Consiglio Nazionale delle Ricerche
Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
AIMH Artificial Intelligence for Media and Humanities
HPC lab
Università di Pisa
University of St Andrews

Experimenting with Information Dissimilarity for Knowledge Distillation at the Edge

Joaquim Mbasa Molo[1,2], Lucia Vadicamo[1], Emanuele Carlini[1], Claudio Gennaro[1], Richard Connor[3]

[1]*Institute of Information Science and Technologies, National Research Council (CNR-ISTI), Pisa, Italy*
[2]*Department of Computer Science, University of Pisa, Pisa, Italy*
[3]*School of Computer Science, University of St Andrews, St Andrews, Scotland*

joaquimbasa/Distributed_KD_Information_Dissimilarity

joaquim.molo@phd.unipi.it lucia.vadicamo@isti.cnr.it emanuele.carlini@isti.cnr.it

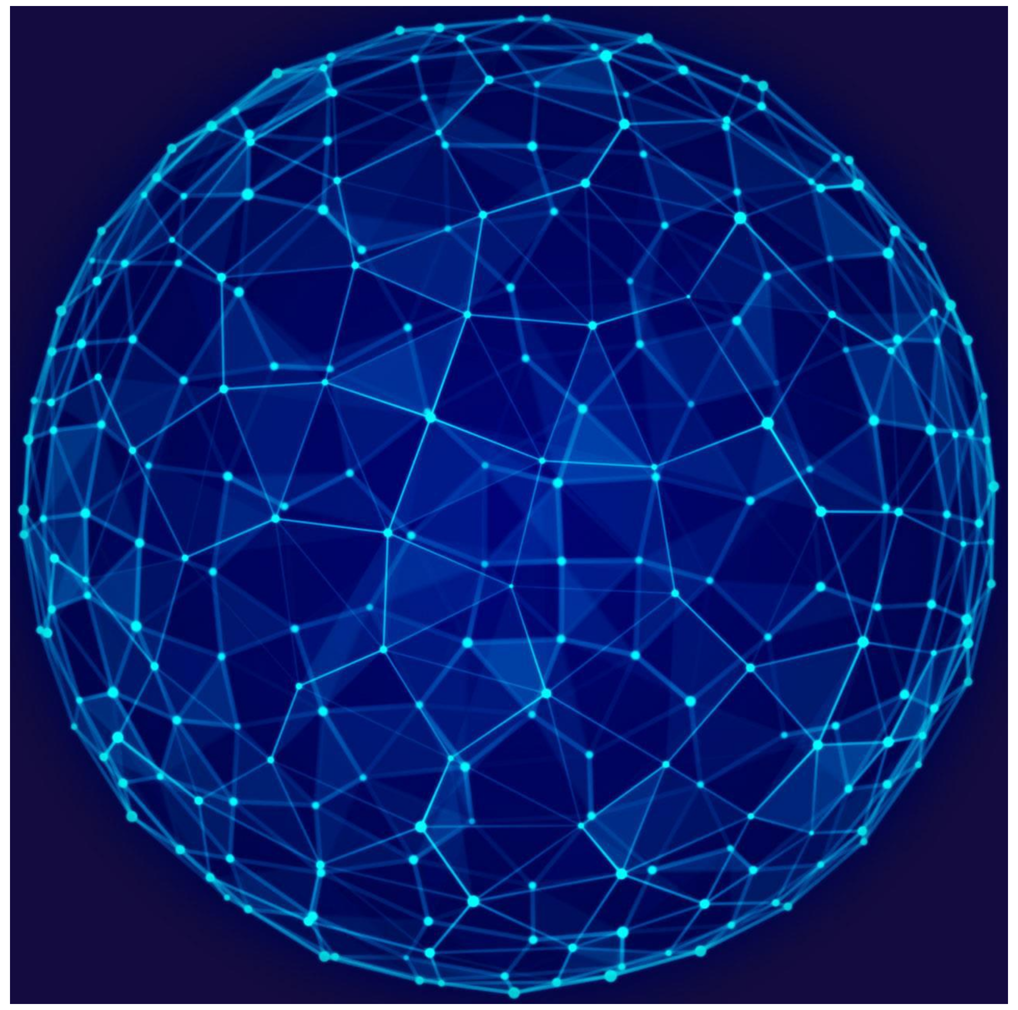
slides from lucia.vadicamo@isti.cnr.it

EDGELESS
HiPEAC

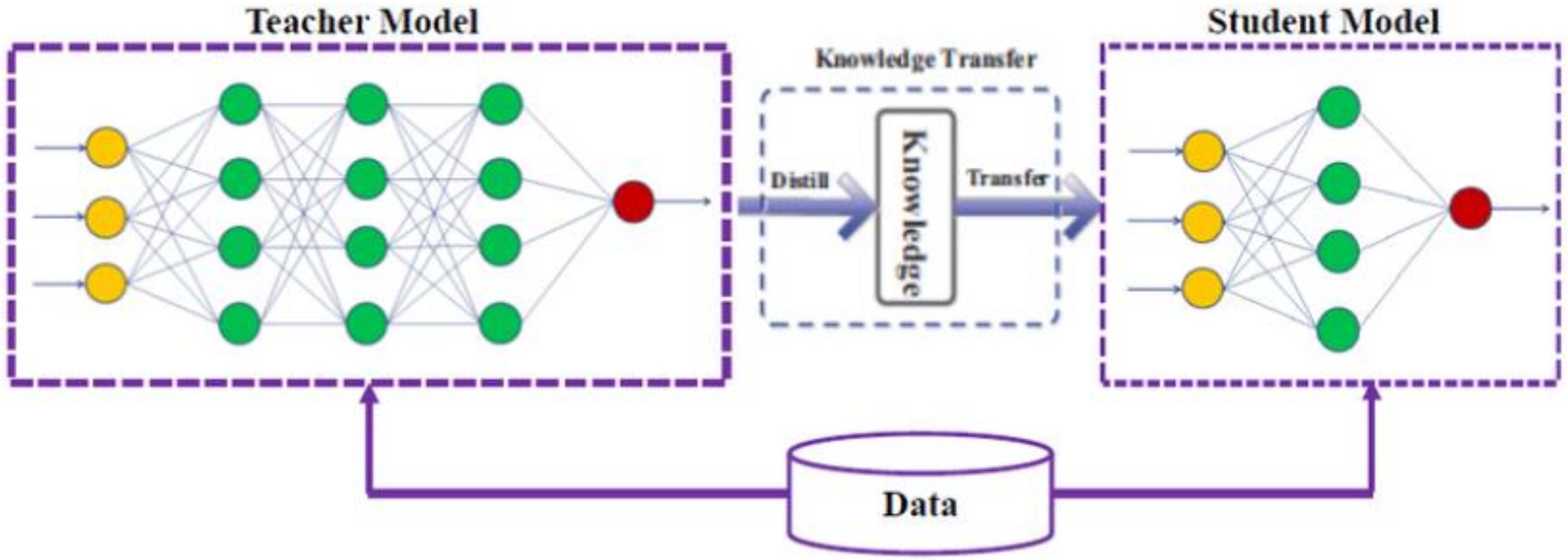
Edge Intelligence and Decentralized DNN

- Integration of AI into **edge devices**, enabling computation closer to data sources
- **Collaborative learning mechanism** composed of software agents, robots, sensors, and computer systems that can collaborate effectively

