dogs |
|---|---|---|
| Teacher | 49.41 | 56.65 |
| Student | 41.44 | 48.61 |
| KD | 11.07 | 10.24 |
| Balanced | 4.56 | 6.42 |
| FitNet | 18.12 | 19.13 |
| Ours | **26.11** | **28.02** |

Table 3: Test accuracy of student networks on fine-grained datasets.



| Method | Data | FLOPs | mIoU |
|---|---|---|---|
| Teacher | NYUv2 | 41G | 0.519 |
| Student | | 5.54G | 0.375 |
| ZSKT | Data-Free | 5.54G | 0.364 |
| DAFL | | 5.54G | 0.105 |
| KD | ImageNet | 5.54G | 0.406 |
| Ours | | 5.54G | **0.454** |

Table 2: Mean Intersection over Union (mIoU) of student models on NYUv2 data set.

# Experiment

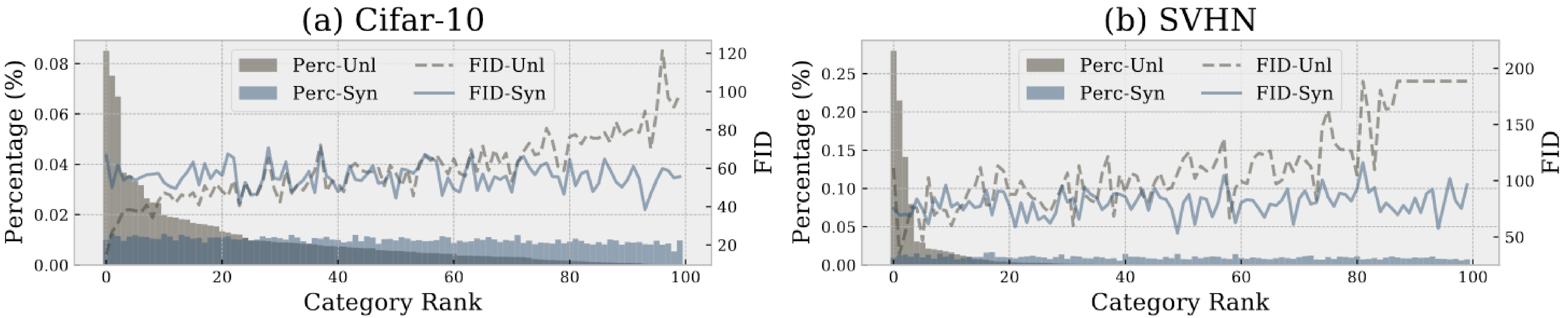
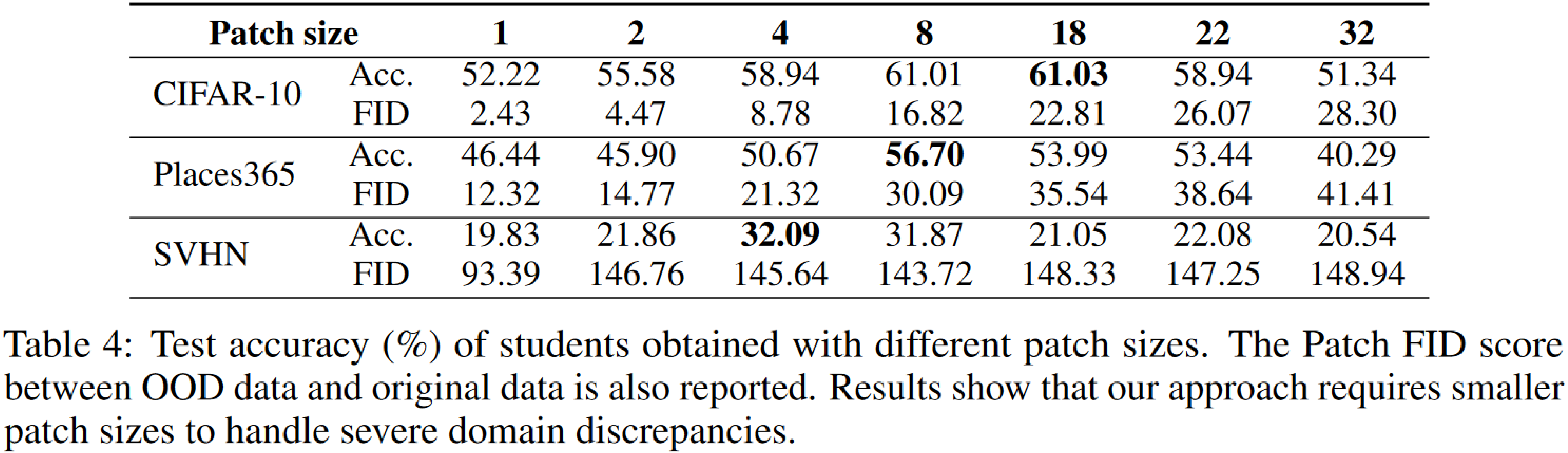


