DART: A Server-side Plug-in for Resource-efficient Robust Federated Learning

Supplementary Material

7. Method Details

7.1. Data Augmentation

The DART method adopts the standard AugMix [15] augmentation strategy, which stochastically mixes diverse transformation operations such as posterize, equalize, rotate, translate, and shear. This procedure introduces multiple sources of randomness, including the selection of operations, their severity levels, the depth of transformation chains, and mixing weights.

First, three operations, op_1, op_2, and op_3, are sampled from the set of available operations, $\mathcal{O}$. The sampled operations are sequentially composed to generate transformation chains of increasing depth:

$$\text{op}_{12} = \text{op}_2 \circ \text{op}_1, \text{ and } \text{op}_{123} = \text{op}_3 \circ \text{op}_2 \circ \text{op}_1 \quad (5)$$

Then one chain is sampled uniformly at random from $\{\text{op}_1, \text{op}_{12}, \text{op}_{123}\}$. This process is repeated S times, where S is a hyperparameter denoting mixture width, resulting in S operation chains $\{\text{chain}_i\}_{i=1}^{S}$. S is set to 3 by default [15].

The augmented sample $\mathbf{x}_{\text{aug}}$ is generated from a sample $\mathbf{x}$ using:

$$\mathbf{x}_{\text{aug}} = \eta\mathbf{x} + (1-\eta)\sum_{i=1}^{S} m_i \times \text{chain}_i(\mathbf{x}) \quad (6)$$

where $(m_1 \ldots m_i)$ and η are random weights sampled from Dirichlet$(\alpha \ldots \alpha)$ and from Beta(α, α) distributions, respectively. As can be seen, the augmentation procedure mixes the output of the operation chains. It also includes direct contribution from the original image $\mathbf{x}$ to prevent the loss of sematic information.

This diverse augmentation process is designed to improve the robustness of machine learning models by exposing them to a wide range of data during training.

7.2. Overall System Algorithm

Section 4.1 presented the overall DART-enhanced FL system design. In a standard FL setting, clients perform local training while a central server aggregates their updates. Our proposed plug-in, **DART**, is integrated on the server side, operating immediately after model aggregation and before the updated global model is redistributed to the clients. Algorithm 1 details the DART-enhanced FL system.

7.3. DART Algorithm

Section 4.2 introduced our novel server-side plug-in, **DART**, designed to enable resource-efficient and robust federated learning. Algorithm 2 details the DART procedure.

Algorithm 1 `System Description`.
K: number of clients; D_k: local private dataset of client k; D_0: public unlabeled dataset; G: number of global updates; E: local rounds per global update; T_{DART}: global rounds per DART update

Server executes:

Initialize $\mathbf{w}_0$;
for each global round $t = 1, 2, \ldots, G$ **do**
  for each client $k = 1, 2, \ldots, K$ **in parallel do**
    $\mathbf{w}_k^{t+1} \leftarrow \texttt{ClientUpdate}(k, \mathbf{w}_0^t)$;
  end for
  $\mathbf{w}_0^{t+1} \leftarrow \texttt{Aggregate}(\mathbf{w}_1^{t+1}, \ldots, \mathbf{w}_k^{t+1})$;
  if $t \bmod T_{\text{DART}} = 0$ **then**
    $\mathbf{w}_{\text{rob}}^{t+1} \leftarrow \texttt{DART}(\mathbf{w}_0^{t+1}, D_0)$;
    $\mathbf{w}_0^{t+1} \leftarrow \mathbf{w}_{\text{rob}}^{t+1}$;
  end if
end for

ClientUpdate(k, $\mathbf{w}$):

for each local epoch $i = 1, 2, \ldots, E$ **do**
  for batch $\mathbf{b} \in D_k$ **do**
    $\mathbf{w} \leftarrow \texttt{LocalTraining}(\mathbf{w}, \mathbf{b})$;
  end for
end for
return $\mathbf{w}$ to server

8. Theorem Proof

The theoretical analysis in this work is inspired by Lin et al. [29], who study ensemble distillation for model fusion in model heterogeneous FL. In their work, a generalization bound is developed that relates the loss of distilled federated models to that of a centrally trained model, explicitly characterizing the impact of distribution mismatch across clients and server. Building on this foundation, we derive a bound that relates the pre-DART (teacher) and post-DART (student) clean risks, enabling us to formally quantify the utility degradation introduced by DART while enhancing robustness.