- Computation and decision-making is **distributed** across multiple nodes or devices in a network (no central node)
- Nodes can **cooperate** for DNN training or inference
- Advantages in
  - Scalability
  - Data privacy
  - Robustness



Knowledge Distillation (KD)



- KD is a machine learning technique designed to transfer knowledge from a *large, complex model* (**Teacher model**) to a *smaller, more efficient one* (**Student model**)
- The Student model **learns to mimic** the behavior of the Teacher (i.e., its outputs or internal representations)
- Key benefits:
  - Model Compression
  - Faster inference
  - Improved generalization

KD - Information Exchange Mechanism

The student model is **trained** using two types of losses:

- **Fully-supervised loss** ($\mathcal{L}_{stu}$)
  - Encourage the *student's "hard" prediction* to align closely with the *ground-truth labels* of the input samples
  - Uses **original training data**
- **Distillation loss** ($\mathcal{L}_{\mathrm{KD}}$)
  - Encourages the student's *output probabilities/representations* to align closely with those of the teacher
  - Uses **Teacher model predictions or intermediate representations** (e.g., *logits*)

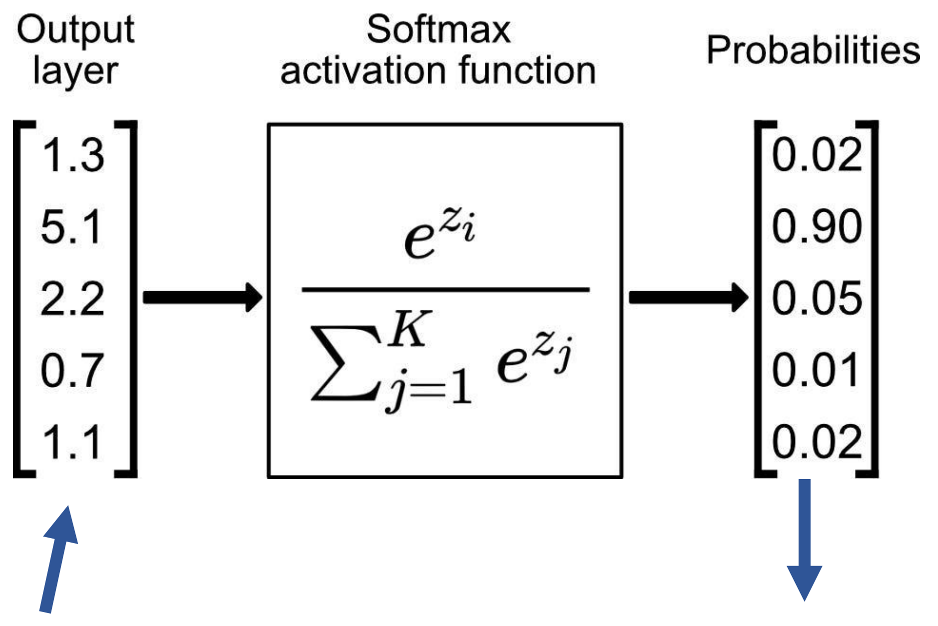
$$\mathcal{L} = \alpha\,\mathcal{L}_{stu} + (1-\alpha)\mathcal{L}_{KD}$$

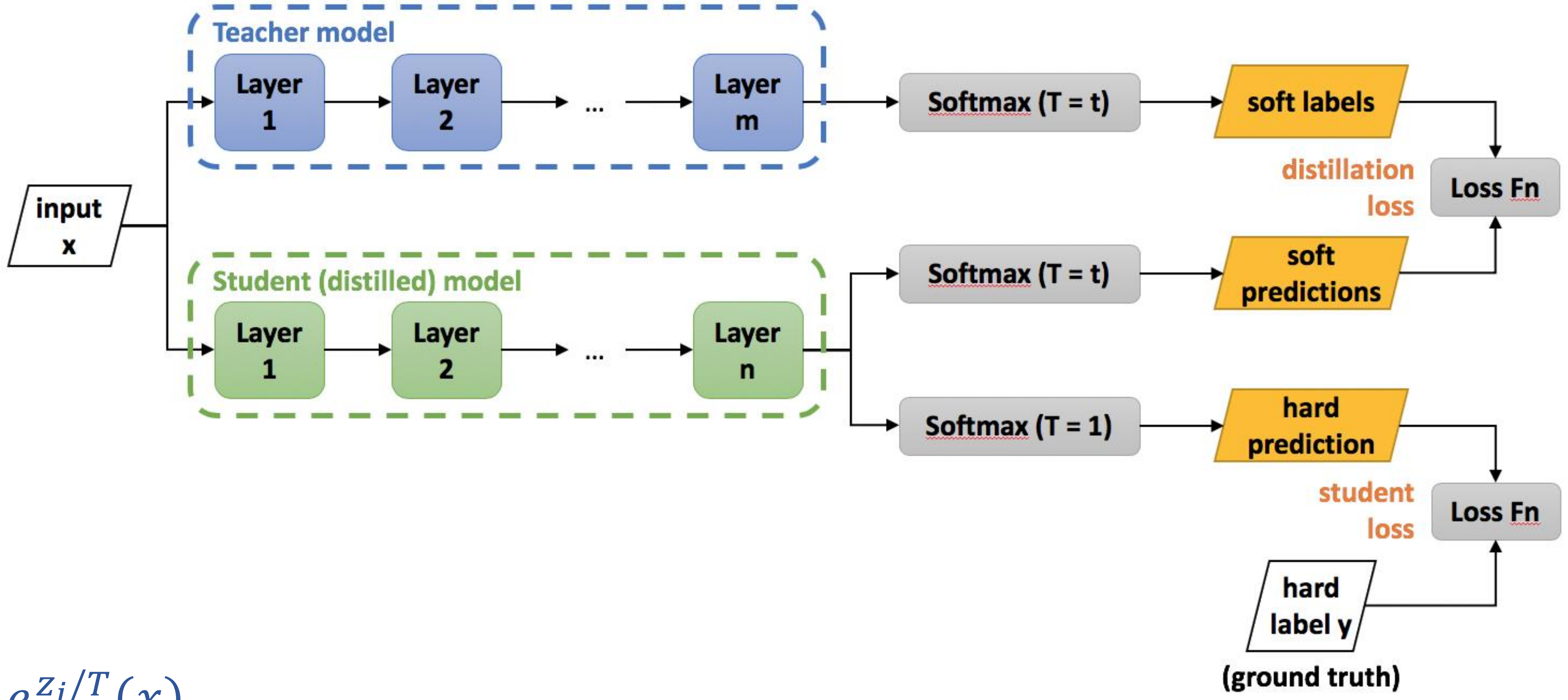
$$\mathcal{L}_{stu} = \boldsymbol{CE}\big(\boldsymbol{y}: \boldsymbol{p_S}(x, T=1)\big)$$

y: "hard" targets from ground-truth
p_S: hard prediction of the student

$$\mathcal{L}_{KD} = \boldsymbol{CE}\big(\boldsymbol{p_{\mathcal{T}}}(x, T=t): \boldsymbol{p_S}(x, T=t)\big)$$