Figure 3: Statistical information of OOD data and out generated data. Category percentage (the first y-axis) and FID score (the second y-axis) to original data (CIFAR-100) is reported.

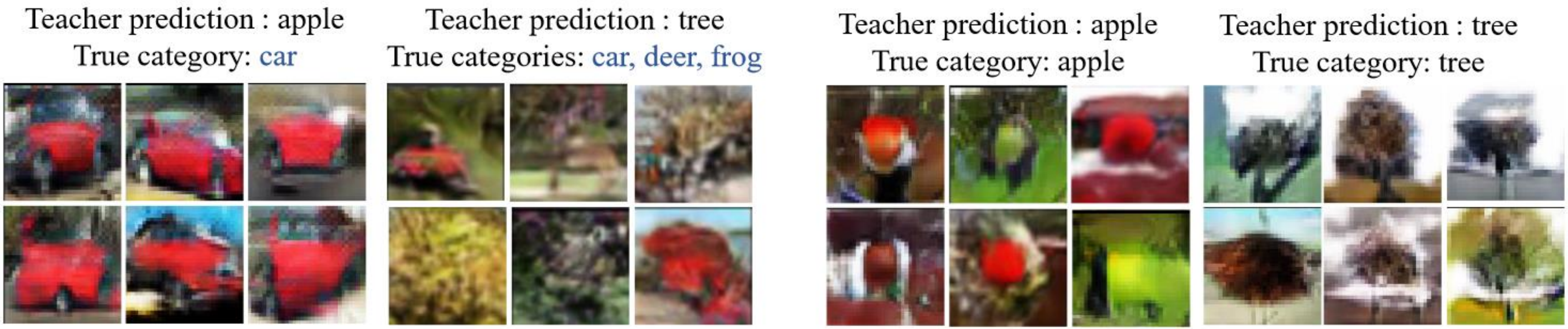
# Experiment



| **Patch size** | | **1** | **2** | **4** | **8** | **18** | **22** | **32** |
|---|---|---|---|---|---|---|---|---|
| CIFAR-10 | Acc. | 52.22 | 55.58 | 58.94 | 61.01 | **61.03** | 58.94 | 51.34 |
| | FID | 2.43 | 4.47 | 8.78 | 16.82 | 22.81 | 26.07 | 28.30 |
| Places365 | Acc. | 46.44 | 45.90 | 50.67 | **56.70** | 53.99 | 53.44 | 40.29 |
| | FID | 12.32 | 14.77 | 21.32 | 30.09 | 35.54 | 38.64 | 41.41 |
| SVHN | Acc. | 19.83 | 21.86 | **32.09** | 31.87 | 21.05 | 22.08 | 20.54 |
| | FID | 93.39 | 146.76 | 145.64 | 143.72 | 148.33 | 147.25 | 148.94 |

Table 4: Test accuracy (%) of students obtained with different patch sizes. The Patch FID score between OOD data and original data is also reported. Results show that our approach requires smaller patch sizes to handle severe domain discrepancies.

# Experiment



Teacher prediction : apple
True category: car

Teacher prediction : tree
True categories: car, deer, frog

Teacher prediction : apple
True category: apple

Teacher prediction : tree
True category: tree

(a) W/O Patch Learning

(b) Patch Learning (Ours)

Figure 4: Visualization of synthetic data with and without patch learning. GANs without patch learning will be trapped by OOD data and fails to present correct semantic for different categories (highlighted in blue). In our method, the semantic can be correctly aligned with target domain.

