Benchmarking Federated Learning & Knowledge Distillation for Point Cloud Classification

Aizierjiang Aiersilan

University of Macau
ezharjan@outlook.com

Abstract. Deploying 3D point cloud analysis in privacy-sensitive and resource-constrained settings faces two coupled barriers: data cannot be centralized for training, and the trained model must run on limited edge hardware. We present a multi-seed benchmark that jointly evaluates federated learning (FL) and knowledge distillation (KD) for 3D point cloud classification. It spans thirteen FL algorithms and ten KD objectives, supporting their full 130-pair teacher–objective cross-product per dataset; every standardized configuration is repeated over three random seeds for 504 training runs in total, with the complete combined grid evaluated at multi-seed scale on the clinical dataset. We characterize federated degradation and the combined-pipeline pitfall on ModelNet40, then validate them on a real-world clinical craniosynostosis dataset of patient head shapes, where the privacy and edge-deployment stakes are concrete. We report three findings. First, under extreme non-independent and identically distributed (non-IID) label skew, standalone FL degrades sharply: on ModelNet40 the strongest method reaches only 76.32% against a 92.26% centralized reference, on the clinical data the best reaches 75.83% against 100%, and the four server-side optimizers collapse to near the chance level; the best algorithm differs by dataset, so none is universally robust. Second, distillation compresses the teacher into a student 74.51% smaller and roughly twice as fast at inference, with five of the seven objectives evaluated on ModelNet40 matching or surpassing the 92.44% teacher. Third, the combined pipeline exposes an evaluation pitfall: when distillation keeps a hard-label cross-entropy term on a labeled proxy split, a collapsed federated teacher at 8.50% paired with Logit-MSE still yields a 92.94% student. This 84.4-point gap reflects the proxy labels rather than the federated model, and the hard-label term reuses the very labels whose privacy motivated federation. Objectives without a hard-label term instead track teacher quality ($r \approx 0.99$ on the clinical grid) and collapse when the teacher does. We therefore recommend evaluating FL-KD pipelines with label-free distillation, so that the reported accuracy reflects the federated teacher rather than the proxy.

Keywords: Federated learning · Knowledge distillation · 3D point cloud · Model compression · Non-IID · Privacy-preserving learning · Medical imaging · Benchmark

Benchmark & Code: https://ezharjan.github.io/FLKD3DBenchmark

1 Introduction

3D point cloud analysis underpins applications including autonomous driving, robotics, augmented reality, and medical imaging, where sensor data are naturally represented as irregular point sets in Euclidean space. Architectures such as PointNet [33] and PointNet++ [34] achieve strong classification performance on standard benchmarks, yet their deployment in privacy-sensitive, resource-constrained settings faces two challenges that have not been studied together in one benchmark.

The first challenge is data privacy. In medical [25, 38, 43], industrial [9, 27, 31], and multi-institutional settings [21, 26, 52], raw sensor data cannot be centralized because of regulatory or competitive constraints. Federated learning (FL) [30] addresses this by coordinating model training across clients without exchanging raw data, using iterative local updates and global aggregation. Even so, gradient inversion attacks [6, 60] show that transmitted gradients can still leak private data, and the broader open problems in FL are surveyed in [15]. The statistical heterogeneity of data across clients, the non-IID problem [12, 14, 28], degrades convergence and final accuracy even for well-established aggregation algorithms. These effects are well documented for image classification, but their impact on hierarchical 3D feature extraction remains largely uncharacterized.

The second challenge is computational efficiency at deployment. A full PointNet++ model is often unsuitable for resource-constrained edge platforms such as embedded robotic controllers or portable clinical scanners. Knowledge distillation (KD) [11] transfers the capacity of a large teacher into a compact student, producing a smaller and faster network that retains much of the teacher's accuracy. Several distillation objectives exist for image recognition, including feature hints [39], attention transfer [56], and self-distillation [5], but their effectiveness for 3D point cloud models has not been systematically benchmarked.

Combining FL and KD is appealing: FL trains a shared model without centralizing data, and KD then compresses that model for edge deployment. A recent survey [36] identifies the absence of domain-specific benchmarks as a critical gap. In practice, the two-stage pipeline is usually evaluated end to end, by reporting the compressed student's accuracy, rather than by asking whether that accuracy actually reflects the federated teacher. To the best of our knowledge, no prior work provides a systematic, reproducible evaluation of FL and KD strategies for 3D point cloud classification across the full cross-product of federated teachers and distillation objectives, let alone on clinical 3D shape data; 3DFFL [17] addresses federated few-shot learning for point clouds but does not evaluate KD losses or the combined pipeline.

We study two complementary datasets. On ModelNet40 [51], a 40-class benchmark, we characterize federated degradation and isolate the combined-pipeline pitfall against a strong centralized reference. We then validate that pitfall on a real-world clinical craniosynostosis dataset [42] of patient head shapes, where privacy mandates federation [38, 43] and edge compression is a deployment necessity; there the full multi-seed cross-product confirms that hard-label masking conceals federated failure on a privacy-critical task. We make four contributions:

- **Benchmark design.** The benchmark provides thirteen FL algorithms, a centralized baseline, and ten KD objectives, and supports the complete two-stage cross-product of all 130 teacher–objective pairs per dataset. Our standardized multi-seed evaluation comprises 504 training runs over three seeds: on ModelNet40 the centralized baseline and the thirteen FL teachers, and on the clinical dataset the centralized baseline, the thirteen FL teachers, the ten KD objectives, and the full 130-pair combined grid; the ModelNet40 distillation and combined results come from a separate focused round (Section 4.3) selected for its wide spread of teacher quality. All results are reported as mean and standard deviation. The framework is open-source and is built to be extended with further FL algorithms, KD objectives, and datasets through one shared interface.
- **Quantified degradation under label skew.** On ModelNet40 the strongest FL method trails the centralized reference by about 16 points (FedNova at 76.32% against 92.26%), on the clinical data by about 24 points (FedProx at 75.83% against 100%), and the four server-side optimizers (FedAvgM, FedAdam, FedYogi, FedAdagrad) collapse to near the chance level on both. The method ranking is dataset-dependent, so no single algorithm is robust to label skew.
- **Effective compression through distillation.** On ModelNet40 distillation transfers the teacher's accuracy into a student that is 74.51% smaller and roughly twice as fast at inference (about 50% lower latency), with five of seven objectives matching or surpassing the 92.44% teacher; the clinical task exhibits the same near-ceiling compression.
- **A controlled diagnosis of the combined pipeline.** The teacher-by-objective grid separates the distillation objectives by their reliance on hard labels. Objectives that keep a hard-label cross-entropy term recover the student to near-centralized accuracy regardless of teacher quality, so that on ModelNet40 a collapsed federated teacher at 8.50% paired with Logit-MSE still yields a 92.94% student, an 84.4-point gap driven by the proxy labels, not by transfer. The feature- and attention-based objectives whose cross-entropy weight is exactly zero instead track the teacher and collapse when it collapses; the full multi-seed cross-product on the clinical data confirms the same split across all thirteen teachers, with the four pure-transfer objectives at $r \approx 0.99$. Hard-label distillation thus masks federated failure while reusing exactly the labels that federation was meant to keep private, and we accordingly recommend two controls (distilling with the cross-entropy term removed, or on an unlabeled proxy split) that distinguish genuine knowledge transfer from this artifact.