Setup and notation. Let $\mathcal{X}$ be the input space and $\mathcal{Y}$ the label set. Let $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}$ be distributions on $\mathcal{X} \times \mathcal{Y}$. A classifier $h : \mathcal{X} \to \mathcal{Y}$ has risk on a distribution Q

$$R_Q(h) = \Pr_{(\mathbf{x},\mathbf{y})\sim Q}\big(h(\mathbf{x}) \neq \mathbf{y}\big) \quad (7)$$

Algorithm 2 DART.
D_{tr} and D_{val}: public unlabeled training and validation datasets partitioned from D_0; $T_{\max}$: maximum number of DART epochs; T_{val}: early stopping threshold; η_s: DART learning rate; $\mathbf{w}_0$: pre-trained weights.

counter $\leftarrow 0$; $\mathbf{w} \leftarrow \mathbf{w}_0$; $\text{loss}_{\min} \leftarrow \infty$
for each epoch $i = 1, 2, \ldots, T_{\max}$ **do**
  for batch $\mathbf{b} \in D_{\text{tr}}$ **do**
    $\mathbf{w} \leftarrow \mathbf{w} - \eta_s \nabla \mathcal{L}_{\text{DART}}(\mathbf{w}; \mathbf{b})$
  end for
  if $\mathcal{L}_{\text{DART}}(\mathbf{w}; D_{\text{val}}) \leq \text{loss}_{\min}$ **then**
    $\text{loss}_{\min} \leftarrow \mathcal{L}_{\text{DART}}(\mathbf{w}; D_{\text{val}})$
    $\mathbf{w}_{\text{best}} \leftarrow \mathbf{w}$; counter $\leftarrow 0$
  else
    counter ++
  end if
  if counter $= T_{\text{val}} - 1$ **then**
    break
  end if
end for
$\mathbf{w}_{\text{rob}} \leftarrow \mathbf{w}_{\text{best}}$
return $\mathbf{w}_{\text{rob}}$

For $h, h' : \mathcal{X} \to \mathcal{Y}$, define their *disagreement probability* on a distribution Q by

$$\text{Dis}_Q(h, h') \;=\; \Pr_{\mathbf{x} \sim Q}\big(h(\mathbf{x}) \neq h'(\mathbf{x})\big) \tag{8}$$

We consider a teacher $f_{\mathbf{w}_t}$ and a student $f_{\mathbf{w}_s}$ that make predictions via posteriors $p_t(\cdot \mid \mathbf{x})$ and $p_s(\cdot \mid \mathbf{x})$, with decisions $f_{\mathbf{w}_t}(\mathbf{x}) = \arg\max_{\mathbf{y}} p_t(\mathbf{y} \mid \mathbf{x})$ and $f_{\mathbf{w}_s}(\mathbf{x}) = \arg\max_{\mathbf{y}} p_s(\mathbf{y} \mid \mathbf{x})$. The client distribution is $\mathcal{D}_{\text{in}}$, while $\mathcal{D}_{\text{out}}$ is the server distribution.

