DarkDistill

2025-08-05

August 3–7, 2025
KDD2025

DarkDistill: Difficulty-Aligned Federated Early-Exit Network Training on Heterogeneous Devices

Lehao Qu[1], Shuyuan Li[2], Zimu Zhou[2],
Boyi Liu[1,2], Yi Xu[1], Yongxin Tong[1]
[1]Beihang University
[2]City University of Hong Kong

Outline

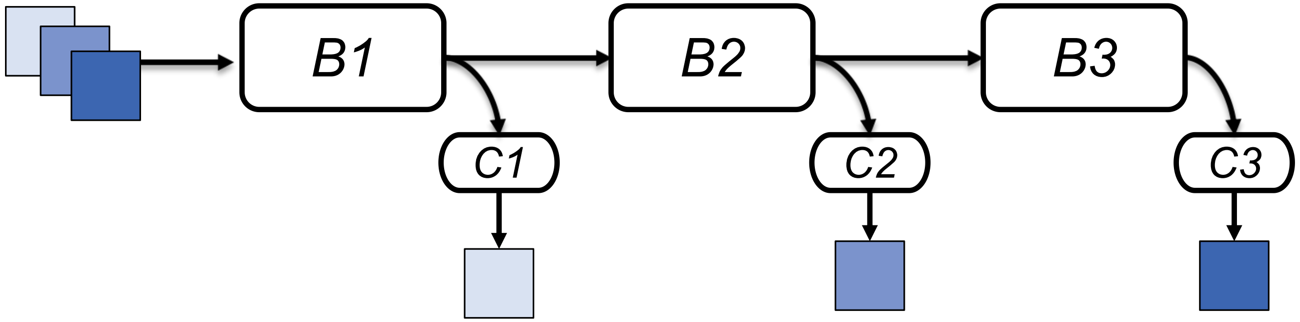
- Background & Motivation
- Problem Statement
- Our Solutions
- Experiments
- Conclusion

Outline

- Background & Motivation
- Problem Statement
- Our Solutions
- Experiments
- Conclusion

Background & Motivation

- **Early-Exit Network (EEN)**



B *Block* **C** *Classifier*

Difficulty-increased inference samples

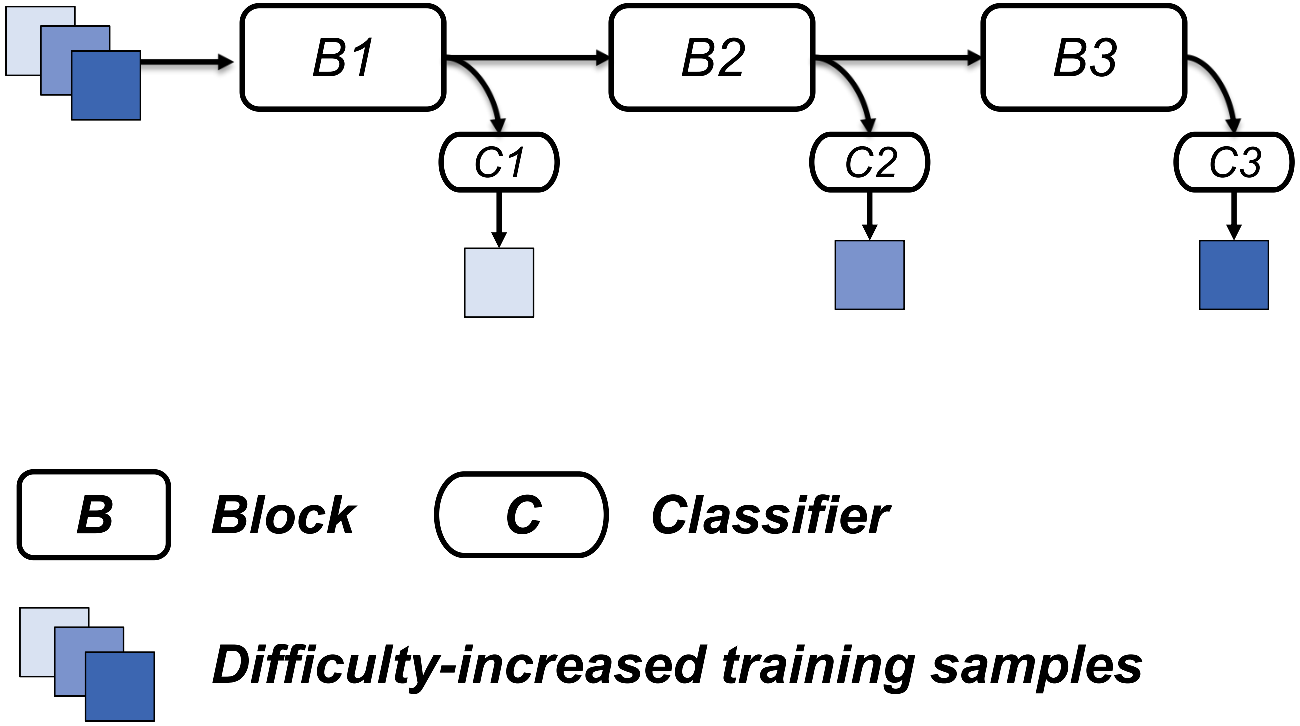
Traffic Analysis

Autonomous Driving

Well-being Monitoring

Background & Motivation

- **Difficulty-aware EEN Training**



Training objective

$$\mathcal{L}(\theta; D) = \sum_{m=1}^{M} \omega^m \mathcal{L}^m(\theta; D) = \sum_{m=1}^{M} \omega^m \sum_{i=1}^{|D|} l_i^m$$