2 Related Work

2.1 3D Point Cloud Learning

Early volumetric methods discretized shapes onto occupancy grids and applied 3D convolutions [51], but their memory cost grows cubically with resolution.

PointNet [33] introduced a permutation-invariant architecture that processes raw point sets via shared multi-layer perceptrons and global max pooling. PointNet++ [34] extended this with hierarchical set abstraction layers that capture local geometry at progressively coarser scales through farthest point sampling and ball-query grouping. Later architectures incorporate graph-based aggregation as in DGCNN [48] and self-attention as in Point Transformer [59]. Further advances include kernel-point convolutions (KPConv) [44], random-sampling segmentation with RandLA-Net [13], the Point Cloud Transformer (PCT) [8], PointNeXt [35], and Point Transformer V2 [50]. Sarker *et al.* [41] survey deep learning for point cloud understanding. We use PointNet++ single-scale grouping (SSG) as the teacher backbone because its hierarchical, locality-sensitive feature extraction provides a demanding setting for federated aggregation under non-IID label skew.

2.2 Federated Learning

We benchmark thirteen FL algorithms spanning the major families of aggregation strategy. FedAvg [30] averages client models weighted by dataset size after local gradient descent. To counter client drift under heterogeneity, FedProx [22] adds a proximal term, SCAFFOLD [16] introduces client-level control variates that reduce gradient variance, FedNova [47] normalizes the accumulated local updates to remove the objective inconsistency caused by differing local-step counts, and FedDyn [1] adds a dynamic per-client regularizer that aligns local and global stationary points. A second family modifies the server-side update: FedAvgM applies server momentum [12], while FedAdam, FedYogi, and FedAdagrad apply adaptive server optimizers [37]. FedMedian [53] replaces averaging with coordinate-wise median aggregation for robustness, and FedBN [23] keeps batch-normalization statistics client-local to mitigate feature shift. Two methods tailored to non-IID data are MOON [19], which adds a model-contrastive term aligning client and global representations, and Ditto [20], which learns a personalized model per client alongside the global one. The effects of non-IID partitions on several of these strategies are documented for image classification [12,14,28]; we characterize them for hierarchical 3D point cloud architectures, including the methods most often recommended for label skew.

2.3 Knowledge Distillation

We benchmark ten KD objectives spanning the major families of distillation loss. Hinton *et al.* [11] showed that a student can match the soft class-probability outputs of a larger teacher via temperature-scaled Kullback–Leibler divergence. FitNets [39] aligned intermediate layer representations through hint-based training. Zagoruyko and Komodakis [56] transferred attention maps derived from intermediate feature tensors. Born-Again Networks [5] showed that self-distillation can yield a student that surpasses its teacher. Our benchmark also evaluates four representative relational and decoupled objectives: contrastive representation distillation (CRD) [45], relational knowledge distillation (RKD) [32], similarity-preserving distillation (SP) [46], and decoupled knowledge distillation (DKD) [58],

alongside simpler logit-matching variants [2, 29]. Other objectives such as the comprehensive feature overhaul [10] and the label-smoothing reinterpretation of KD [55] are not evaluated here. A general survey is given by Gou *et al.* [7]. For 3D data, structured distillation for detection [57] and adversarial distillation [40] have been explored, but neither benchmarks multiple loss objectives nor connects KD with federated pre-training. Qin *et al.* [36] survey the combination of FL and KD and identify the absence of domain-specific evaluations as a key open problem. KD has been integrated into FL to mitigate non-IID degradation and catastrophic forgetting in clinical settings [18, 54], yet systematic benchmarking for 3D hierarchical models remains unexplored. A complementary line distills on unlabeled or server-side proxy data [24], the approach our diagnosis ultimately motivates.

2.4 Clinical Shape Analysis

Craniosynostosis is the premature fusion of cranial sutures, which alters head shape. Schaufelberger *et al.* [42] released a statistical shape model of patients, with instances for the principal suture-fusion types and a control group, giving a privacy-conscious source of patient-like head shapes. Because such scans are distributed across clinics and cannot be pooled freely [25, 38, 43], the setting motivates federated training and compact models for point-of-care devices.

3 Benchmark Design

3.1 Datasets and Non-IID Partitioning

We evaluate on two complementary datasets. ModelNet40 [51] provides point clouds across 40 shape categories; the standard pre-processed normal-resampled split provides 9,843 training and 2,468 test shapes. It serves as a widely used benchmark on which to measure federated degradation at scale. The craniosynostosis dataset [42] provides 3D head-surface instances sampled from a statistical shape model, with four classes (a healthy control group and the coronal, metopic, and sagittal suture-fusion pathologies) and 100 instances per class; we use a deterministic stratified 80/20 split, giving 320 training and 80 test shapes. Each shape is represented by 1,024 points with three spatial coordinates and no surface normals, an input tensor of shape $3 \times 1{,}024$. The two datasets are deliberately contrasting: a large-vocabulary collection where the centralized task is moderately hard, and a small four-class clinical collection where the centralized task is essentially solved. We report the primary empirical results on ModelNet40, the more demanding benchmark, and use the clinical collection as a corroborating study; its near-perfect centralized ceiling lets us read off teacher quality unambiguously. Beyond these two, the data loader is dataset-agnostic and currently supports ModelNet10 [51], OmniObject3D [49], YCB [3], and GazeboSim [4] under the same interface, so the protocol can be transferred to other point-cloud sources.

To model federated deployment, the training set of each dataset is partitioned among five clients with a label-skew non-IID strategy: all training samples are

sorted by class label and the ordered sequence is divided into five consecutive equal slices. Each client therefore concentrates on a disjoint segment of the label space, reflecting the class imbalance that arises when different facilities collect data for different purposes. This is a deliberately severe, worst-case form of heterogeneity; we adopt it because it most directly isolates the failure modes the benchmark is designed to surface, and we return to its implications in Section 5.

Multi-seed protocol. To make the comparison statistically interpretable rather than dependent on a single run, every configuration (baseline, FL, KD, and combined) is trained independently with three random seeds, $\{7, 42, 123\}$. These seeds control the model initialization and the stochastic elements of training, namely mini-batch ordering and point-cloud augmentation; the label-skew partition is deterministic and held fixed across seeds, so the reported variance isolates training-time randomness rather than partition luck. Unless stated otherwise, all reported numbers are the mean over the three seeds, and tables give the corresponding standard deviation.

3.2 Model Architectures

Teacher: PointNet++ SSG. The teacher follows the PointNet++ single-scale grouping architecture [34] on raw coordinates without surface normals, with three hierarchical set-abstraction levels followed by a two-layer classification head. It occupies 5.65 MB on ModelNet40 and 5.62 MB on the four-class clinical data, measured from its parameter and buffer tensors.

Student: a compact PointNet++ SSG. The student, SmallPointNet2, keeps the teacher's hierarchical structure with narrower set-abstraction and classification layers. It occupies 1.44 MB on ModelNet40 (a 74.51% reduction from the teacher) and 1.42 MB on the clinical data, and exposes a penultimate feature so that feature-based distillation has a signal.