# THANKS

# Noise-Robust Bidirectional Learning with Dynamic Sample Reweighting

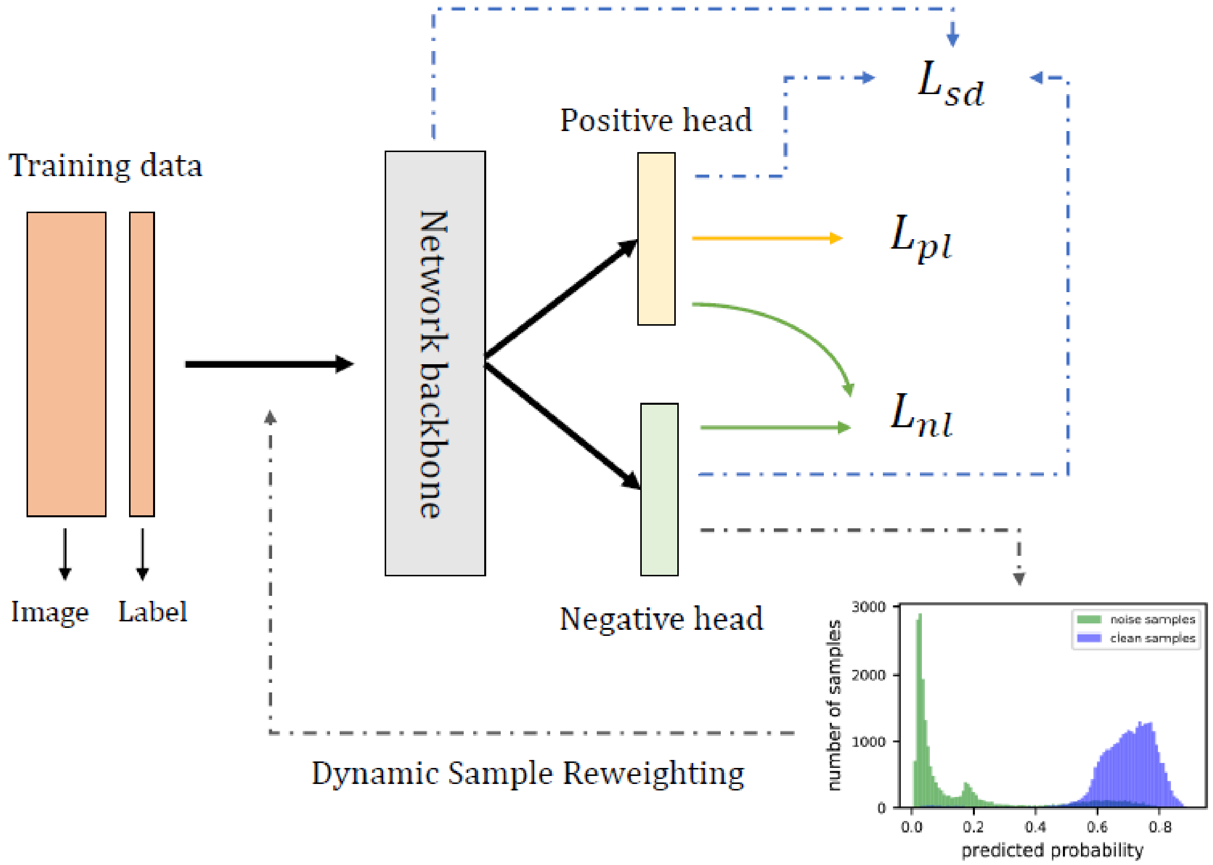


Figure 1: Overview of the BLDR method. We adopt a two-head model, which has a positive head trained in the normal way and a negative head trained with Equation 2. By normalizing the predicted probability of the negative head, we assign a weight to each sample.

$$\mathcal{L}_{pl} = \mathcal{L}_{ce} + \beta\mathcal{L}_{SOP} + \gamma\mathcal{L}_{consistency} + \delta\mathcal{L}_{class}$$

$$\mathcal{L}_{sd} = \sum_{j}^{t} CrossEntropy(p^j, \bar{y}) + \alpha \sum_{j}^{t} KL(p^j, p_{ens}) + \lambda \sum_{j}^{t} \|F_j - F_{pl}\|_2^2$$

$$\mathcal{L}_{nl}(f, \hat{y}) = -\sum_{k=1}^{c} \hat{y}_k \log\left(1 - p_k\right)$$

$\hat{y}$ is a label other than the given label $y$.