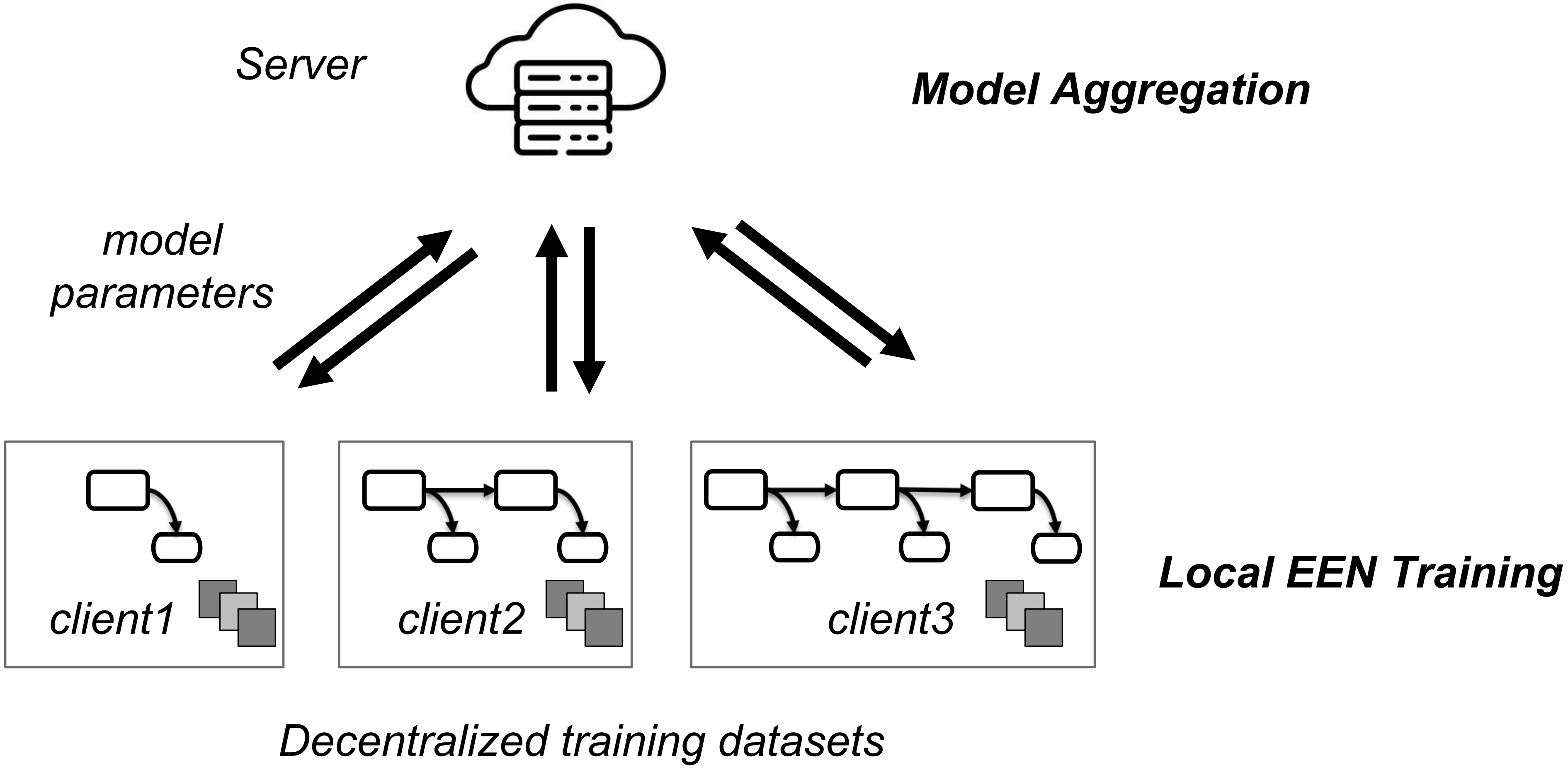
Core ideas

1) BoostNet: Directing samples misclassified by shallow exits to deep ones

2) L2w: Increasing the weight of complex samples on training deep exits

Background & Motivation

- **Federated learning EEN training**

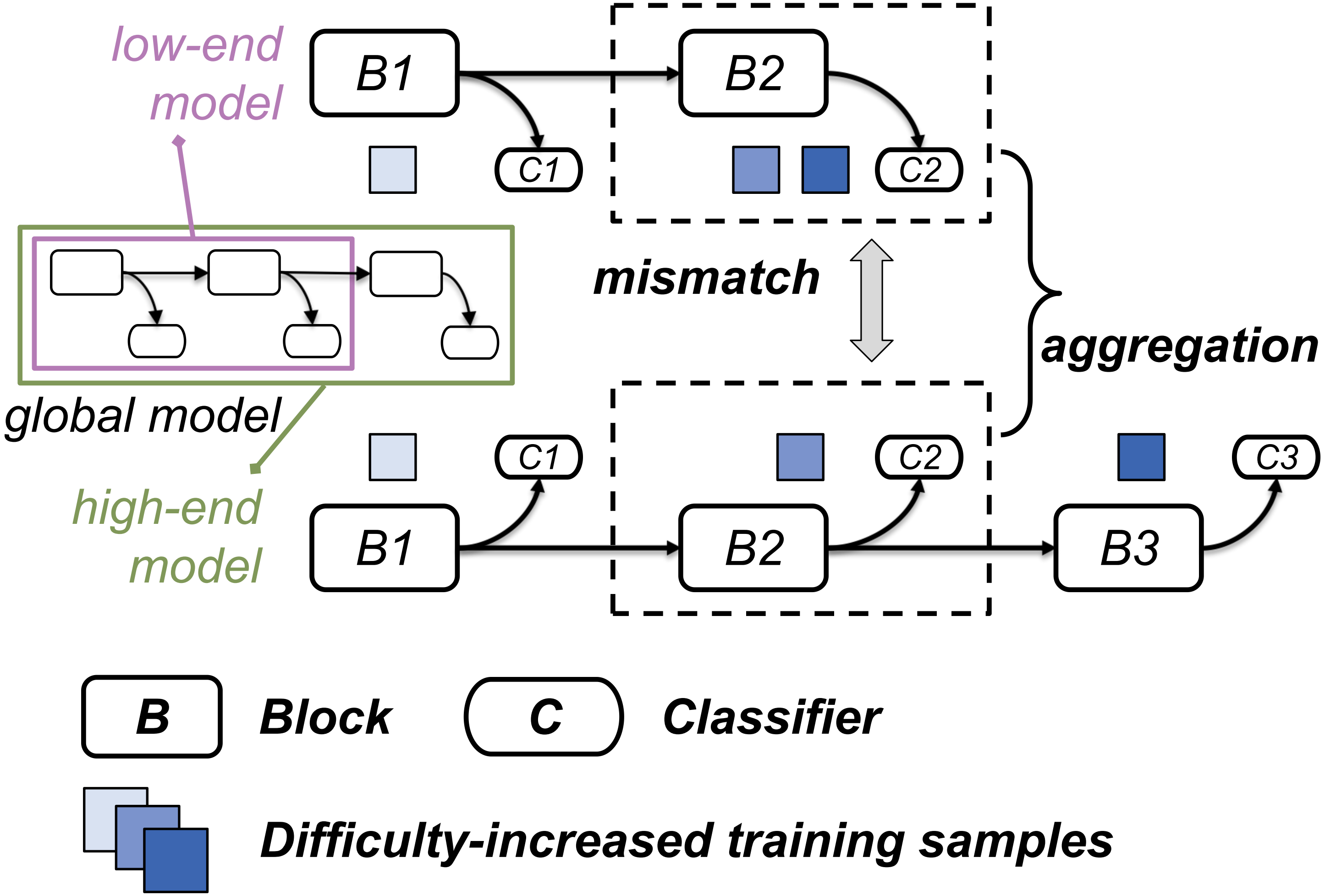


Outline

- Background & Motivation
- Problem Statement
- Our Solutions
- Experiments
- Conclusion