3.3 Federated Learning Strategies

We evaluate thirteen FL aggregation algorithms on the teacher architecture, chosen to span the major mechanism families: drift correction, server-side optimization, robust aggregation, and personalization. All configurations share the same federated setting of five clients, five local epochs per round, and twenty communication rounds, with local training using the Adam optimizer at learning rate 0.001 and batch size 24. Each strategy is listed with its distinguishing mechanism alongside its accuracy in Table 1.

3.4 Knowledge Distillation Objectives

We evaluate ten distillation objectives, all transferring knowledge from the PointNet++ SSG teacher to the compact student, spanning the major families of KD loss. Logit matching is represented by Vanilla KD [11], Logit-MSE [2], Cosine [29], and DKD [58]; intermediate-feature and attention alignment by

Feature KD [39], Attention Transfer [56], and SP [46]; relational and contrastive transfer by RKD [32] and CRD [45]; and self-distillation by the Born-Again protocol [5]. All but DKD share a soft-target weight α=0.5, and the feature, attention, SP, and RKD objectives add an intermediate-representation weight β=0.5; the soft-target objectives use temperature T=2.0. For Self-Distillation, an initial compact student trained from the teacher then supervises the final reported student.

Central to our combined-pipeline diagnosis is the weight on the hard-label cross-entropy term, $\mathcal{L}_{\mathrm{CE}}$. Vanilla KD, Logit-MSE, Cosine, CRD, and Self-Distillation weight it by $(1-\alpha)$=0.5, and DKD adds it at full weight, so all six retain a direct supervised signal from the ground-truth labels. The feature, attention, SP, and RKD objectives weight it by $(1-\alpha-\beta)$, which is exactly 0 at α=β=0.5, so these four are pure teacher-transfer objectives with no hard-label term.

3.5 Combined Pipeline

The two-stage combined pipeline first trains a federated teacher (PointNet++ SSG) for twenty rounds under the non-IID partition; the best checkpoint then supervises the compact student in a KD run on the centralized training set. Figure 1 illustrates the workflow. This setup mirrors deployments where a periodically aggregated federated model is compressed for edge hardware using a server-side proxy dataset. We run Stage 2 on the fully labeled centralized data precisely to test the pitfall it can create, so that the spread of teacher quality and the role of the hard-label term can both be measured directly. The pipeline is dataset-agnostic; on ModelNet40 we pair federated teachers spanning the full quality range with the distillation objectives. The same two stages apply unchanged to the clinical dataset, where we run the complete multi-seed cross-product of all thirteen FL teachers with all ten KD objectives (130 teacher–objective pairs) and reproduce the same recovery illusion.

4 Experiments

4.1 Centralized Baseline

We train the teacher on the full training split of each dataset over three seeds, using the common 200-epoch training protocol. The teacher reaches $\mathbf{92.26 \pm 0.04}$% instance accuracy on ModelNet40 (89.58 ± 0.10% mean class accuracy) and $\mathbf{100.00 \pm 0.00}$% on the four-class clinical data, using the best checkpoint per run averaged over seeds. The clinical task is essentially solved when trained centrally, which is useful for our purposes: it fixes an unambiguous reference for teacher quality against which the FL degradation and the combined-pipeline behavior can be read. We compare all FL and KD results against these centralized references.

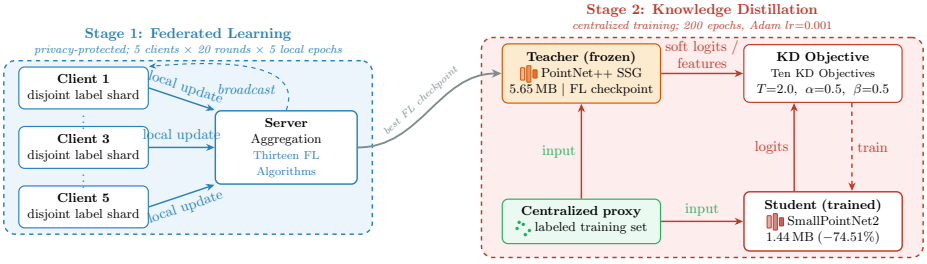


Fig. 1: Overview of the two-stage FL-KD benchmark pipeline. **Stage 1** (blue, privacy-protected): Five non-IID clients train on label-skewed slices of the dataset; the server applies one of the thirteen benchmarked FL algorithms to produce a shared PointNet++ SSG teacher. **Stage 2** (red): The best FL teacher checkpoint supervises a compact student on the centralized training set with a KD objective, yielding a smaller and faster student model.

4.2 Federated Learning Results

Table 1 reports the best instance accuracy of each FL strategy over twenty communication rounds under the non-IID partition, for both datasets. Per-round convergence trajectories appear in the supplementary material (Appendix C, Figure A7).

Federated training degrades severely under the label-skew partition on both datasets (Table 1). On ModelNet40 the strongest method, FedNova, reaches only 76.32%, 15.94 points below centralized; the drift-correction methods (FedProx 71.04%, FedDyn 66.00%, SCAFFOLD 64.26%) improve on plain FedAvg (58.51%), whereas MOON (45.84%) falls below it and Ditto (61.36%) exceeds it only modestly. FedMedian reaches 39.13%, and the four server-side optimizers fall to near-trivial accuracy.

The ranking is dataset-dependent, itself a finding: no single algorithm is robust. FedNova, the ModelNet40 leader, drops to mid-pack on the clinical data (50.83%), where FedProx leads at 75.83% and is the most consistent across both; SCAFFOLD falls from strong to 36.67% on the four-class problem, while Ditto and MOON rise into the upper group. The key result is constant: the best method trails centralized by 15.94 points on ModelNet40 and 24.17 on the clinical data, and the four server-side optimizers collapse to near-chance accuracy on both (4.05–6.28% near 2.5% ModelNet40 floor, 25.00–26.25% near the four-class floor).

These results point to a vulnerability of this hierarchical 3D extractor to label skew. When each client sees only a few categories, the early set-abstraction layers likely calibrate their geometric priors to locally dominant shapes and transfer poorly after aggregation. The server-side optimizers are the most fragile: for the three adaptive variants, even at a tuned server rate of 0.05 the averaged client updates form an unreliable pseudo-gradient whose second-moment estimates become ill-conditioned, likely amplifying rather than correcting divergence [37]. Adaptive server aggregation is thus a practical risk in label-skewed edge settings,

Table 1: Standalone FL instance accuracy (%, best per run, mean±std over three seeds) on both datasets, with each method's distinguishing mechanism (5 clients, non-IID label-skew, 20 rounds, 5 local epochs per round). Methods are ordered by ModelNet40 accuracy. The ranking is dataset-dependent: FedNova leads on ModelNet40 but drops to mid-pack on the clinical data, where FedProx leads, while the four server-side optimizers fail on both. On the balanced clinical test set the mean-class accuracy equals the instance accuracy.