$p_{\mathcal{T}}$: soft prediction of the teacher
p_S : soft prediction of the student





Logits: z_i

Given an input data x, trained neural networks produce peaky probability which are less informative. So a **Temperature scaling** is used → $p = \frac{e^{z_i/T}(x)}{\sum_j e^{z_j/T}(x)}$ $\quad T \equiv temperature$

Information distances

*A wide range of **information distance functions** remains underexplored in distributed learning literature*

Information distance measures the dissimilarity between two sources of information

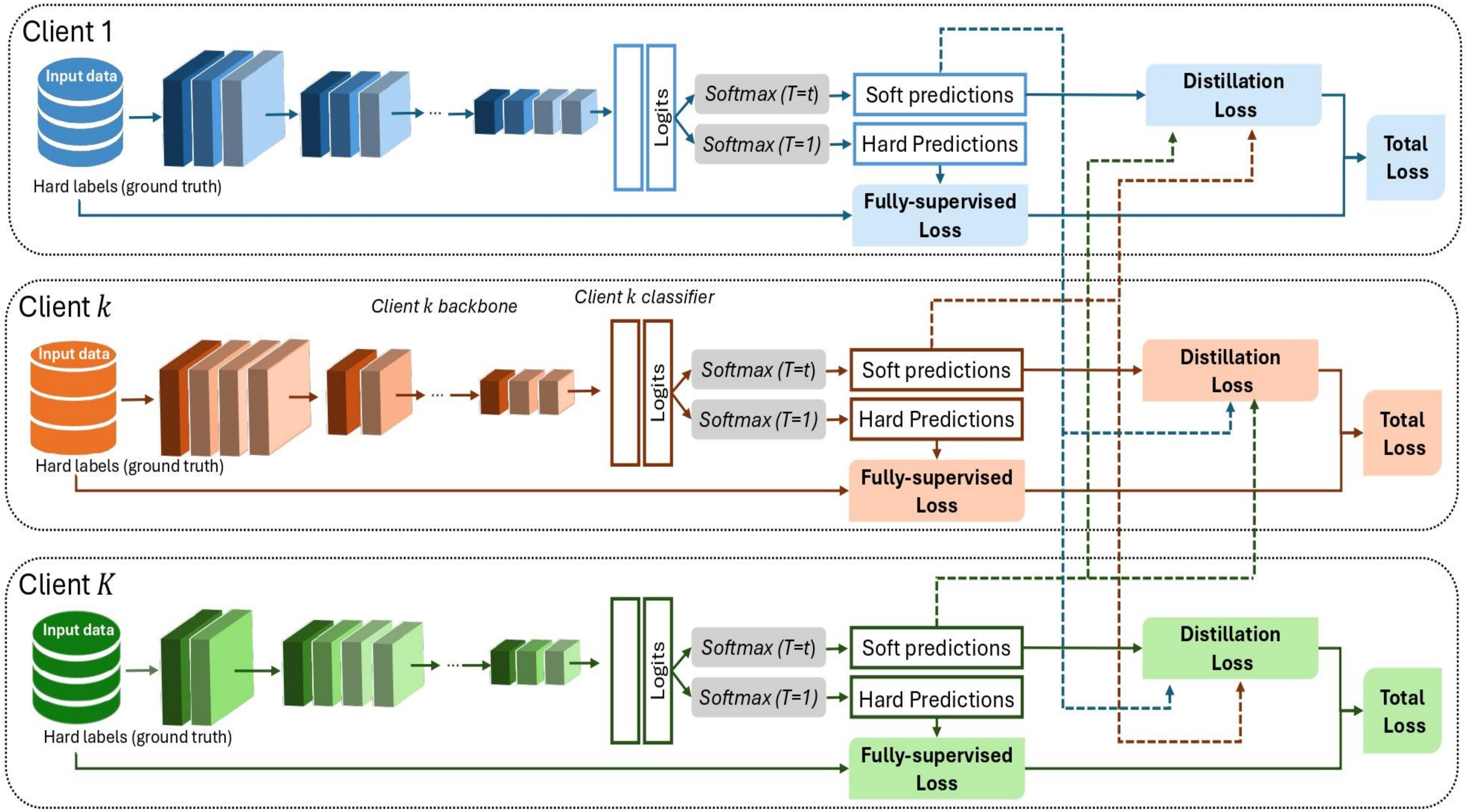
- Cross Entropy: $CE(\boldsymbol{q}:\boldsymbol{p}) = -\sum_{i=1}^{N} q_i \log p_i$
- Kullback-Leibler Divergence: $KL(\boldsymbol{q}:\boldsymbol{p}) = \sum_{i}^{N} q_i \log \frac{q_i}{p_i}$
- Jensen-Shannon Divergence: $JS(\boldsymbol{q},\boldsymbol{p}) = \frac{1}{2}\left(KL\left(\boldsymbol{q}:\frac{\boldsymbol{q}+\boldsymbol{p}}{2}\right) + KL\left(\boldsymbol{p}:\frac{\boldsymbol{q}+\boldsymbol{p}}{2}\right)\right)$
- Structural Entropic Distance: $SED(\boldsymbol{q},\boldsymbol{p}) = \frac{C\left(\frac{\boldsymbol{q}+\boldsymbol{p}}{2}\right)}{\sqrt{C(\boldsymbol{p})C(\boldsymbol{q})}}$ $\quad C(\boldsymbol{p}) = b^{-\sum_{i=1}^{N} p_i \log_b p_i}$
- Triangular Divergence: $TD(\boldsymbol{q},\boldsymbol{p}) = 1 - \sum_{i=1}^{N} \frac{2q_i p_i}{q_i+p_i}$

Note that CE, KL, JS, TD shows very tight correlations! [1]

[1] Connor, R., Dearle, A., Claydon, B., Vadicamo, L.: Correlations of cross-entropy loss in machine learning. Entropy 26(6) (2024)