Define the *teacher margin* at $\mathbf{x}$:

$$m_t(\mathbf{x}) \;=\; p_t(\mathbf{y}_t^* \mid \mathbf{x}) - \max_{\mathbf{y} \neq \mathbf{y}_t^*} p_t(\mathbf{y} \mid \mathbf{x}) \tag{9}$$

where $\mathbf{y}_t^* = \arg\max_{\mathbf{y}} p_t(\mathbf{y} \mid x)$ and $m_t(\mathbf{x}) \in [0, 1]$. Define the *margin tail* on $\mathcal{D}_{\text{out}}$ by $\theta_{\mathcal{D}_{\text{out}}}(\gamma) = \Pr_{\mathbf{x} \sim \mathcal{D}_{\text{out}}}\big(m_t(\mathbf{x}) \leq \gamma\big)$ for $\gamma \in (0, 1]$. The total variation distance is

$$\text{TV}(\mathbf{x}) \;=\; \tfrac{1}{2}\big\|p_t(\cdot \mid \mathbf{x}) - p_s(\cdot \mid \mathbf{x})\big\|_1 \tag{10}$$

and the expected distillation KL divergence on $\mathcal{D}_{\text{out}}$:

$$\delta \;=\; \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\text{out}}}\Big[\text{KL}\big(p_t(\cdot \mid \mathbf{x}) \,\|\, p_s(\cdot \mid \mathbf{x})\big)\Big] \tag{11}$$

Finally, the $\mathcal{H}\Delta\mathcal{H}$-divergence [3] between $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}$ is

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\text{in}}, \mathcal{D}_{\text{out}}) \;=\; 2 \sup_{h, h' \in \mathcal{H}} \Big|\text{Dis}_{\mathcal{D}_{\text{in}}}(h, h') - \text{Dis}_{\mathcal{D}_{\text{out}}}(h, h')\Big|$$

for some hypothesis class $\mathcal{H}$.

Lemma 1. *For any $\mathbf{x} \in \mathcal{X}$, if $f_{\mathbf{w}_s} \neq f_{\mathbf{w}_t}$ then*

$$m_t(\mathbf{x}) \;\leq\; 2\,\text{TV}(\mathbf{x})$$

Proof. Let $\mathbf{y}_t^* = f_{\mathbf{w}_t}(\mathbf{x})$ and let $\mathbf{y}' = f_s(x) \neq \mathbf{y}_t^*$. Since the student prefers $\mathbf{y}'$, $p_s(\mathbf{y}' \mid \mathbf{x}) \geq p_s(\mathbf{y}_t^* \mid \mathbf{x})$. Let $\Delta_{\mathbf{y}} := p_s(\mathbf{y} \mid \mathbf{x}) - p_t(\mathbf{y} \mid \mathbf{x})$ for each class $\mathbf{y}$, then

$$\begin{aligned}
m_t(\mathbf{x}) &\leq p_t(\mathbf{y}_t^* \mid \mathbf{x}) - p_t(\mathbf{y}' \mid \mathbf{x}) && (12)\\
&= \big(p_t(\mathbf{y}_t^* \mid \mathbf{x}) - p_s(\mathbf{y}_t^* \mid \mathbf{x})\big) && (13)\\
&\quad + \big(p_s(\mathbf{y}_t^* \mid x) - p_s(\mathbf{y}' \mid \mathbf{x})\big)\\
&\quad + \big(p_s(\mathbf{y}' \mid \mathbf{x}) - p_t(\mathbf{y}' \mid \mathbf{x})\big)\\
&\leq |\Delta_{\mathbf{y}_t^*}| + 0 + |\Delta_{\mathbf{y}'}| && (14)\\
&\leq \sum_{\mathbf{y} \in \mathcal{Y}} |\Delta_{\mathbf{y}}| && (15)\\
&= \big\|p_s(\cdot \mid \mathbf{x}) - p_t(\cdot \mid \mathbf{x})\big\|_1 && (16)\\
&= 2\,\text{TV}(\mathbf{x}) && (17)
\end{aligned}$$

which proves the claim. □

Theorem (Clean Student Risk Bound under Distillation)**.** *For any $\gamma \in (0, 1]$,*

$$\begin{aligned}
\mathcal{R}_{\mathcal{D}_{in}}(f_{\mathbf{w}_s}) \leq\; & \mathcal{R}_{\mathcal{D}_{in}}(f_{\mathbf{w}_t}) + \theta_{\mathcal{D}_{out}}(\gamma)\\
& + \frac{2}{\gamma}\sqrt{\frac{\delta}{2}} + \frac{1}{2}\, d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{in}, \mathcal{D}_{out})
\end{aligned} \tag{18}$$