Method	Distinguishing Mechanism	ModelNet40 ↑	Craniosynostosis ↑
Centralized	Non-federated reference	92.26 ±0.04	100.00 ±0.00
FedNova [47]	Normalized local averaging	**76.32** ±1.55	50.83 ±8.04
FedProx [22]	Proximal regularization	71.04 ±0.87	**75.83** ±10.63
FedDyn [1]	Dynamic regularization	66.00 ±1.28	58.33 ±8.13
SCAFFOLD [16]	Control-variate correction	64.26 ±3.75	36.67 ±1.44
Ditto [20]	Personalized plus global model	61.36 ±4.11	71.25 ±9.92
FedAvg [30]	Weighted parameter averaging	58.51 ±0.25	69.58 ±2.60
FedBN [23]	Client-local batch-norm	55.28 ±2.37	65.42 ±4.73
MOON [19]	Model-contrastive term	45.84 ±7.51	69.17 ±8.78
FedMedian [53]	Coordinate-wise median	39.13 ±5.79	64.58 ±7.53
FedAdagrad [37]	Server-side Adagrad	6.28 ±1.09	26.25 ±2.17
FedAdam [37]	Server-side Adam	4.71 ±0.98	25.00 ±0.00
FedAvgM [12]	Server-side momentum	4.05 ±0.00	25.00 ±0.00
FedYogi [37]	Server-side Yogi	4.05 ±0.00	25.00 ±0.00

where validation data for tuning is rarely available. This is a worst-case regime; softer partitions would narrow the gap (Section 5).

Communication is identical across strategies: the full model occupies 5.65 MB per transmission, and each communication round sends and receives across five clients, totaling 56.5 MB per round, or approximately 1.13 GB over twenty rounds. Wall-clock cost varies with the local objective: MOON and Ditto roughly double FedAvg's per-round time by training an auxiliary model on every client. Full payloads and times are reported in the supplementary material (Appendix C, Table A6).

4.3 Knowledge Distillation Results

We report the ModelNet40 distillation and combined stages on our primary testbed. The combined stage pairs federated teachers spanning the full quality range, from strong aggregators down to collapsed teachers, with seven distillation objectives (including a Basic KD variant that keeps the cross-entropy term at full weight as a probe); this spread is what makes the recovery illusion measurable. These teachers are individual checkpoints, so their standalone accuracies differ from the multi-seed means in Table 1: SCAFFOLD and FedDyn sit at 8.50% and 16.67% here against multi-seed means of 64.26% and 66.00%, and the centralized reference reaches 92.44%, close to the 92.26% multi-seed mean. The benchmark

Table 2: KD results on ModelNet40. Teacher: PointNet++ SSG (5.65 MB, 92.44%). Student: SmallPointNet2 (1.44 MB, 74.51% size reduction, approximately 50% lower inference latency). Instance accuracy is the peak over all training epochs. All strategies are run from the same pre-trained teacher checkpoint. The seven shown are six of the ten benchmark objectives plus a full-weight Basic KD probe; all ten are evaluated on the clinical task (Section 4.5).

Strategy	Inst. Acc. (%) ↑	Class Acc. (%) ↑	vs. Teacher (%) ↑
Teacher (PointNet++ SSG) [34]	92.44	90.24	0.00
Attention Transfer [56]	**92.74**	90.05	**+0.30**
Basic KD [7]	92.69	89.30	+0.25
Vanilla KD [11]	92.61	89.43	+0.17
Self-Distillation [5]	92.61	89.66	+0.17
Feature KD [39]	92.57	89.52	+0.13
Cosine Similarity [29]	92.35	**90.12**	−0.09
Logit-MSE [2]	87.07	75.08	−5.37

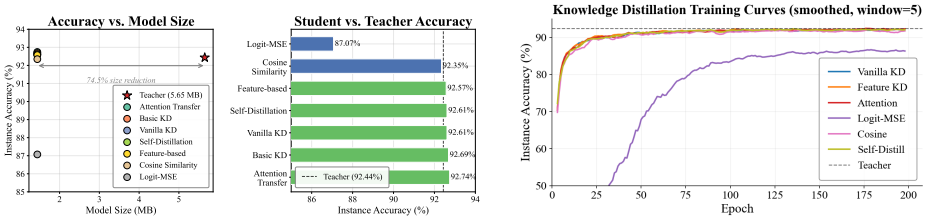
supports the full 13×10 cross-product, which Section 4.5 reports at full multi-seed scale on the clinical task.

Table 2 reports instance accuracy for each distillation objective distilled from the centralized ModelNet40 teacher (92.44%), and Figure 2 relates accuracy to model size together with the student-versus-teacher training dynamics.

Distillation is an effective compressor for this architecture (Figure 2). Five of the seven objectives meet or surpass the 92.44% teacher on the compact student: Attention Transfer (92.74%), Basic KD (92.69%), Vanilla KD (92.61%), Self-Distillation (92.61%), and Feature KD (92.57%); Cosine Similarity (92.35%) trails by 0.09 points. The lone outlier is Logit-MSE (87.07%): a mean-squared error on the output vectors is scale-sensitive relative to the cross-entropy term and unbalances the loss, slowing convergence (Figure 2b). Compression is uniform across objectives (74.51% smaller, from 5.65 to 1.44 MB, and about 50% lower inference latency). This near-uniform success under a strong teacher is exactly what makes standalone KD a poor probe of the federated teacher: when the labels alone reach the ceiling, the objective is barely tested. The combined experiment exposes that.

4.4 Combined Federated Learning and Knowledge Distillation Results

Table 3 shows instance accuracy for the two-stage combined pipeline across the five FL teachers and the seven KD objectives, and Figure 3 plots each objective's sensitivity to teacher quality. These five teachers span the full standalone-accuracy range, from FedProx at 67.61% down to SCAFFOLD at 8.50% (FedAvg reaches 60.83% with size-weighted and 56.38% with uniform averaging, and FedDyn 16.67%); we also evaluated FedMedian at 31.16% and the server-side FedAvgM and FedAdam at 4.05%, excluding them as teachers because their near-chance



(a) Accuracy vs. model size **(b)** Training-epoch convergence

Fig. 2: Efficiency and distillation performance on ModelNet40. (a) Instance accuracy versus model size (left) and each objective's student accuracy relative to the 92.44% teacher (right): all SmallPointNet2 students (1.44 MB) are 74.51% smaller than the teacher (5.65 MB), with most objectives within 0.3 points of the teacher. **(b)** Training-epoch accuracy for six of the seven objectives (Basic KD overlaps Vanilla KD and is omitted for clarity): Logit-MSE converges slower and plateaus below the teacher baseline, whereas the others asymptote at or above 92.35% within 100 epochs.

logits carry no usable distillation gradient. It shows a consistent dichotomy that turns on a single design choice, the weight of the hard-label cross-entropy term.

Objectives that keep a hard-label term mask the teacher. The five objectives with a positive cross-entropy coefficient (Vanilla KD, Basic KD, Self-Distillation, Logit-MSE, and Cosine Similarity) recover near-centralized accuracy from every teacher: setting aside the runs that stopped before epoch 200 (marked †), every completed cell for these five lies between 92.13% and 92.94%. The peak cell, SCAFFOLD + Logit-MSE at 92.94%, exceeds the 92.44% centralized reference by 0.50 points despite an FL teacher that converges to only 8.50% on its own, a gap of 84.4 points above the teacher it distilled from, which cannot reflect genuine teacher-to-student transfer. Seven of the ten highest-accuracy configurations pair a hard-label objective with a collapsed teacher (three from SCAFFOLD at 8.50%, four from FedDyn at 16.67%).