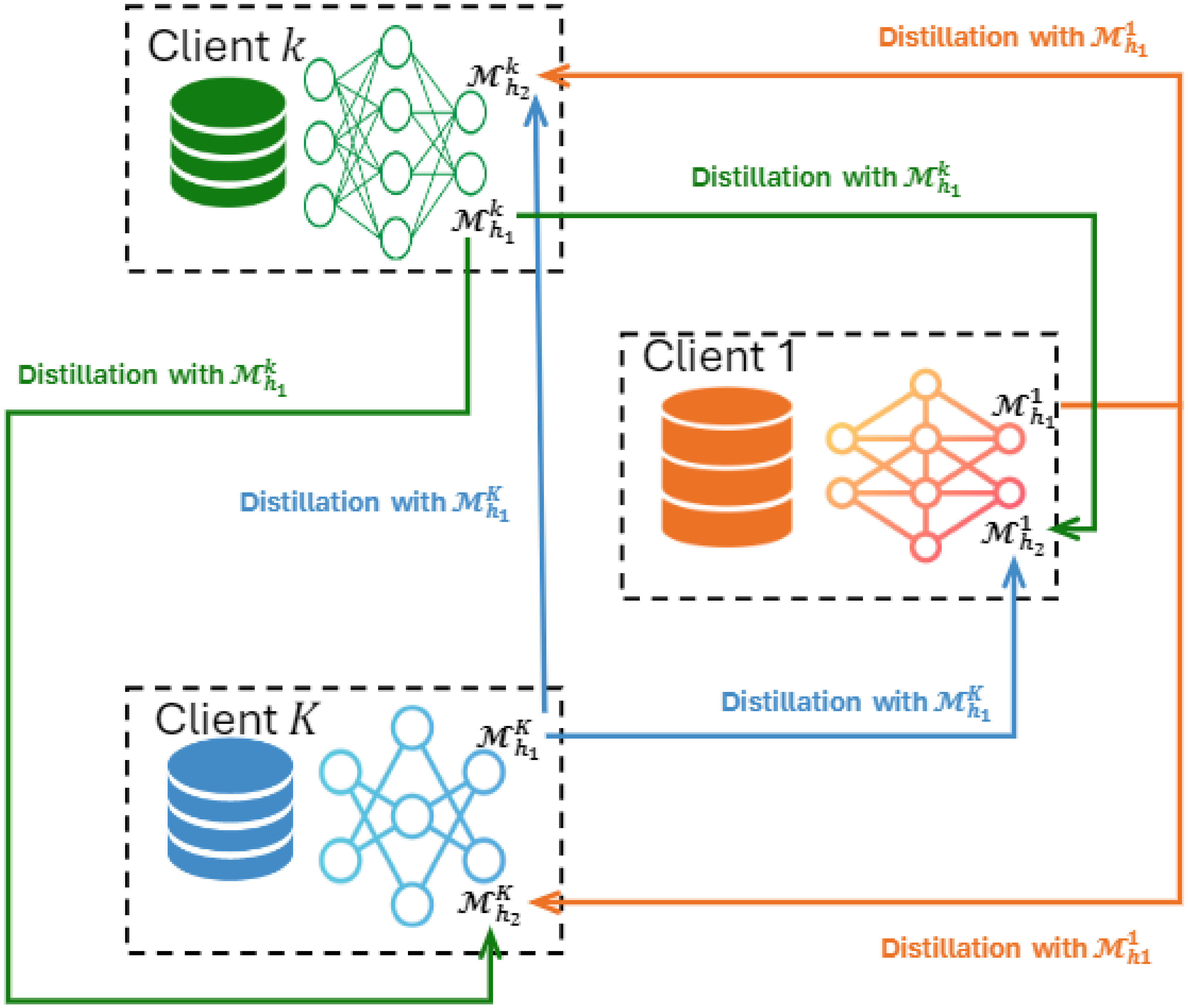
Do these different information distance exhibit similar behavior in distributed learning contexts?

KD-based Distributed Learning Framework



- Clients exchange information to enhance their learning
- Each client acts as both **learner** (**student**) and **source of knowledge (teacher)** for others
- **Decentralized** system: No central model or teacher
- Clients train on **local datasets** and share knowledge with peers

Fully Decentralized Learning Model



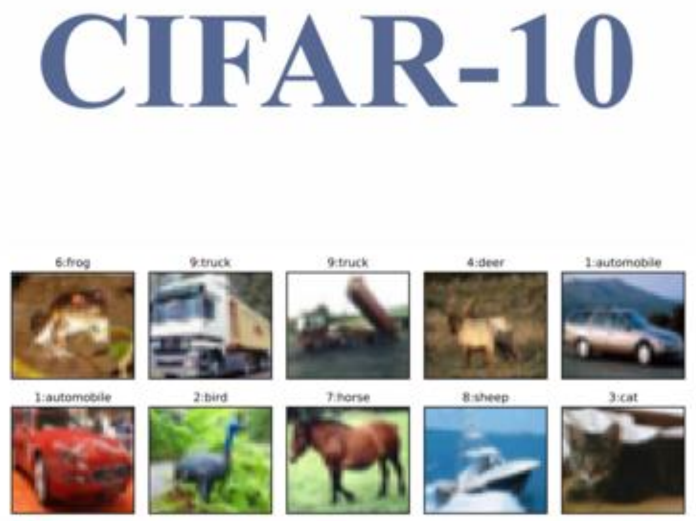
- Network with K clients
- Each client $\boldsymbol{C^k}$ holds a local dataset $\boldsymbol{D^k}$ and a multi-head neural network $\boldsymbol{\mathcal{M}^k}$, composed of :
  - **Backbone**: Extracts feature representations from input data
  - **Head 1**: Model $\boldsymbol{\mathcal{M}^k_{h1}}$(Backbone + Head 1) trained on local distribution $\boldsymbol{D^k}$
  - **Head 2**: Model $\boldsymbol{\mathcal{M}^k_{h2}}$(Backbone + Head 2) trained on $\boldsymbol{D^k}$ using *knowledge distillation* from connected clients
- Clients are trained concurrently, allowing them to share knowledge through distillation to improve overall model performance

Experimental Setup

- Decentralized network with **3** interconnected clients
- Comparing different information dissimilarity measures (CE, KL, SED, TD, JS)
- Different levels of data heterogeneity [2]
  - Each client C^k receives a subset of labels $\{\ell_i\}$, referred to as *primary labels for* C^k
  - Labels outside $\{\ell_i\}$ are considered *secondary labels* for Client C^k
  - Data samples are distributed randomly among clients. The probability of assigning a sample with label ℓ to a client C^k is chosen to be $(1 + \gamma)$ higher for clients that have ℓ as their primary label
  - $\boldsymbol{\gamma}$ controls dataset skewness:
    - $\gamma = 0$: data distribution is uniform across all clients (**iid**)
    - Higher γ: **Non-iid** distribution (more primary label focus)