⇩

$\hat{y}$ is a label other than the given label $y$ and the corrected label $\tilde{y}$ for every iteration during training.

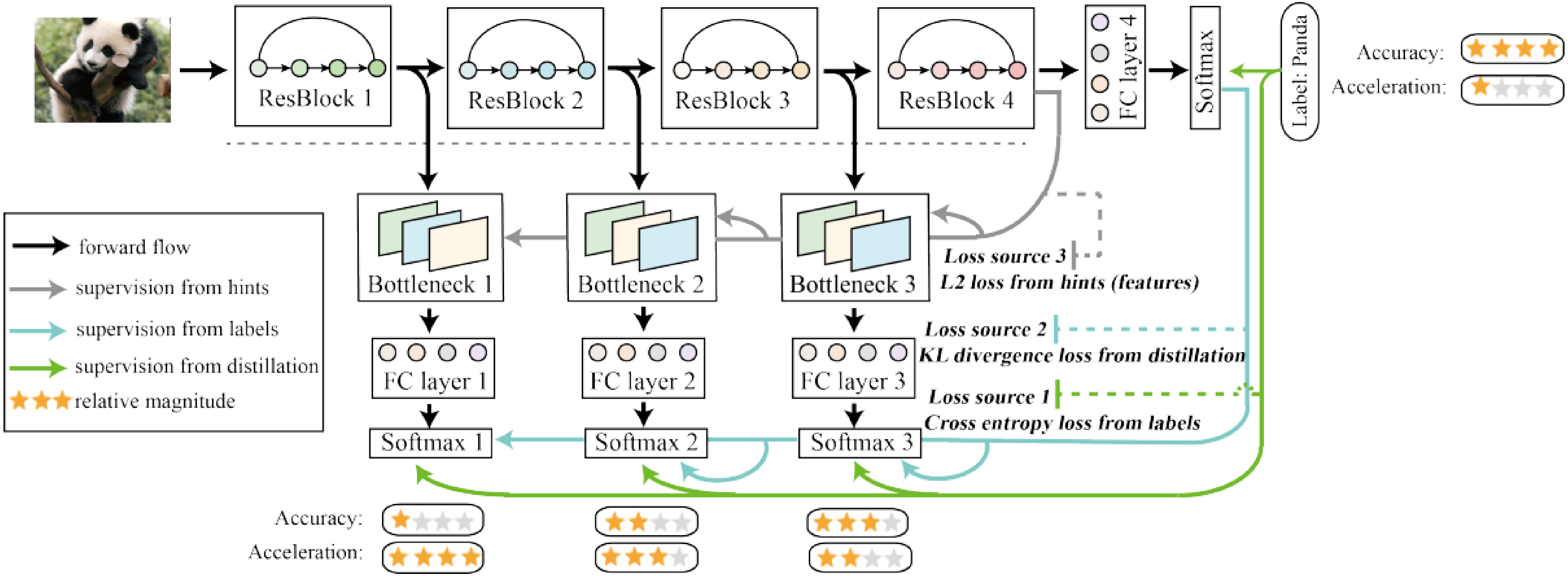


Figure 2. This figure shows the details of a ResNet equipped with proposed self distillation. (i) A ResNet has been divided into four sections according to their depth. (ii) Additional bottleneck and fully connected layers are set after each section, which constitutes multiple classifiers. (iii) All of the classifiers can be utilized independently, with different accuracy and response time. (iv) Each classifier is trained under three kinds of supervision as depicted. (v) Parts under the dash line can be removed in inference.