Problem Statement

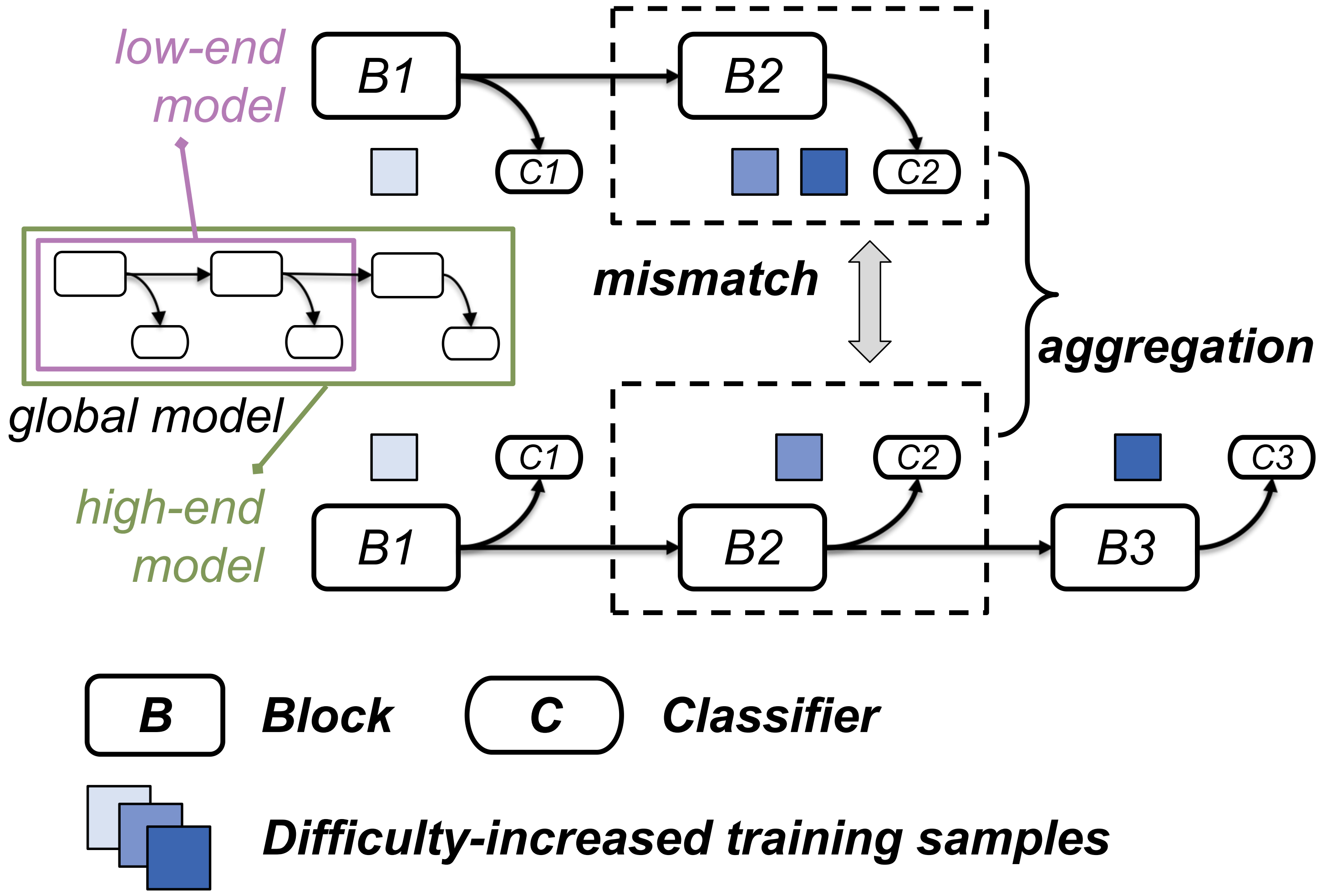
- **Federated EEN Training on Heterogeneous Devices**



Heterogeneous Resource

Problem Statement

- **Federated EEN Training on Heterogeneous Devices**



Heterogeneous Resource

Cross-model Exit Unalignment

Exits at equivalent depths may handle samples from disparate difficulty ranges across models

Challenge:
How to solve the Unalignment?

Outline

- Background & Motivation
- Problem Statement
- Our Solutions
- Experiments
- Conclusion