- In experiments:
  - $\gamma = 15$ for CIFAR-10
  - $\gamma = 10$ for SUN397

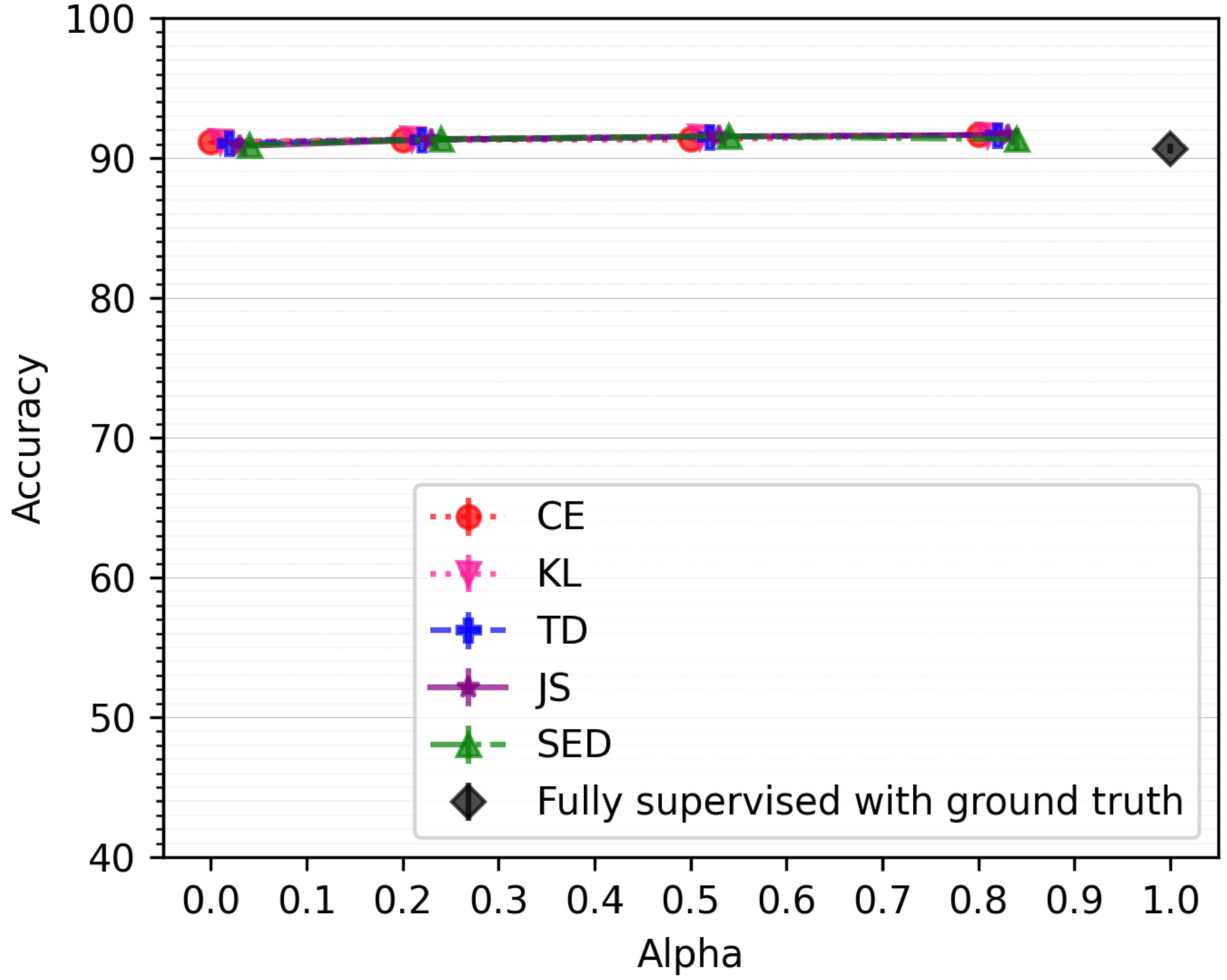
  $\Big\}$ temperature T: 1, 10 and 100



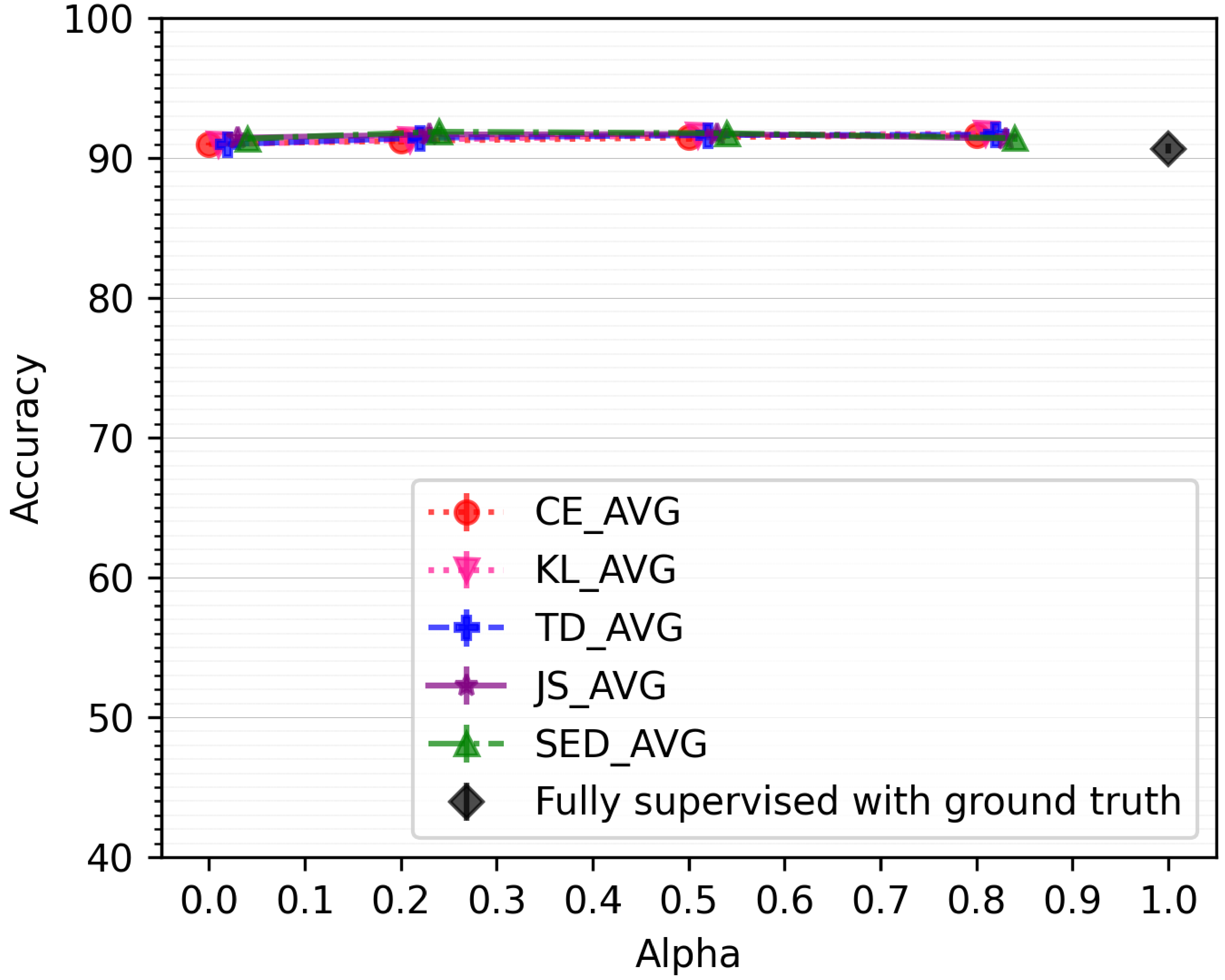
[2] Zhmoginov, Andrey, Mark Sandler, Nolan Miller, Gus Kristiansen, and Max Vladymyrov. "Decentralized Learning with Multi-Headed Distillation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8053-8063. 2023.