Proof. **(1) Risk decomposition.** For any $(\mathbf{x}, \mathbf{y})$,

$$\mathbf{1}\big[f_{\mathbf{w}_s}(\mathbf{x}) \neq \mathbf{y}\big] \;\leq\; \mathbf{1}\big[f_{\mathbf{w}_t}(\mathbf{x}) \neq \mathbf{y}\big] + \mathbf{1}\big[f_{\mathbf{w}_s}(\mathbf{x}) \neq f_{\mathbf{w}_t}(\mathbf{x})\big] \tag{19}$$

Taking expectation over $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}_{\text{in}}$ gives

$$R_{\mathcal{D}_{\text{in}}}(f_{\mathbf{w}_s}) \;\leq\; R_{\mathcal{D}_{\text{in}}}(f_{\mathbf{w}_t}) + \text{Dis}_{\mathcal{D}_{\text{in}}}(f_{\mathbf{w}_s}, f_{\mathbf{w}_t}) \tag{20}$$

(2) Move disagreement from $\mathcal{D}_{\text{in}}$ to $\mathcal{D}_{\text{out}}$. By the definition of $d_{\mathcal{H}\Delta\mathcal{H}}$,

$$\text{Dis}_{\mathcal{D}_{\text{in}}}(f_{\mathbf{w}_s}, f_{\mathbf{w}_t}) \;\leq\; \text{Dis}_{\mathcal{D}_{\text{out}}}(f_{\mathbf{w}_s}, f_{\mathbf{w}_t}) + \frac{1}{2}\, d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\text{in}}, \mathcal{D}_{\text{out}}) \tag{21}$$

Combining with (20),

$$\begin{aligned}
R_{\mathcal{D}_{\text{in}}}(f_{\mathbf{w}_s}) \leq\; & R_{\mathcal{D}_{\text{in}}}(f_{\mathbf{w}_t}) + \text{Dis}_{\mathcal{D}_{\text{out}}}(f_{\mathbf{w}_s}, f_{\mathbf{w}_t})\\
& + \frac{1}{2}\, d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\text{in}}, \mathcal{D}_{\text{out}})
\end{aligned} \tag{22}$$

(3) Control disagreement on $\mathcal{D}_{\text{out}}$ via margin and TV. By Lemma 1, for any $\mathbf{x}$,

$$\big\{f_{\mathbf{w}_s}(\mathbf{x}) \neq f_{\mathbf{w}_t}(\mathbf{x})\big\} \;\subseteq\; \big\{m_t(\mathbf{x}) \leq 2\,\text{TV}(\mathbf{x})\big\}$$

Fix $\gamma \in (0, 1]$. Then $\{m_t \leq 2\,\mathrm{TV}\} \subseteq \{m_t \leq \gamma\} \cup \{2\,\mathrm{TV} \geq \gamma\}$ and,

$$\mathrm{Dis}_{\mathcal{D}_{\text{out}}}(f_{\mathbf{w}_s}, f_{\mathbf{w}_t}) \;\leq\; \theta_{\mathcal{D}_{\text{out}}}(\gamma) \;+\; \Pr_{\mathbf{x}\sim\mathcal{D}_{\text{out}}}\big(\,2\,\mathrm{TV}(\mathbf{x}) \geq \gamma\,\big)$$