Our Solutions

- **DarkDistill: Framework**
  - **Progressive Knowledge Distillation**

Generator

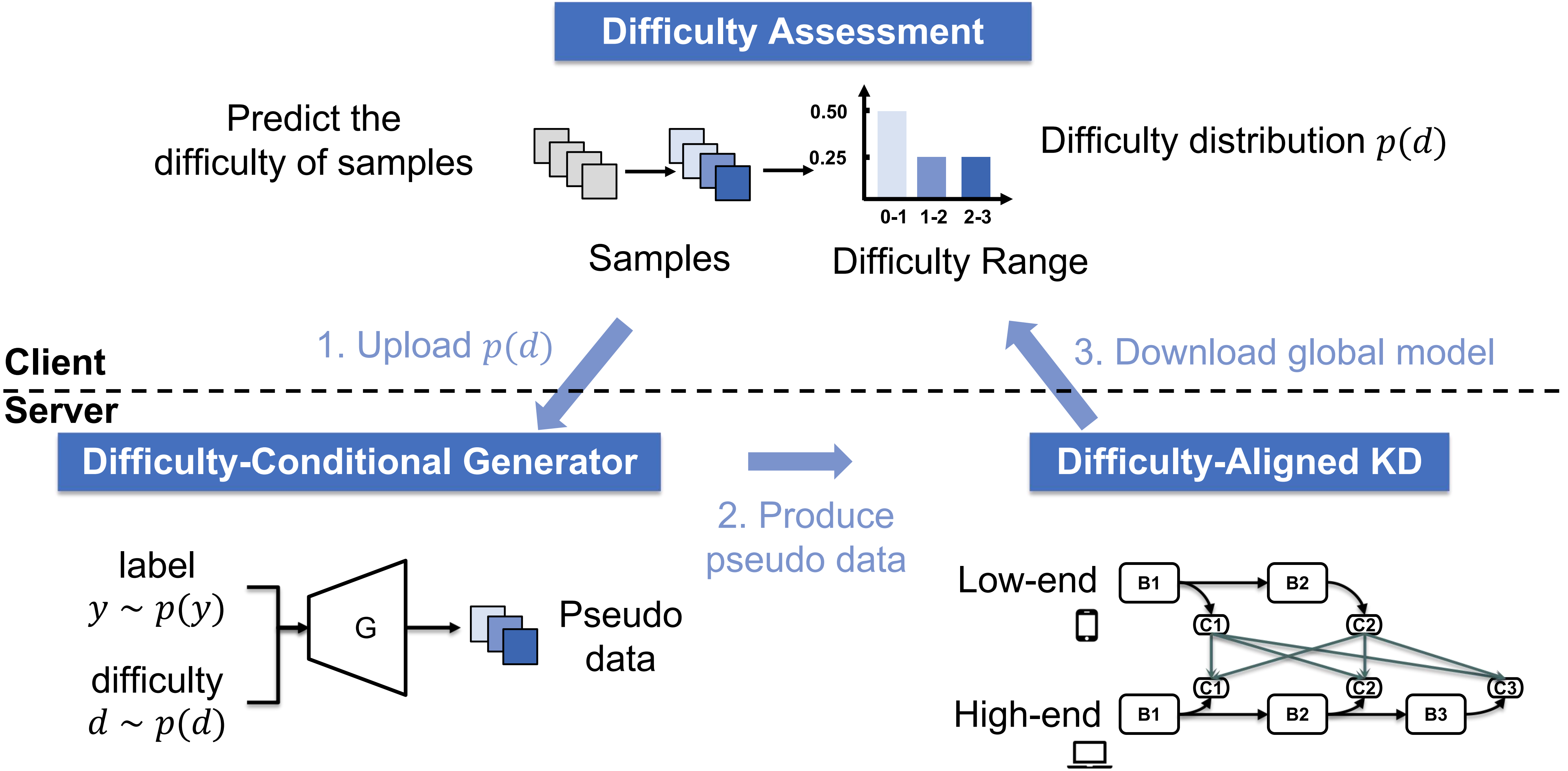
Create pseudo-data for specific difficulties, supporting the knowledge distillation process

Progressive KD

Transfer knowledge from shallow to deep exits in adjacent layers across varied depth local EENs

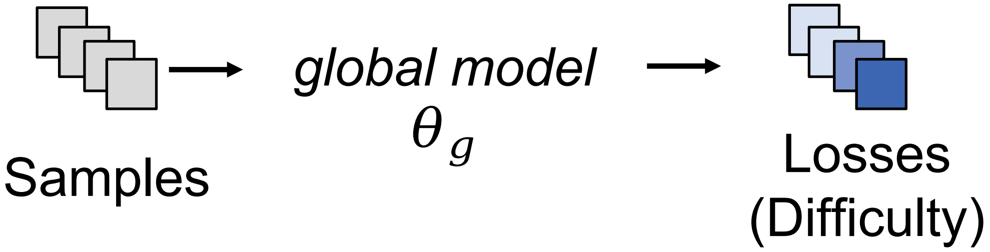
Our Solutions

- **DarkDistill: Workflow**



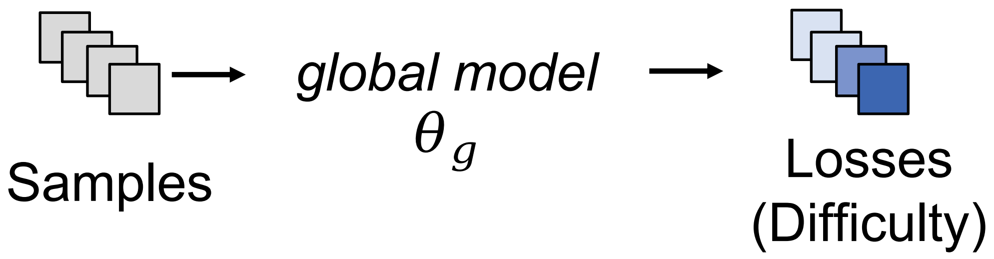
Our Solutions

- **DarkDistill: Difficulty Assessment**
  - Inspired by Curriculum Learning, we utilize the loss of sample to respect its difficulty. *The bigger the loss, the harder it is.*
  - In order to uniformly measure difficulty across clients, we leverage the global model to calculate the loss.

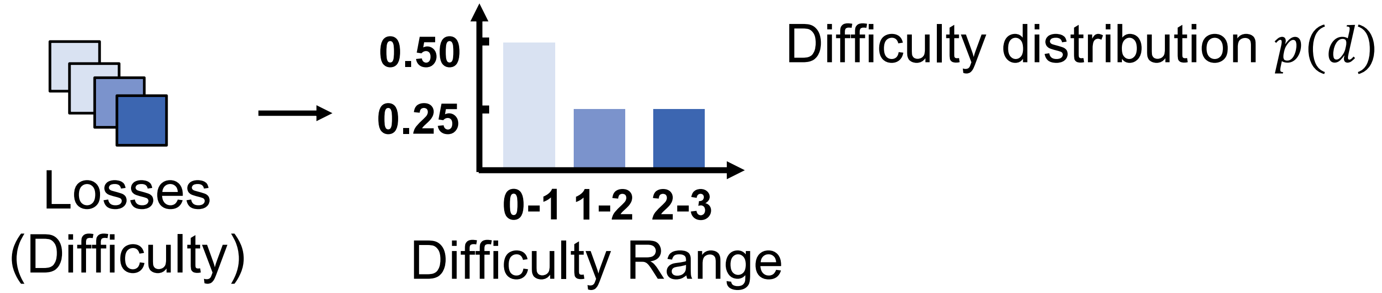


Our Solutions

- **DarkDistill: Difficulty Assessment**
  - Inspired by Curriculum Learning, we utilize the loss of sample to respect its difficulty. *The bigger the loss, the harder it is.*
  - In order to uniformly measure difficulty across clients, we leverage the global model to calculate the loss.