Results: CIFAR-10 iid

$$\mathcal{L}_{k,KD}(\mathbf{w}_2^k, x) = \sum_{\phi \in \Phi_k} \mathbb{E}_{x \sim X^k} f\left(\mathbf{p}^k(\mathbf{w}_2^k, x), \mathbf{p}^\phi(x)\right) \quad \text{(SUM)}$$



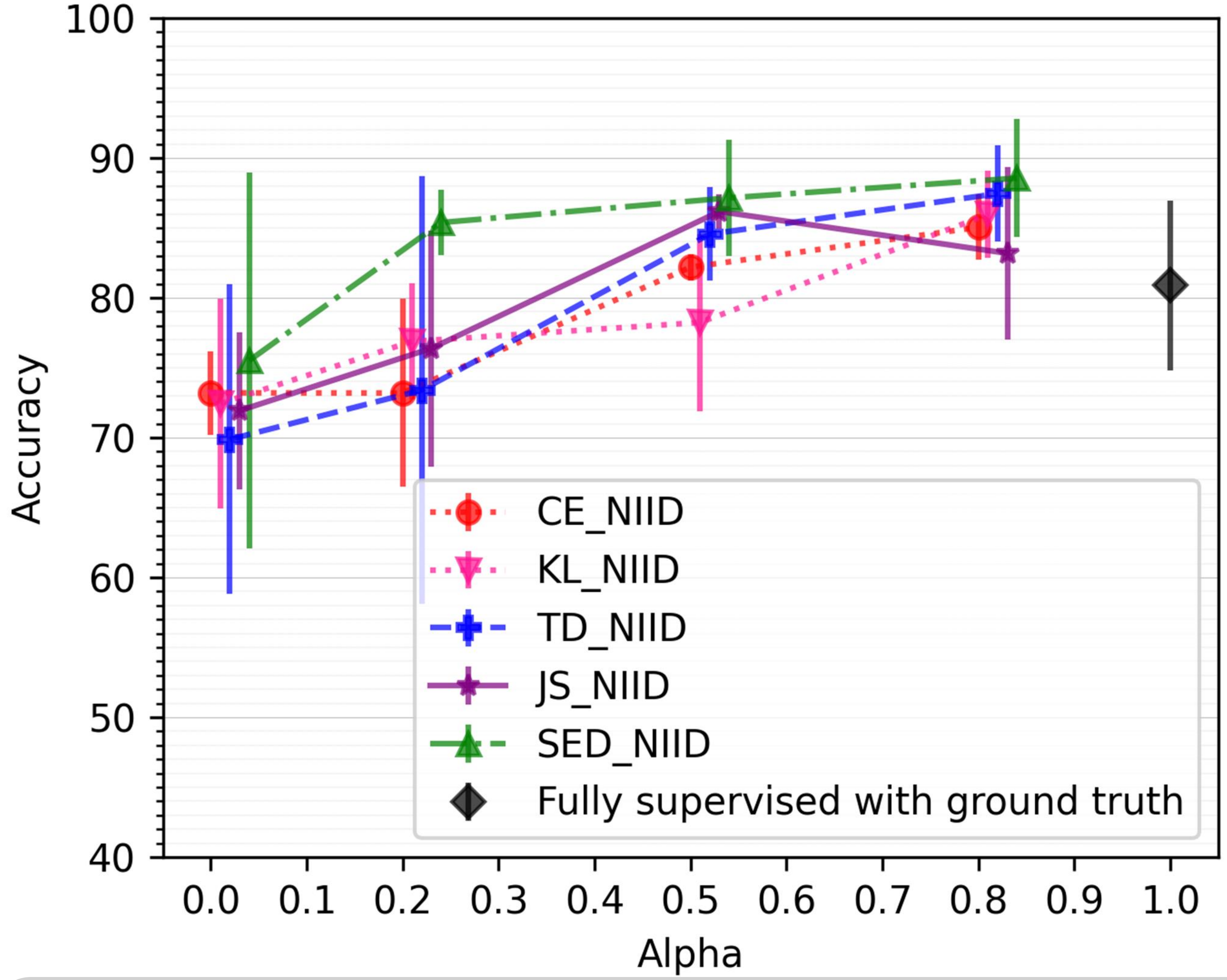
$$\mathcal{L}_{k,KD}(\mathbf{w}_2^k, x) = \mathbb{E}_{x \sim X^k} f\left(\mathbf{p}^k(\mathbf{w}_2^k, x), \frac{\sum_{\phi \in \Phi_k} \mathbf{p}^\phi(x)}{|\Phi|}\right) \quad \text{(AVG)}$$