Objectives with no hard-label term reveal the teacher. Feature- and attention-based objectives behave oppositely. For collapsed teachers (SCAFFOLD at 8.50%, FedDyn at 16.67%) they reach only 15.82% to 18.20%, comparable to the teacher itself, and even with the FedProx teacher (67.61%) they stay in the 53% to 55% range. At $\alpha{=}\beta{=}0.5$ their cross-entropy coefficient $(1{-}\alpha{-}\beta)$ is exactly 0, removing the labeled-data fallback: both remaining terms depend entirely on teacher quality, so a collapsed teacher drives the student toward meaningless representations. These objectives are thus a label-free probe of teacher quality, showing genuine transfer unpropped by proxy labels.

Interpretation. Both behaviors follow from the loss $\mathcal{L} = w_{\mathrm{CE}}\,\mathcal{L}_{\mathrm{CE}}(\hat{y}, y) + \mathcal{L}_{\mathrm{transfer}}$ on the labeled centralized set. When $w_{\mathrm{CE}}{>}0$ and the teacher has collapsed, its soft targets give little useful gradient and the well-conditioned cross-entropy term dominates, so the student learns the proxy labels directly; when $w_{\mathrm{CE}}{=}0$, the teacher is the only signal, so a failed teacher yields a failed student. *An FL-KD evaluation that uses hard labels therefore cannot measure the federated teacher: it reduces the pipeline to centralized training on the proxy labels, reusing*

Table 3: Two-stage combined pipeline: instance accuracy (%) on ModelNet40 (focused round, individual checkpoints). Each cell is the best instance accuracy of a SmallPointNet2 student trained with the given KD objective after federated pre-training of the teacher with the given FL strategy; "Std." is the teacher's own standalone instance accuracy. Column headers: Vanilla KD (VKD) [11], Basic KD (BKD) [7], Self-Distillation (SD) [5], Attention Transfer (AT) [56], Feature KD (FKD) [39], Logit-MSE (MSE) [2], Cosine Similarity (Cos) [29]. FedAvg (van.) uses uniform (unweighted) client averaging; the other FedAvg uses size-weighted averaging. **Bold** marks the best value in each KD-objective column. † Denotes runs stopped before 200 epochs; values are the best accuracy recorded to that point. These five teachers were chosen to span the full standalone-accuracy range ("Std."); the complete 13×10 multi-seed cross-products appear in the supplementary material (ModelNet40) and in Section 4.5 (clinical task).

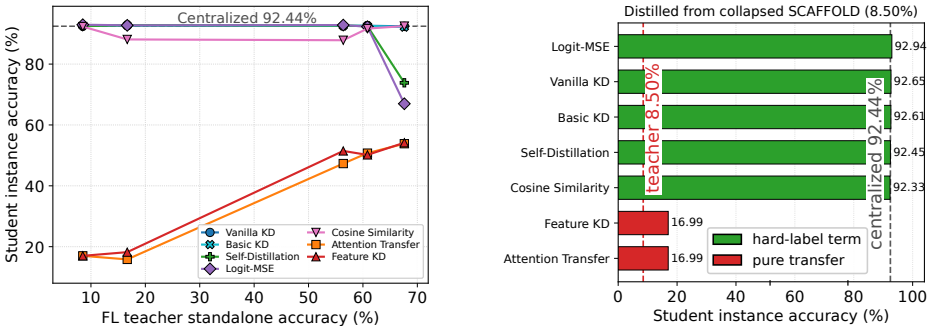
FL Teacher	Std.	VKD ↑	BKD ↑	SD ↑	AT ↑	FKD ↑	MSE ↑	Cos ↑
FedAvg [30]	60.83	92.61	92.57	92.45	50.74	50.16	92.29†	91.67†
FedAvg (van.) [30]	56.38	92.57	92.53	92.65	47.31	51.46	92.82	87.84†
FedProx [22]	67.61	92.41	92.13	73.85†	**53.84**	**54.16**	66.97†	**92.49**
FedDyn [1]	16.67	**92.65**	**92.77**	**92.78**	15.82†	18.20	92.73	88.10†
SCAFFOLD [16]	8.50	**92.65**	92.61	92.45	16.99†	16.99	**92.94**	92.33

the very labels whose privacy motivated federation. A sound evaluation distills without hard labels, by aggregating client soft predictions on unlabeled data [24] or omitting the cross-entropy term; the feature-based curves of Figure 3 are exactly that measurement. Two controls expose the illusion: a *no-cross-entropy* control (α=1, $\mathcal{L}_{\mathrm{CE}}$ removed) and an *unlabeled-proxy* control. Both leave the teacher as the sole signal, so a sound student falls back toward its teacher's standalone accuracy (a SCAFFOLD-conditioned student toward 8.50%, not 92.94%); residual recovery signals cross-entropy-driven masking.

4.5 Real-World Validation: Clinical Craniosynostosis

The ModelNet40 grid isolates the recovery illusion under controlled conditions; the clinical craniosynostosis dataset shows that it survives on real patient data, where the stakes are concrete. These head shapes sit in separate clinics, cannot be pooled freely [38, 43], and must run on portable point-of-care scanners, so federation and compression are operational requirements rather than conveniences. We therefore run the complete cross-product of all thirteen FL teachers and ten KD objectives (130 teacher–objective pairs) at full multi-seed scale (three seeds). Standalone distillation compresses the 100% centralized teacher into a 1.42 MB student (74.7% smaller, roughly 2× faster) at near-ceiling accuracy across all ten objectives (supplementary material, Appendix B, Table A4); the unambiguous ceiling lets us read teacher quality directly.

The combined grid reproduces the ModelNet40 dichotomy, and the hard-label cross-entropy term again decides the outcome (Table 4; see also the supplementary material, Appendix B, Figure A4). The six objectives that keep a cross-entropy



(a) Student vs. teacher accuracy

(b) Recovery from a collapsed teacher

Fig. 3: The recovery illusion on ModelNet40 (focused round). (a) Student instance accuracy versus FL-teacher standalone accuracy for the seven distillation objectives of Table 3: the five objectives with a positive cross-entropy coefficient (Vanilla KD, Basic KD, Self-Distillation, Logit-MSE, Cosine Similarity) stay near the 92% centralized line for every teacher, even one collapsed to 8.50%, because the labeled proxy split supervises the student directly, whereas Attention Transfer and Feature KD, whose cross-entropy coefficient is exactly 0, fall in step with the teacher. **(b)** Student accuracy distilled from the collapsed SCAFFOLD teacher (8.50% standalone) under each objective, ordered by accuracy: the five hard-label objectives recover to the 92.44% centralized ceiling, a gap of up to 84.4 points above the teacher, while Attention Transfer and Feature KD stay near the teacher's own level. The full 13×10 multi-seed heatmaps appear in the supplementary material.

term recover the student from any teacher: their accuracy varies by at most 7.5 points across the thirteen teachers and does not correlate with teacher quality (r from -0.53 to $+0.37$). A teacher frozen at the 25% chance level (FedAvgM) still yields a 97.5% student under Vanilla KD and 100% under CRD. The four pure-transfer objectives, whose cross-entropy coefficient is 0 at $\alpha{=}\beta{=}0.5$, instead track the teacher almost perfectly ($r{=}0.986$ to 0.990): they span 57.5 to 58.3 points across teachers, fall to chance when the teacher does (FedAvgM with Attention Transfer or SP yields 25.0%), and recover only behind a strong teacher (FedProx with Feature KD reaches 84.2%, tracking its 75.83% teacher).