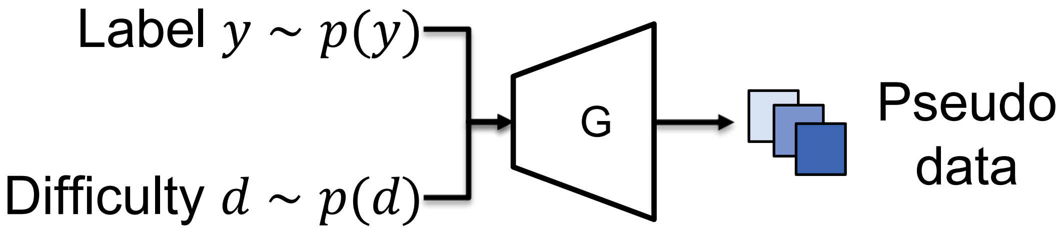


  - Calculate difficulty distribution $p(d)$ for *privacy*



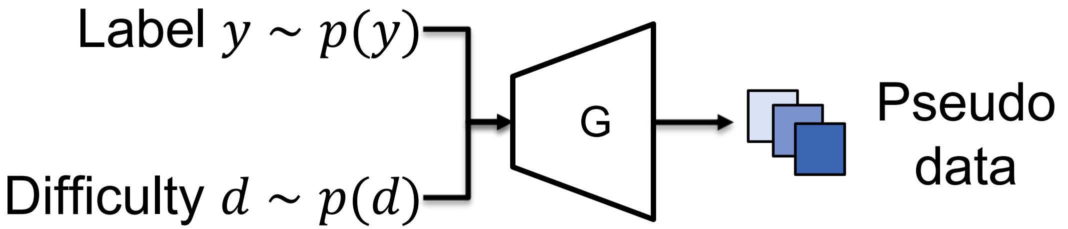
Our Solutions

- **DarkDistill: Difficulty-Conditional Generator**
  - *Create pseudo-data for specific difficulty and label to simulate local datasets, supporting the knowledge distillation process*



Our Solutions

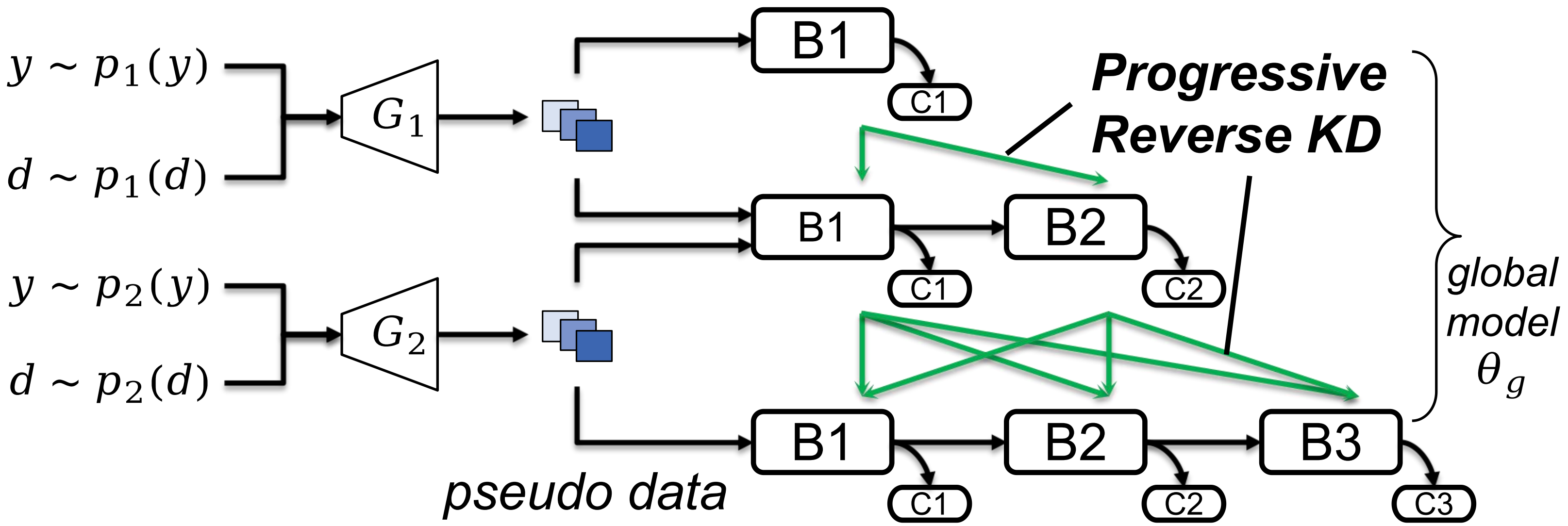
- **DarkDistill: Difficulty-Conditional Generator**
  - *Create pseudo-data for specific difficulty and label to simulate local datasets, supporting the knowledge distillation process*



  - Training objectives
    - Classification: $\mathcal{L}_{ce}(\phi_m, \theta_m) = \mathbb{E}_{\tilde{x} \sim G_m(y,d,\epsilon;\, \phi_m)} \sum_{i=1}^{m} \mathrm{CE}(\hat{y}, y)$
      - $\hat{y}$ is the predicted label for pseudo data $\tilde{x}$, minimizing m exits loss.
    - Difficulty Simulation: $\mathcal{L}_{dif}(\phi_m, \theta_m) = \mathbb{E}_{\tilde{x} \sim G_m(y,d,\epsilon;\, \phi_m)} \left| d - \widehat{d} \right|$
      - Give difficulty $d \sim p(d)$, $\widehat{d}$ is the predicted difficulty, minimizing the $|d - \widehat{d}|$

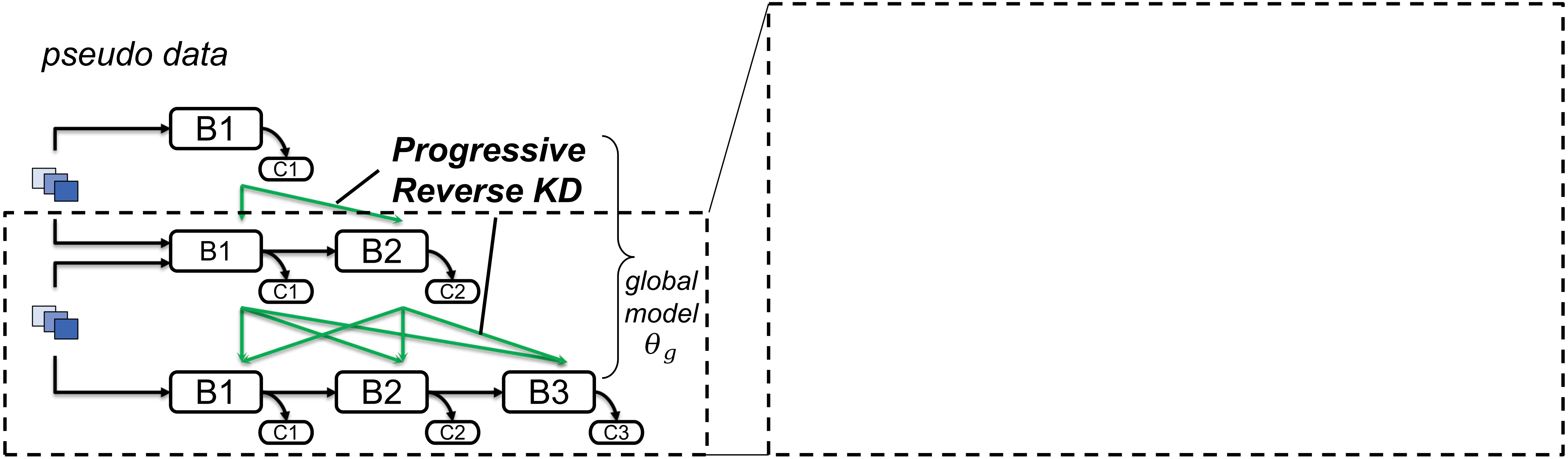
Our Solutions

- **DarkDistill: Difficulty-Aligned Reverse KD**
  - *Model-wise: Transfer knowledge from shallow to deep exits in adjacent layers across varied depth local EENs*