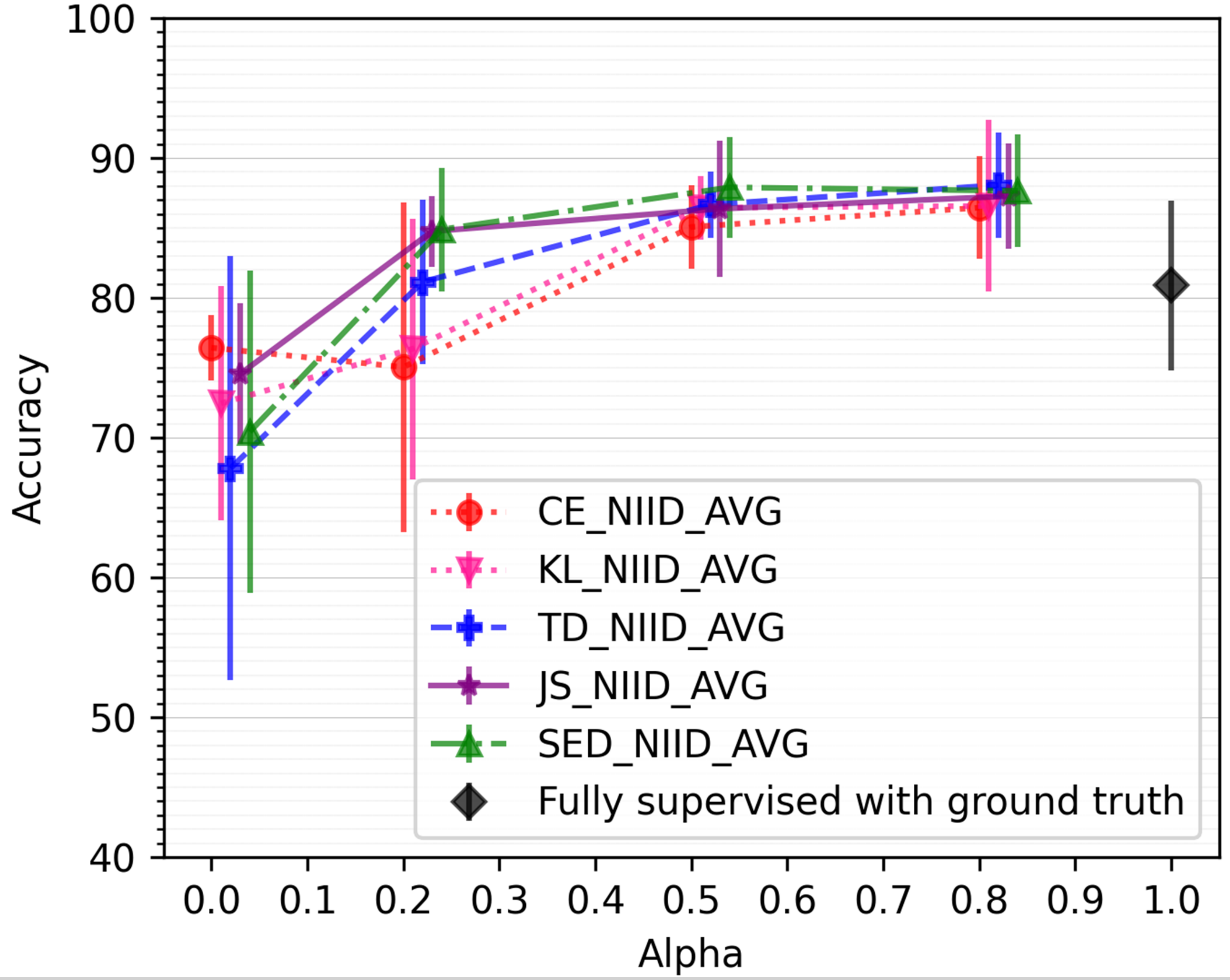
- KD does not significantly enhance overall accuracy when the input data is sufficient and balanced
  - all tested dissimilarity measures exhibited performance similar to CE
- AVG approach achieve same performance while reducing computational complexity

Results: CIFAR-10 non-iid

$$\mathcal{L}_{k,KD}(\mathbf{w}_2^k, x) = \sum_{\phi \in \Phi_k} \mathbb{E}_{x \sim X^k} f\left(\mathbf{p}^k(\mathbf{w}_2^k, x), \mathbf{p}^{\phi}(x)\right) \quad \text{(SUM)}$$



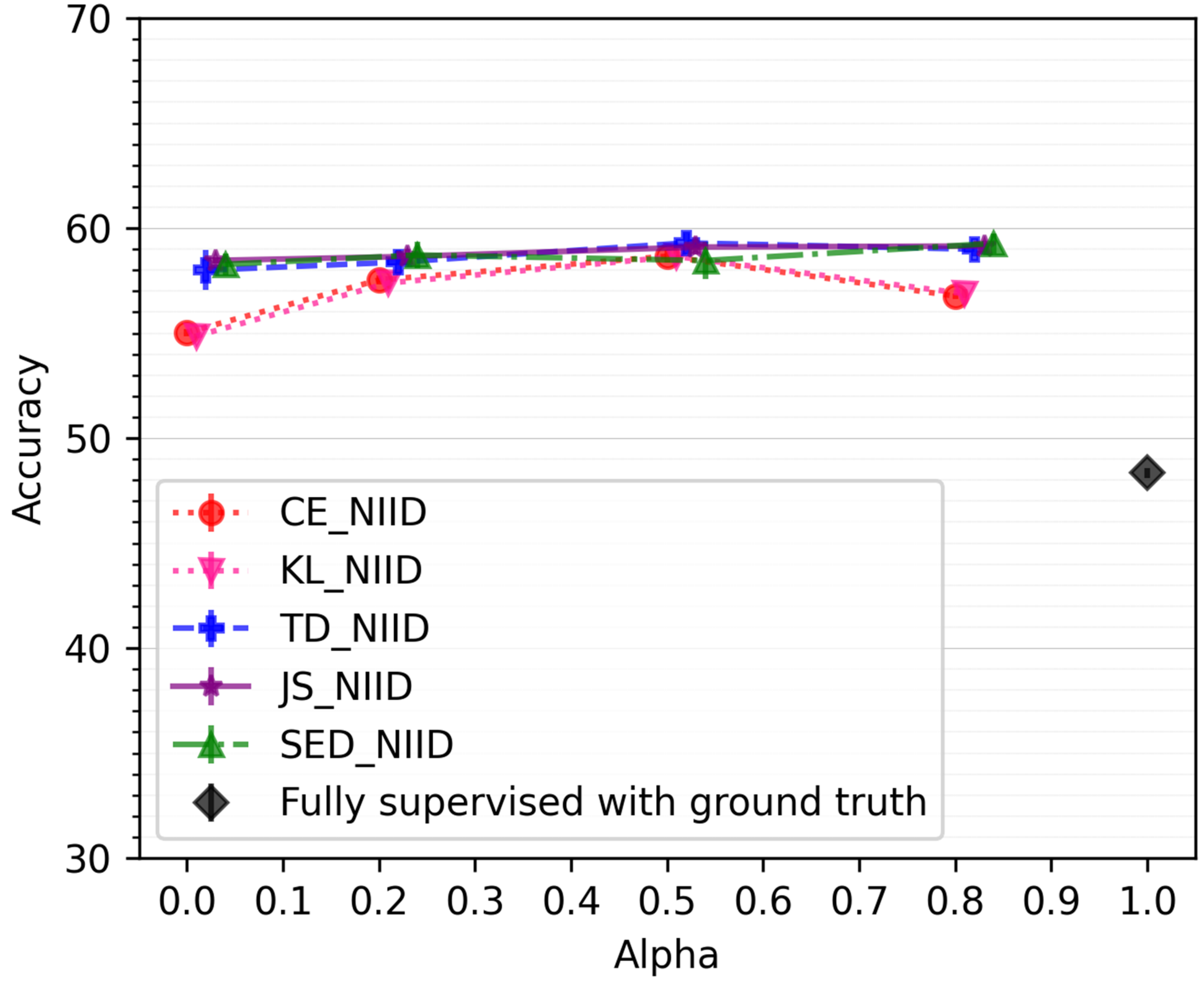
$$\mathcal{L}_{k,KD}(\mathbf{w}_2^k, x) = \mathbb{E}_{x \sim X^k} f\left(\mathbf{p}^k(\mathbf{w}_2^k, x), \frac{\sum_{\phi \in \Phi_k} \mathbf{p}^{\phi}(x)}{|\Phi|}\right) \quad \text{(AVG)}$$