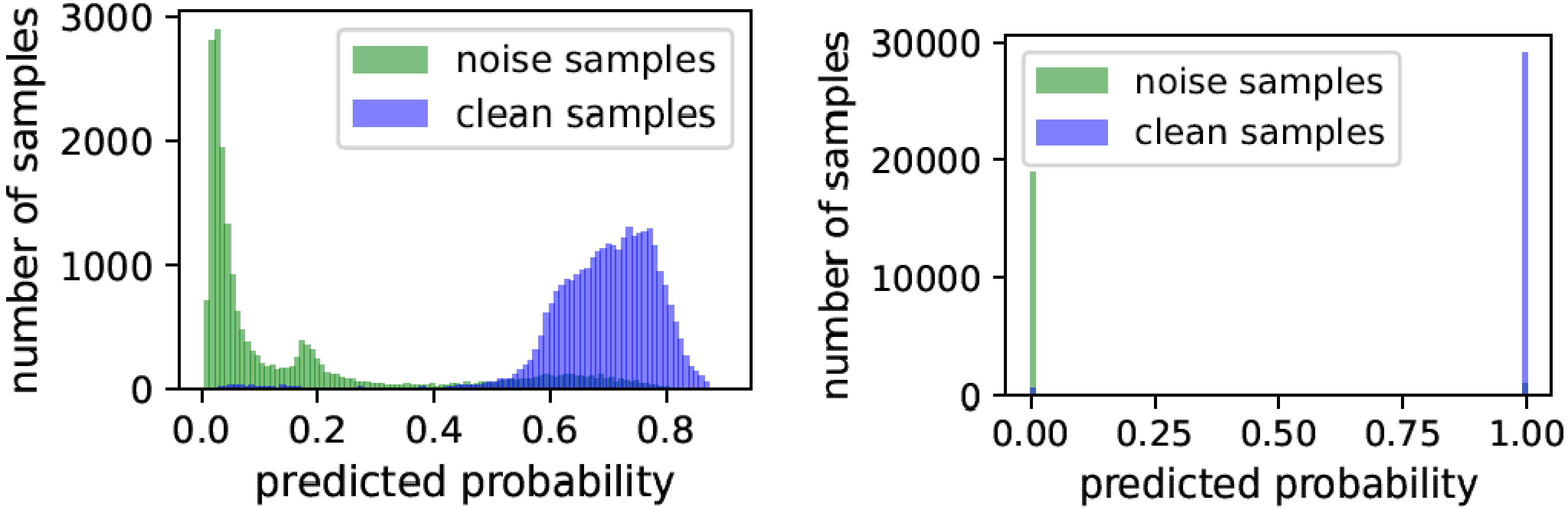


Figure 2: The predicted probability statistics for all CIFAR-10N-Worst training samples on the noisy labels. Figure left only uses bidirectional learning scheme to train the model, while right one combines with the dynamic sample reweighting strategy.