Our Solutions

- **DarkDistill: Difficulty-Aligned Reverse KD**
  - *Exit-wise: adaptive KD based on difficulty distance between exits across adjacent EENs*



Model-wise KD ***Exit-wise KD***

Our Solutions

- **DarkDistill-PL: Framework**
  - **Parallel Variant for DarkDistill**

Difficulty-Increased Generator

Generate pseudo data with increasing difficulty to simulate the difficulty range across various depth exits

Parallel KD

Directly transfer the ensemble knowledge in same depth exits across intermediate models into global model

Outline

- Background & Motivation
- Problem Statement
- Our Solutions
- Experiments
- Conclusion

Experiments

- **Configuration**

  - Dataset: CIFAR100，SVHN，SpeechCommands
  - Settings：100 clients, divided into 4 levels with increasing compute capabilities (4 sizes of local model)
  - Base Model：Deit-tiny (Transformer, 12 layers)
  - Exit distribution： add exits at 3th, 6th, 9th, 12th layer
  - Finetune methods: Full parameters, LoRA
  - Total Epoch： 500

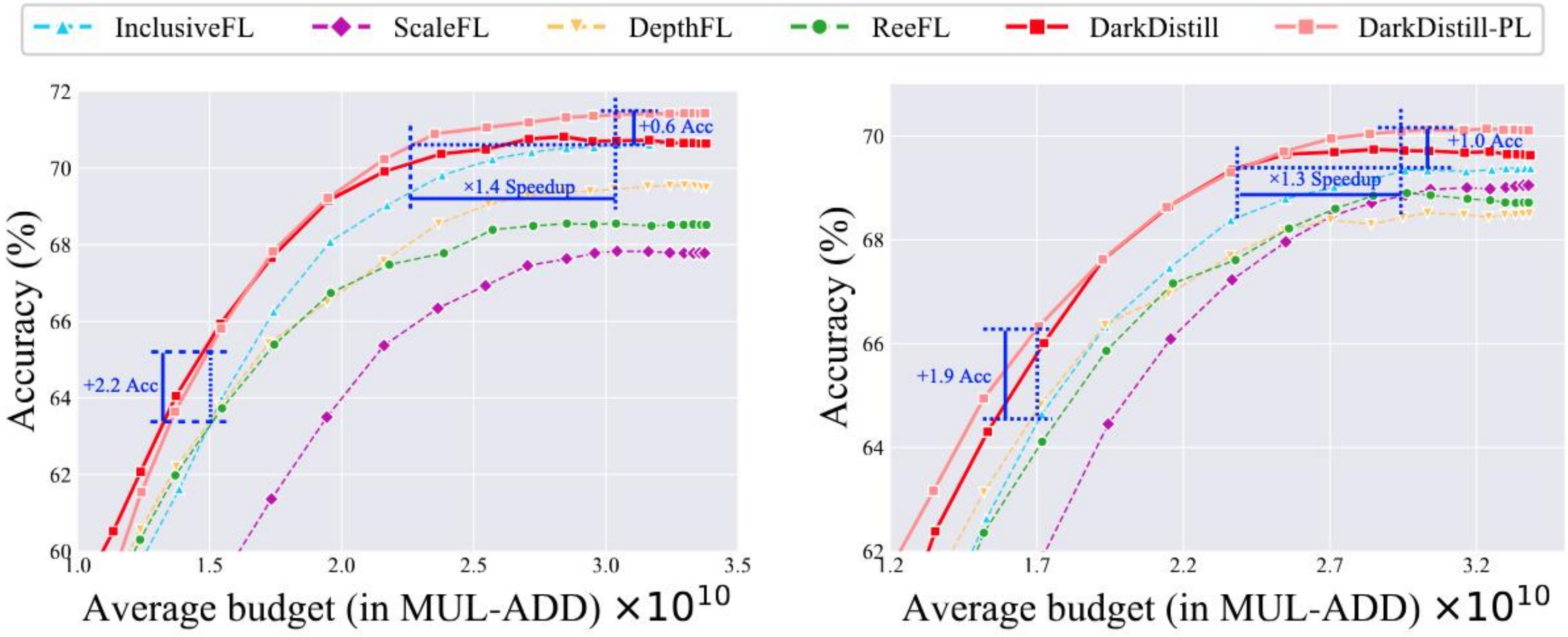
Experiments-Main Results

- **Performance of Anytime Inference**
  - Measures the accuracy of each exit assuming sufficient budgets
  - DarkDistill and DarkDistill-PL with BoostNet are the top 2 on all datasets, and increase 2 percents in general