The clinical data thus turns the diagnosis into a deployment warning. A hard-label-distilled student here logs near-100% accuracy from a federated teacher no better than chance, while the cross-entropy term silently reuses the patient labels that federation was meant to protect. Only label-free distillation reports the federated teacher's true quality. Per-cell means, standard deviations, and per-seed values appear in the supplementary material (Appendix C, Tables A8 and A9; Appendix D, Table A13).

5 Discussion and Limitations

Why does FL degrade severely for 3D architectures? Extreme label skew forces each client's layers to overfit local categories, producing gradient

Table 4: Two-stage combined pipeline on the clinical task, summarized by KD objective over all thirteen FL teachers (mean instance accuracy, %, three seeds). "CE wt." is the weight on the hard-label cross-entropy term. "Chance teacher" uses FedAvgM [12], a teacher at the 25% chance level; "best teacher" uses FedProx [22] at 75.83%. "Range" is the spread of student accuracy across the thirteen teachers, and r is the Pearson correlation between student accuracy and the teacher's standalone accuracy. The six objectives that keep a cross-entropy term recover the student regardless of teacher quality ($r \approx 0$); the four with zero cross-entropy weight track the teacher ($r \approx 0.99$). Per-cell values appear in the supplementary material (Appendix C, Tables A8 and A9).

KD Objective	CE wt.	Chance teacher	Best teacher	Range	r
Objectives retaining a hard-label term					
Vanilla KD [11]	0.5	97.5	99.6	2.5	+0.37
Self-Distillation [5]	0.5	99.2	99.6	1.2	+0.19
Cosine Similarity [29]	0.5	99.6	100.0	0.8	−0.05
Contrastive (CRD) [45]	0.5	100.0	99.6	0.8	−0.53
Logit-MSE [2]	0.5	97.5	99.2	3.8	+0.10
Decoupled KD [58]	1.0	93.8	96.7	7.5	−0.10
Pure teacher-transfer objectives ($1 - \alpha - \beta = 0$)					
Feature KD [39]	0	27.1	84.2	57.5	+0.99
Attention Transfer [56]	0	25.0	83.3	58.3	+0.99
Relational KD [32]	0	27.5	81.7	57.5	+0.99
Similarity-Preserving [46]	0	25.0	82.9	58.3	+0.99

heterogeneity that standard aggregation cannot handle. Adaptive server optimizers suffer most, likely from ill-conditioned second-moment estimates under heterogeneous updates [37]. That this holds on both a 40-class benchmark and a four-class clinical dataset shows it is not an artifact of label-space size.

No universally robust algorithm. The dataset-dependent ranking (Table 1) cautions against choosing an FL method from a single dataset: the leader on one dataset slips on the other, and even FedProx, the most consistent of the thirteen, never closes the gap to centralized training on both. That the contrastive, personalized, and adaptive server-side methods all leave this gap indicates these objectives do not by themselves overcome the extreme label skew studied here.

The hard-label term governs the combined pipeline. An FL-KD pipeline's reported accuracy is governed by whether the distillation loss keeps a hard-label term, not by the federated teacher: with such a term the student reaches the ceiling from any teacher. A faithful assessment must therefore distill on unlabeled data or drop the cross-entropy term, and report teacher-tracking accuracy.

Hierarchical versus transformer backbones. The degradation may also depend on the backbone. The fixed-radius neighborhood aggregation that may underlie PointNet++'s sensitivity to label skew is specific to hierarchical models. Transformer backbones such as PCT [8] and Point Transformer V2 [50], and kernel-point methods such as KPConv [44], aggregate features through different mechanisms whose behavior under parameter averaging is not obvious in advance.

Whether such architectures degrade less than hierarchical models under non-IID federation is an open question that our single-backbone study cannot settle.

Limitations. Several boundaries remain. First, we evaluate two datasets, ModelNet40 and a statistical-shape-model craniosynostosis set; larger real-world scanned collections remain to be tested. Second, the ModelNet40 combined diagnosis uses teachers spanning the full quality range down to collapsed checkpoints, and the clinical dataset confirms it at full multi-seed scale over the complete 13×10 grid, whose 100% ceiling saturates the standalone KD numbers and leaves the combined grid as the diagnostic. Third, we study one non-IID regime, the extreme disjoint label-skew partition; softer partitions would narrow the gap, and a heterogeneity sweep is left to future work. Fourth, the seeds vary initialization and optimization but not the deterministic partition, so the reported deviations capture training-time variance. Fifth, the architecture is a fixed PointNet++ SSG teacher and compact student; whether the effects persist for transformer or kernel-point backbones is open. Finally, all runs fix five clients, twenty rounds, and homogeneous client architectures; heterogeneous client models would require aggregation across dissimilar parameter spaces.

6 Conclusion

We presented a multi-seed benchmark for FL and KD in 3D point cloud classification, spanning thirteen FL algorithms, ten KD objectives, and their 130-pair cross-product. Standalone federated training degrades sharply under extreme label skew on both datasets, the method ranking is dataset-dependent, and the four server-side optimizers collapse to near the chance level, so no single algorithm is robust. Distillation compresses the strong centralized teacher into a student 74.51% smaller and roughly twice as fast at near-ceiling accuracy. The combined teacher-by-objective grid then exposes the central caution: any objective that keeps a hard-label cross-entropy term recovers the student to the centralized ceiling whatever the federated teacher, so its reported accuracy measures the proxy labels, not the federated model; only label-free objectives track teacher quality. The clinical craniosynostosis data shows the effect is not specific to ModelNet40: on a privacy-critical task, hard-label distillation reports a near-100% student from a chance-level teacher, while reusing the very labels federation was meant to protect. We therefore recommend evaluating FL-KD pipelines with label-free distillation, so that the reported accuracy measures the federated teacher.

For reproducibility, the source code and the trained model checkpoints for every reported run are publicly available at `https://ezharjan.github.io/FLKD3DBenchmark`, and new backbones, tasks, datasets, algorithms, and student architectures integrate through one shared interface. We envision this benchmark and its label-free evaluation protocol as a robust foundation for advancing privacy-preserving, edge-deployable 3D point cloud systems, and we welcome community contributions to expand its scope.