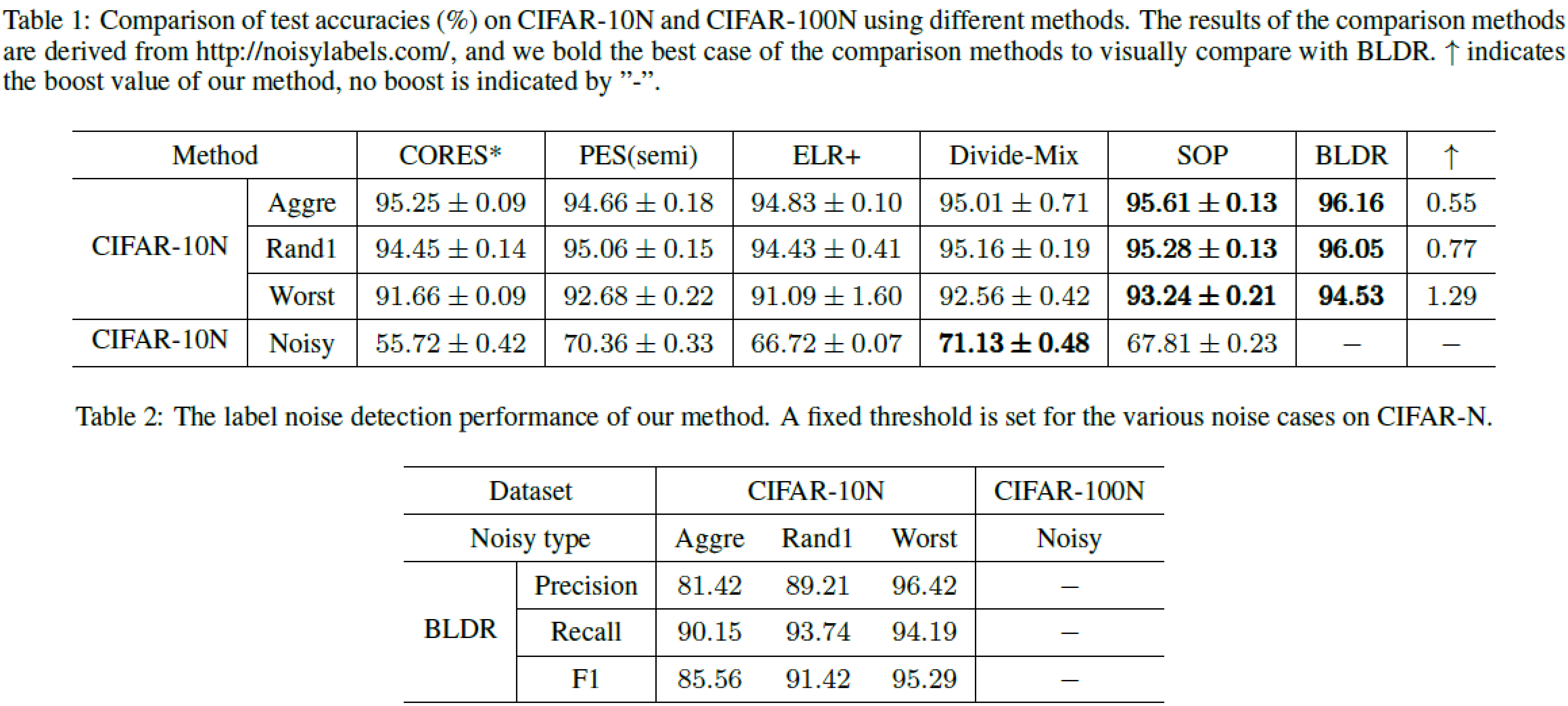


Table 1: Comparison of test accuracies (%) on CIFAR-10N and CIFAR-100N using different methods. The results of the comparison methods are derived from http://noisylabels.com/, and we bold the best case of the comparison methods to visually compare with BLDR. ↑ indicates the boost value of our method, no boost is indicated by ”-”.

| Method | | CORES* | PES(semi) | ELR+ | Divide-Mix | SOP | BLDR | ↑ |
|---|---|---|---|---|---|---|---|---|
| CIFAR-10N | Aggre | 95.25 ± 0.09 | 94.66 ± 0.18 | 94.83 ± 0.10 | 95.01 ± 0.71 | **95.61 ± 0.13** | **96.16** | 0.55 |
| | Rand1 | 94.45 ± 0.14 | 95.06 ± 0.15 | 94.43 ± 0.41 | 95.16 ± 0.19 | **95.28 ± 0.13** | **96.05** | 0.77 |
| | Worst | 91.66 ± 0.09 | 92.68 ± 0.22 | 91.09 ± 1.60 | 92.56 ± 0.42 | **93.24 ± 0.21** | **94.53** | 1.29 |
| CIFAR-10N | Noisy | 55.72 ± 0.42 | 70.36 ± 0.33 | 66.72 ± 0.07 | **71.13 ± 0.48** | 67.81 ± 0.23 | – | – |

Table 2: The label noise detection performance of our method. A fixed threshold is set for the various noise cases on CIFAR-N.

| Dataset | | CIFAR-10N | | | CIFAR-100N |
|---|---|---|---|---|---|
| Noisy type | | Aggre | Rand1 | Worst | Noisy |
| BLDR | Precision | 81.42 | 89.21 | 96.42 | – |
| | Recall | 90.15 | 93.74 | 94.19 | – |
| | F1 | 85.56 | 91.42 | 95.29 | – |