Finetune	Difficulty-aware	Approach	CIFAR-100 [19] $\alpha = 0.1$	CIFAR-100 [19] $\alpha = 1$	CIFAR-100 [19] $\alpha = 1000$	SVHN [30]	SpeechCmds [44]
Full	None	ExclusiveFL	$26.60_{\pm3.10}$	$49.96_{\pm11.48}$	$41.58_{\pm7.01}$	$85.28_{\pm2.97}$	$87.00_{\pm2.88}$
		InclusiveFL [26]	$40.10_{\pm2.03}$	$58.83_{\pm6.98}$	$61.40_{\pm7.01}$	$82.95_{\pm0.34}$	$91.90_{\pm1.42}$
		ScaleFL [16]	$54.99_{\pm10.61}$	$63.21_{\pm9.14}$	$63.82_{\pm9.87}$	$88.24_{\pm0.78}$	$92.56_{\pm0.26}$
		DepthFL [18]	$40.70_{\pm1.57}$	$59.01_{\pm5.18}$	$61.71_{\pm5.73}$	$83.45_{\pm0.43}$	$92.05_{\pm0.60}$
		ReeFL [23]	$59.24_{\pm8.00}$	$63.37_{\pm7.72}$	$63.90_{\pm8.68}$	$88.37_{\pm1.27}$	$93.12_{\pm1.14}$
	BoostNet [45]	ExclusiveFL	$48.68_{\pm13.66}$	$57.57_{\pm15.12}$	$58.65_{\pm15.31}$	$87.30_{\pm2.89}$	$91.07_{\pm2.58}$
		InclusiveFL [26]	$57.10_{\pm7.21}$	$62.96_{\pm8.12}$	$64.01_{\pm8.24}$	$87.86_{\pm1.66}$	$92.91_{\pm1.10}$
		ScaleFL [16]	$52.74_{\pm13.82}$	$60.55_{\pm11.93}$	$60.73_{\pm10.80}$	$87.91_{\pm0.77}$	$92.03_{\pm0.37}$
		DepthFL [18]	$58.15_{\pm6.73}$	$63.81_{\pm6.34}$	$64.19_{\pm6.73}$	$87.74_{\pm1.01}$	$92.72_{\pm0.64}$
		ReeFL [23]	$59.01_{\pm7.98}$	$63.08_{\pm9.03}$	$63.66_{\pm7.31}$	$88.39_{\pm1.28}$	$93.01_{\pm1.18}$
		DarkDistill	$\underline{60.48}_{\pm7.93}$	$\underline{64.50}_{\pm7.97}$	$\mathbf{65.67}_{\pm7.48}$	$\underline{88.41}_{\pm1.46}$	$\underline{93.31}_{\pm1.13}$
		DarkDistill-PL	$\mathbf{61.05}_{\pm8.19}$	$\mathbf{65.12}_{\pm7.02}$	$\underline{65.49}_{\pm7.88}$	$\mathbf{88.48}_{\pm1.57}$	$\mathbf{93.42}_{\pm0.98}$

Experiments-Main Results

- **Performance of Budget Inference**
  - Measures the accuracy of a batch samples within given budgets
  - DarkDistill and DarkDistill-PL can improve the accuracy over the baselines at various computation budgets.



(a) **Full** $\alpha = 1000$ (b) **Full** $\alpha = 1$

Experiments-Module Ablation

- **Difficulty Assessment Module**
  - Verify the efficiency of difficulty assessment module
    - Left images are easier predicted by module
    - Right image are more difficult predicted by module

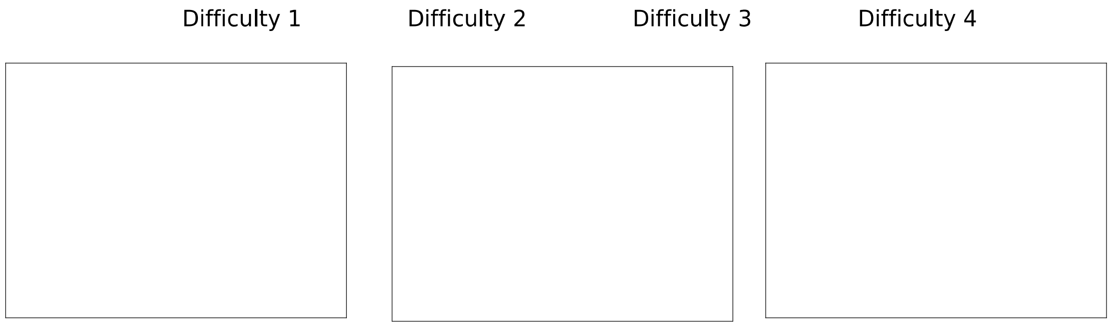
Single pigment, easy to judge

Complex contour, difficult to judge

Consistent with human intuition, the difficulty assessment module is designed reasonably

Experiments-Module Ablation

- **Difficulty-Conditional Generator**
  - Pseudo data of the same category are divided into four different levels of clustering clusters



Fox **Mouse** **Wolf**

Experiments-Module Ablation

- **Difficulty-Conditional Generator**
  - Pseudo data of the same category are divided into four different levels of clustering clusters

Difficulty 1 Difficulty 2 Difficulty 3 Difficulty 4

Fox **Mouse** **Wolf**

  - The robustness of generator architecture

d_ϵ	d_h 64	128	256	512
2	64.79|64.88	65.05|64.93	65.32|65.25	65.10|65.11
16	65.38|64.91	65.20|65.09	65.37|64.65	65.18|65.51
32	65.05|64.79	65.06|64.79	65.28|65.08	65.60|64.90
64	65.15|64.92	65.74|64.93	64.91|64.55	65.06|65.05
SOTA	$64.19_{\pm 6.73}$			

Outline

- Background & Motivation
- Problem Statement
- Our Solutions
- Experiments
- Conclusion

Conclusion

- *This paper introduces DarkDistill, a novel heterogeneous federated learning scheme dedicated for early-exit networks (EENs) and its parallel variant DarkDistill-PL for acceleration.*

- *We identify the cross-model exit unalignment problem, an unexplored challenge when extending difficulty-aware EEN training to federated contexts.*

- *We develop a difficulty-conditional generator training strategy and a difficulty-aligned reverse distillation scheme to aggregate EENs of varying depths into a global model that retains its difficulty-specific specialization.*