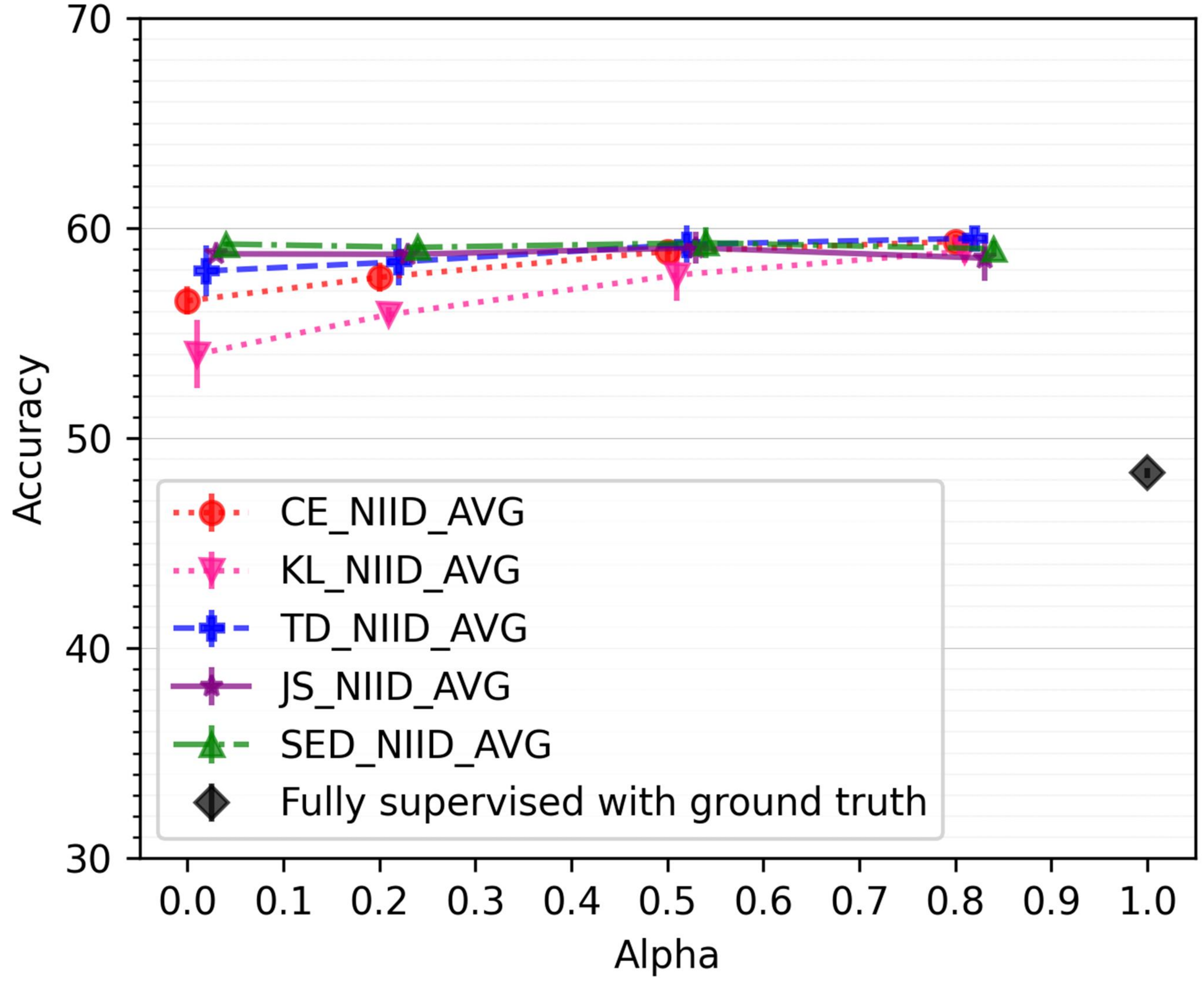
- KD led to an increase in the average accuracy of the clients' models compared to the fully supervised approach
- AVG case : For $\alpha > 0$ minimal diff. between SED and JS; for α=0.2 CE and KL perform worse; for α=0.8 all measures perform similarly, with KL having higher variance across clients

Results: SUN397 non-iid

$$\mathcal{L}_{k,KD}(\mathbf{w}_2^k, x) = \sum_{\phi\in\Phi_k} \mathbb{E}_{x\sim X^k}\, f\left(\mathbf{p}^k(\mathbf{w}_2^k,x), \mathbf{p}^{\phi}(x)\right) \quad \text{(SUM)}$$



$$\mathcal{L}_{k,KD}(\mathbf{w}_2^k, x) = \mathbb{E}_{x\sim X^k}\, f\left(\mathbf{p}^k(\mathbf{w}_2^k,x), \frac{\sum_{\phi\in\Phi_k}\mathbf{p}^{\phi}(x)}{|\Phi|}\right) \quad \text{(AVG)}$$



- This confirms the argument that when the client's training data is scarce (leading to model overfitting) communication between clients can enhance generalization and improve client's performance
- CE and KL are outperformed by SED, TD, and JS distances in many of the tested configurations

Conclusions

- We evaluated **different information dissimilarity measures** in a **distributed KD setting** across various data distributions

- **Key findings:**
  - The **KD-loss** based on the dissimilarity between the current client's soft-predictions and the ***average*** of soft-predictions from remote clients showed the **best trade-off between accuracy and efficiency**
  - In the *iid* case, all measures have similar accuracy, so Triangular Dist. is preferred as it is more efficient
  - The distance measures impact model training on *non-iid* data
  - The commonly used cross-entropy and Kullback-Leibler divergences are not always the most effective

- **Future work:**
  - Investigate gradient stability (exploding/vanishing gradients)
  - Evaluate performance with more nodes and diverse network topologies

Thank you!

joaquim.molo@phd.unipi.it

lucia.vadicamo@isti.cnr.it

Emanuele.carlini@isti.cnr.it

joaquimbasa/Distributed_KD_Information_Dissimilarity

Molo, M. J., Vadicamo, L., Carlini, E., Gennaro, C., & Connor, R. (2024, October). Information Dissimilarity Measures in Decentralized Knowledge Distillation: A Comparative Analysis. In *International Conference on Similarity Search and Applications* (pp. 140-154).