References

1. Acar, D.A.E., Zhao, Y., Navarro, R.M., Mattina, M., Whatmough, P.N., Saligrama, V.: Federated learning based on dynamic regularization. arXiv preprint arXiv:2111.04263 (2021), https://arxiv.org/abs/2111.04263, accessed: 2026-06-29
2. Ba, L.J., Caruana, R.: Do deep nets really need to be deep? Advances in neural information processing systems **27** (2014), https://proceedings.neurips.cc/paper/2014/hash/b0c355a9dedccb50e5537e8f2e3f0810-Abstract.html, accessed: 2026-06-29
3. Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., Dollar, A.M.: The ycb object and model set: Towards common benchmarks for manipulation research. In: 2015 international conference on advanced robotics (ICAR). pp. 510–517. IEEE (2015), https://doi.org/10.1109/ICAR.2015.7251504, accessed: 2026-06-29
4. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. IEEE (2022), https://doi.org/10.1109/ICRA46639.2022.9811809, accessed: 2026-06-29
5. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: International conference on machine learning. pp. 1607–1616. PMLR (2018), https://proceedings.mlr.press/v80/furlanello18a.html, accessed: 2026-06-29
6. Geiping, J., Bauermeister, H., Dröge, H., Moeller, M.: Inverting gradients-how easy is it to break privacy in federated learning? Advances in neural information processing systems **33**, 16937–16947 (2020), https://proceedings.neurips.cc/paper/2020/hash/c4ede56bbd98819ae6112b20ac6bf145-Abstract.html, accessed: 2026-06-29
7. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International journal of computer vision **129**(6), 1789–1819 (2021), https://doi.org/10.1007/s11263-021-01453-z, accessed: 2026-06-29
8. Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M.: Pct: Point cloud transformer. Computational visual media **7**(2), 187–199 (2021), https://doi.org/10.1007/s41095-021-0229-5, accessed: 2026-06-29
9. Hao, M., Li, H., Luo, X., Xu, G., Yang, H., Liu, S.: Efficient and privacy-enhanced federated learning for industrial artificial intelligence. IEEE Transactions on Industrial Informatics **16**(10), 6532–6542 (2019), https://doi.org/10.1109/TII.2019.2945367, accessed: 2026-06-29
10. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1921–1930 (2019), https://openaccess.thecvf.com/content_ICCV_2019/html/Heo_A_Comprehensive_Overhaul_of_Feature_Distillation_ICCV_2019_paper.html, accessed: 2026-06-29
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015), https://arxiv.org/abs/1503.02531, accessed: 2026-06-29
12. Hsu, T.M.H., Qi, H., Brown, M.: Measuring the effects of non-identical data distribution for federated visual classification. arXiv preprint arXiv:1909.06335 (2019), https://arxiv.org/abs/1909.06335, accessed: 2026-06-29

13. Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A.: Randla-net: Efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11108–11117 (2020), https://openaccess.thecvf.com/content_CVPR_2020/html/Hu_RandLA-Net_Efficient_Semantic_Segmentation_of_Large-Scale_Point_Clouds_CVPR_2020_paper.html, accessed: 2026-06-29
14. Jimenez G., D.M., Solans, D., Heikkila, M., Vitaletti, A., Kourtellis, N., Anagnostopoulos, A., Chatzigiannakis, I.: Non-iid data in federated learning: A survey with taxonomy, metrics, methods, frameworks and future directions. arXiv preprint arXiv:2411.12377 (2024), https://arxiv.org/abs/2411.12377, accessed: 2026-06-29
15. Kairouz, P., McMahan, H.B., et al.: Advances and open problems in federated learning. Foundations and trends in machine learning **14**(1-2), 1–210 (2021), https://doi.org/10.1561/2200000083, accessed: 2026-06-29
16. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: Scaffold: Stochastic controlled averaging for federated learning. In: International conference on machine learning. pp. 5132–5143. PMLR (2020), https://proceedings.mlr.press/v119/karimireddy20a.html, accessed: 2026-06-29
17. Khan, A.A., Ahmad, K.M., Shafiq, S., Amin, W., Kumar, R.: 3dffl: privacy-preserving federated few-shot learning for 3d point clouds in autonomous vehicles. Scientific Reports **14**(1), 19589 (2024), https://doi.org/10.1038/s41598-024-70326-5, accessed: 2026-06-29
18. Kim, S., Park, H., Kang, M., Jin, K.H., Adeli, E., Pohl, K.M., Park, S.H.: Federated learning with knowledge distillation for multi-organ segmentation with partially labeled datasets. Medical image analysis **95**, 103156 (2024), https://doi.org/10.1016/j.media.2024.103156, accessed: 2026-06-29
19. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10713–10722 (2021), https://openaccess.thecvf.com/content/CVPR2021/html/Li_Model-Contrastive_Federated_Learning_CVPR_2021_paper.html, accessed: 2026-06-29
20. Li, T., Hu, S., Beirami, A., Smith, V.: Ditto: Fair and robust federated learning through personalization. In: International conference on machine learning. pp. 6357–6368. PMLR (2021), https://proceedings.mlr.press/v139/li21h.html, accessed: 2026-06-29
21. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: Challenges, methods, and future directions. IEEE signal processing magazine **37**(3), 50–60 (2020), https://doi.org/10.1109/MSP.2020.2975749, accessed: 2026-06-29
22. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems **2**, 429–450 (2020), https://proceedings.mlsys.org/paper_files/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html, accessed: 2026-06-29
23. Li, X., Jiang, M., Zhang, X., Kamp, M., Dou, Q.: Fedbn: Federated learning on non-iid features via local batch normalization. arXiv preprint arXiv:2102.07623 (2021), https://arxiv.org/abs/2102.07623, accessed: 2026-06-29
24. Lin, T., Kong, L., Stich, S.U., Jaggi, M.: Ensemble distillation for robust model fusion in federated learning. Advances in neural information processing systems **33**, 2351–2363 (2020), https://proceedings.neurips.cc/paper/2020/hash/18df51b97ccd68128e994804f3eccc87-Abstract.html, accessed: 2026-06-29

25. Liu, S., Yang, H.H., Tao, Y., Feng, Y., Hao, J., Liu, Z.: Privacy-preserved federated learning for 3d tooth segmentation in intra-oral mesh scans. Frontiers in Communications and Networks **3**, 907388 (2022), https://doi.org/10.3389/frcmn.2022.907388, accessed: 2026-06-29
26. Long, G., Tan, Y., Jiang, J., Zhang, C.: Federated learning for open banking. In: Federated learning: privacy and incentive, pp. 240–254. Springer (2020), https://doi.org/10.1007/978-3-030-63076-8_17, accessed: 2026-06-29
27. Lu, Y., Huang, X., Dai, Y., Maharjan, S., Zhang, Y.: Blockchain and federated learning for privacy-preserved data sharing in industrial iot. IEEE Transactions on Industrial Informatics **16**(6), 4177–4186 (2019), https://doi.org/10.1109/TII.2019.2942190, accessed: 2026-06-29
28. Lu, Z., Pan, H., Dai, Y., Si, X., Zhang, Y.: Federated learning with non-iid data: A survey. IEEE Internet of Things Journal **11**(11), 19188–19209 (2024), https://doi.org/10.1109/JIOT.2024.3376548, accessed: 2026-06-29
29. Luo, C., Zhan, J., Xue, X., Wang, L., Ren, R., Yang, Q.: Cosine normalization: Using cosine similarity instead of dot product in neural networks. In: International conference on artificial neural networks. pp. 382–391. Springer (2018), https://doi.org/10.1007/978-3-030-01418-6_38, accessed: 2026-06-29
30. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. Pmlr (2017), https://proceedings.mlr.press/v54/mcmahan17a.html, accessed: 2026-06-29
31. Nguyen, D.C., Ding, M., Pathirana, P.N., Seneviratne, A., Li, J., Poor, H.V.: Federated learning for internet of things: A comprehensive survey. IEEE communications surveys & tutorials **23**(3), 1622–1658 (2021), https://doi.org/10.1109/COMST.2021.3075439, accessed: 2026-06-29
32. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3967–3976 (2019), https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html, accessed: 2026-06-29
33. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017), https://openaccess.thecvf.com/content_cvpr_2017/html/Qi_PointNet_Deep_Learning_CVPR_2017_paper.html, accessed: 2026-06-29
34. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017), https://proceedings.neurips.cc/paper/2017/hash/d8bf84be3800d12f74d8b05e9b89836f-Abstract.html, accessed: 2026-06-29
35. Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H., Elhoseiny, M., Ghanem, B.: Pointnext: Revisiting pointnet++ with improved training and scaling strategies. Advances in neural information processing systems **35**, 23192–23204 (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/9318763d049edf9a1f2779b2a59911d3-Abstract-Conference.html, accessed: 2026-06-29
36. Qin, L., Zhu, T., Zhou, W., Yu, P.S.: Knowledge distillation in federated learning: A survey on long lasting challenges and new solutions. International Journal of Intelligent Systems **2025**(1), 7406934 (2025), https://doi.org/10.1155/int/7406934, accessed: 2026-06-29

37. Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečnỳ, J., Kumar, S., McMahan, H.B.: Adaptive federated optimization. arXiv preprint arXiv:2003.00295 (2020), https://arxiv.org/abs/2003.00295, accessed: 2026-06-29
38. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al.: The future of digital health with federated learning. NPJ digital medicine **3**(1), 119 (2020), https://doi.org/10.1038/s41746-020-00323-1, accessed: 2026-06-29
39. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014), https://arxiv.org/abs/1412.6550, accessed: 2026-06-29
40. Sanjay, S., Akash, J., Dimple, A.S., et al.: Adversarial learning based knowledge distillation on 3d point clouds. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2932–2941. IEEE (2025), https://openaccess.thecvf.com/content/WACV2025/html/J_Adversarial_Learning_Based_Knowledge_Distillation_on_3D_Point_Clouds_WACV_2025_paper.html, accessed: 2026-06-29
41. Sarker, S., Sarker, P., Stone, G., Gorman, R., Tavakkoli, A., Bebis, G., Sattarvand, J.: A comprehensive overview of deep learning techniques for 3d point cloud classification and semantic segmentation. Machine Vision and Applications **35**(4) (2024), https://doi.org/10.1007/s00138-024-01543-1, accessed: 2026-06-29
42. Schaufelberger, M., Kühle, R., Wachter, A., Weichel, F., Hagen, N., Ringwald, F., Eisenmann, U., Hoffmann, J., Engel, M., Freudlsperger, C., et al.: A statistical shape model of craniosynostosis patients and 100 model instances of each pathology. Zenodo: Geneve, Switzerland (2021), https://doi.org/10.5281/zenodo.10167123, accessed: 2026-06-29
43. Sheller, M.J., Edwards, B., Reina, G.A., et al.: Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific Reports **10**, 12598 (Jul 2020), https://doi.org/10.1038/s41598-020-69250-1, accessed: 2026-06-29
44. Thomas, H., Qi, C.R., Deschaud, J.E., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6411–6420 (2019), https://openaccess.thecvf.com/content_ICCV_2019/html/Thomas_KPConv_Flexible_and_Deformable_Convolution_for_Point_Clouds_ICCV_2019_paper.html, accessed: 2026-06-29
45. Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019), https://arxiv.org/abs/1910.10699, accessed: 2026-06-29
46. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1365–1374 (2019), https://openaccess.thecvf.com/content_ICCV_2019/html/Tung_Similarity-Preserving_Knowledge_Distillation_ICCV_2019_paper.html, accessed: 2026-06-29
47. Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the objective inconsistency problem in heterogeneous federated optimization. Advances in neural information processing systems **33**, 7611–7623 (2020), https://proceedings.neurips.cc/paper/2020/hash/564127c03caab942e503ee6f810f54fd-Abstract.html, accessed: 2026-06-29
48. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics (tog) **38**(5), 1–12 (2019), https://doi.org/10.1145/3326362, accessed: 2026-06-29

49. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 803–814 (2023), https://doi.org/10.1109/CVPR52729.2023.00084, accessed: 2026-06-29

50. Wu, X., Lao, Y., Jiang, L., Liu, X., Zhao, H.: Point transformer v2: Grouped vector attention and partition-based pooling. Advances in Neural Information Processing Systems **35**, 33330–33342 (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/d78ece6613953f46501b958b7bb4582f-Abstract-Conference.html, accessed: 2026-06-29

51. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015), https://openaccess.thecvf.com/content_cvpr_2015/html/Wu_3D_ShapeNets_A_2015_CVPR_paper.html, accessed: 2026-06-29

52. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST) **10**(2), 1–19 (2019), https://doi.org/10.1145/3298981, accessed: 2026-06-29

53. Yin, D., Chen, Y., Kannan, R., Bartlett, P.: Byzantine-robust distributed learning: Towards optimal statistical rates. In: International conference on machine learning. pp. 5650–5659. Pmlr (2018), https://proceedings.mlr.press/v80/yin18a.html, accessed: 2026-06-29

54. Yu, L., Huang, J.: Cyclic federated learning method based on distribution information sharing and knowledge distillation for medical data. Electronics **11**(23), 4039 (2022), https://doi.org/10.3390/electronics11234039, accessed: 2026-06-29

55. Yuan, L., Tay, F.E., Li, G., Wang, T., Feng, J.: Revisiting knowledge distillation via label smoothing regularization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3903–3911 (2020), https://openaccess.thecvf.com/content_CVPR_2020/html/Yuan_Revisiting_Knowledge_Distillation_via_Label_Smoothing_Regularization_CVPR_2020_paper.html, accessed: 2026-06-29

56. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016), https://arxiv.org/abs/1612.03928, accessed: 2026-06-29

57. Zhang, L., Dong, R., Tai, H.S., Ma, K.: Pointdistiller: Structured knowledge distillation towards efficient and compact 3d detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21791–21801 (2023), https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_PointDistiller_Structured_Knowledge_Distillation_Towards_Efficient_and_Compact_3D_Detection_CVPR_2023_paper.html, accessed: 2026-06-29

58. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 11953–11962 (2022), https://openaccess.thecvf.com/content/CVPR2022/html/Zhao_Decoupled_Knowledge_Distillation_CVPR_2022_paper.html, accessed: 2026-06-29

59. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16259–16268 (2021), https://openaccess.thecvf.com/content/ICCV2021/html/Zhao_Point_Transformer_ICCV_2021_paper.html, accessed: 2026-06-29
60. Zhu, L., Liu, Z., Han, S.: Deep leakage from gradients. Advances in neural information processing systems **32** (2019), https://proceedings.neurips.cc/paper/2019/hash/60a6c4002cc7b29142def8871531281a-Abstract.html, accessed: 2026-06-29