THANK YOU

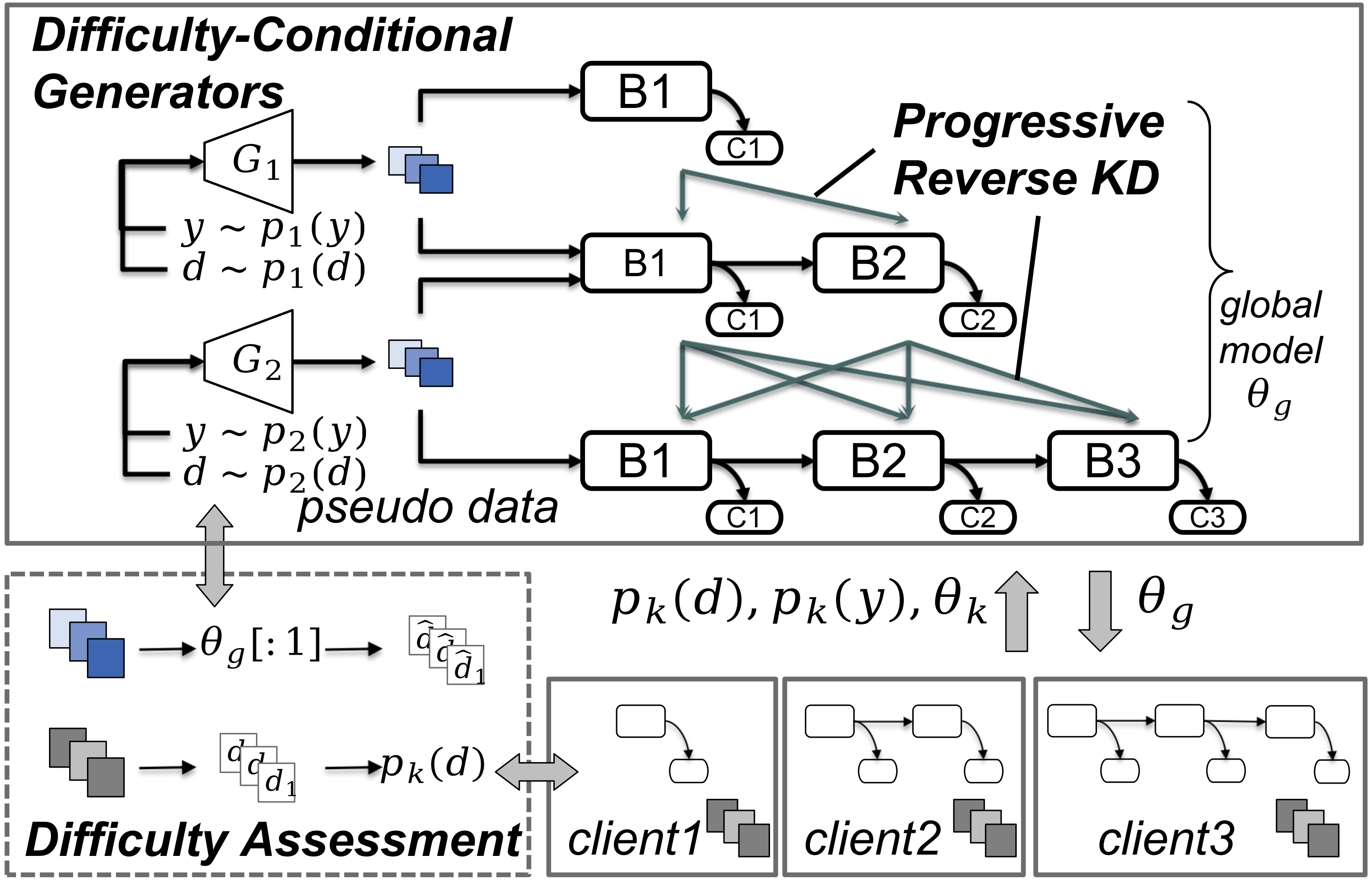
if you have problems, feel free to email

lehaoqv@buaa.edu.cn

or talk with me at Poster 201, 5th August

Our Solutions

- **DarkDistill: Framework**



Progressive Reverse KD

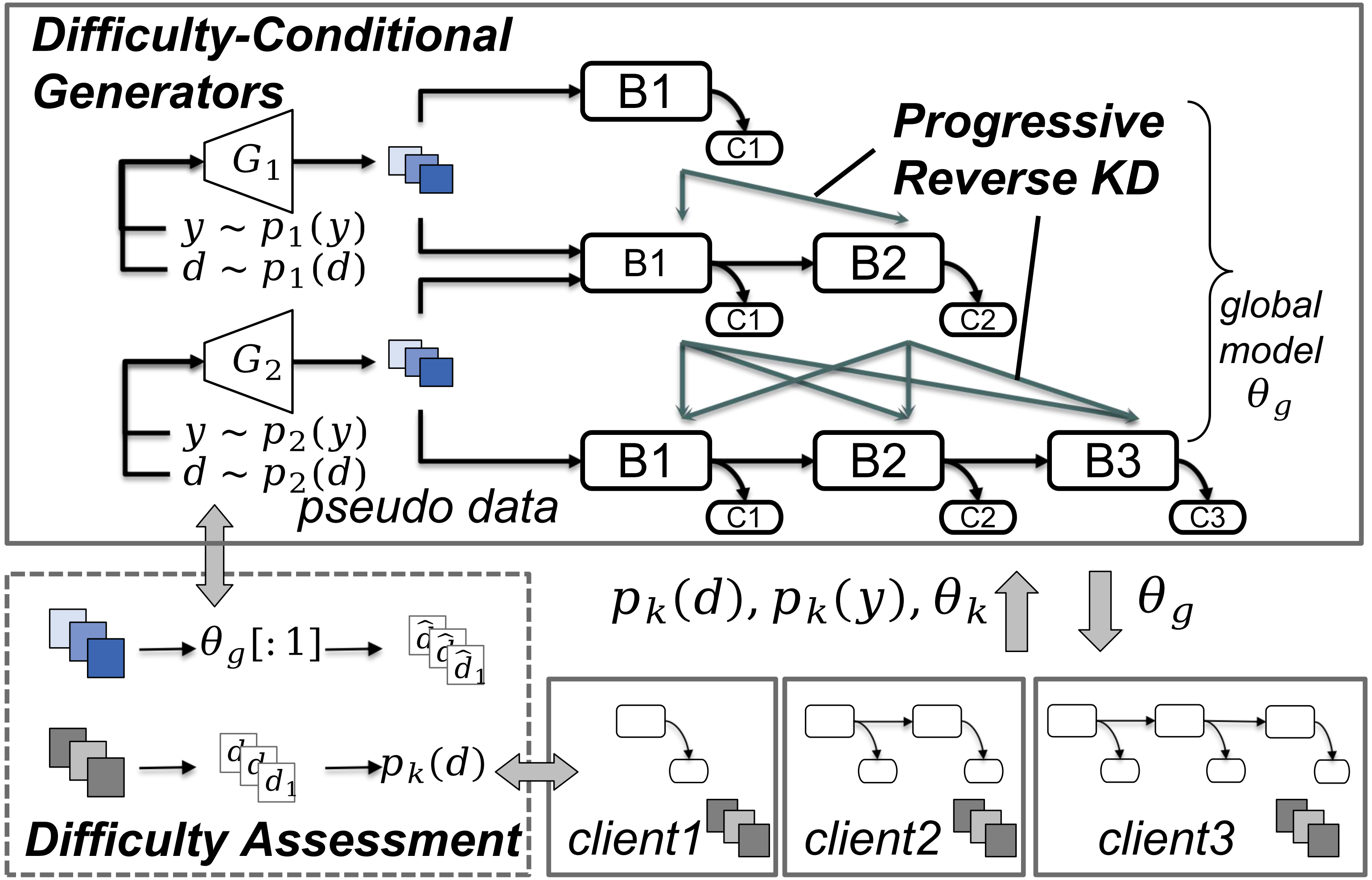
transfer knowledge from shallow to deep exits in adjacent layers across varied depth local EENs

Difficulty-Conditonal Generators

create pseudo-data for specific difficulties, supporting the knowledge distillation process

Our Solutions

- **DarkDistill: Difficulty-Aligined Reverse Knowledge Distillation**



Progressive Reverse KD

transfer knowledge from shallow to deep exits in adjacent layers across varied depth local EENs