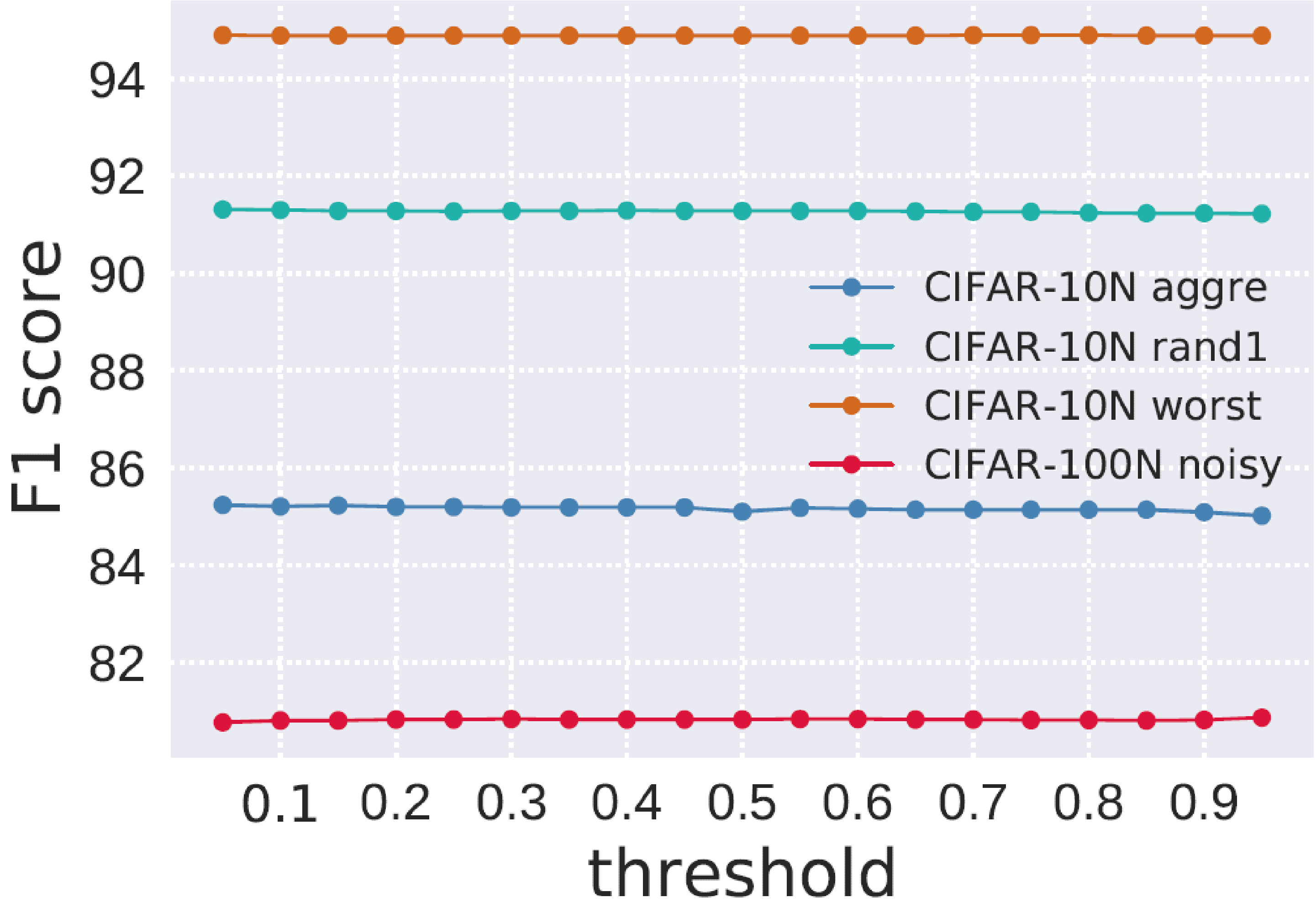


Figure 3: The relationship between the threshold value and the final F1 score obtained among different datasets and different noise cases.