Now apply Markov's inequality (since $\mathrm{TV} \geq 0$), Pinsker and Jensen:

$$\Pr(\,2\,\mathrm{TV} \geq \gamma\,) \;\leq\; \frac{2}{\gamma}\,\mathbb{E}[\mathrm{TV}] \tag{23}$$

$$\leq\; \frac{2}{\gamma}\,\mathbb{E}\Big[\sqrt{\tfrac{1}{2}\,\mathrm{KL}(p_t\|p_s)}\Big] \tag{24}$$

$$\leq\; \frac{2}{\gamma}\sqrt{\tfrac{1}{2}\,\mathbb{E}[\mathrm{KL}(p_t\|p_s)]} \;=\; \frac{2}{\gamma}\sqrt{\frac{\delta}{2}} \tag{25}$$

Hence

$$\mathrm{Dis}_{\mathcal{D}_{\text{out}}}(f_{\mathbf{w}_s}, f_{\mathbf{w}_t}) \;\leq\; \theta_{\mathcal{D}_{\text{out}}}(\gamma) \;+\; \frac{2}{\gamma}\sqrt{\frac{\delta}{2}} \tag{26}$$

(4) Combine. Insert (26) into (22) to obtain the stated bound. □

9. Resource Setup

9.1. Profiling Client Hardware

Each client is represented by an NVIDIA Jetson Orin Nano [33], equipped with an Ampere GPU, a 6-core ARM CPU, and 8 GB of RAM, delivering up to 67 TOPS of compute. We profile batch execution time and batch energy consumption directly on the hardware. Power is measured using Jetson Stats [4], and energy is computed by integrating power over time, approximated as the product of sampled power values and their corresponding time intervals. Execution time is measured as wall-clock time. Table 6 reports the measured batch time and energy. While our experiments are executed on NVIDIA L40S GPUs, we use measurements collected on the Jetson Orin Nano to estimate the per-client training time and energy consumption for large-scale federated learning.

Table 6. Client-side batch time and batch energy, measured on NVIDIA Jetson Orin Nano using ResNet-18 with batch size 64 and input dimension $32 \times 32 \times 3$.

	Clean FL		Robust FL		
	FedAvg	FedProx	FedAFA	FedAugMix	FedPrime
Batch Time (s)	0.39	0.41	0.75	1.02	8.28
Batch Energy (J)	1.43	1.53	2.76	4.15	8.20

9.2. Time Model

Since multiple clients perform training in parallel during each global round, the training time at round i can be estimated as

$$\mathcal{T}_{\text{tr}}^{i} = i \times t_{\text{batch}} \times b_{\text{avg}} \tag{27}$$

where t_{batch} is the profiled batch execution time and b_{avg} is the average number of batches processed by a client. b_{avg} is computed as the average number of samples per client divided by the batch size:

$$b_{\text{avg}} = \frac{\sum_{k=1}^{K} |D_k|}{K \times \text{batch\_size}} \tag{28}$$

where K is the number of clients and $|D_k|$ is denotes the number of samples held by client k.

9.3. Energy Model

We consider the total compute energy consumed by clients during federated training, which can be estimated for global round i as

$$\mathcal{E}_{\text{tr}}^{i} = i \times r \times K \times e_{\text{batch}} \times b_{\text{avg}} \tag{29}$$

where e_{batch} denotes the profiled per-batch energy consumption, r denotes the client participation rate, K is the number of clients, and b_{avg} is computed using Eq. 28.

10. CIFAR-10-C Results

10.1. CIFAR-10-C Corruptions

CIFAR-10-C [14] is a standard benchmark for evaluating model robustness to common image corruptions. Table 7 lists the corruption types included in the benchmark along with the abbreviations used throughout the paper.

10.2. Full CIFAR-10-C Results

Section 5.2 evaluates popular FL methods and their DART-enhanced variants on the CIFAR-10-C robustness benchmark. Table 8 extends the results by reporting performance across all corruption types. As discussed earlier, integrating DART consistently improves robustness on CIFAR-10-C at zero client-side overhead and at varying degrees of heterogeneity. Notably, DART improves performance across most corruption categories, with the exception of brightness and saturation, which introduce minimal visual distortion and thus exhibit smaller robustness gains.

11. Ablation Studies

11.1. Generalization Across Model Architectures

To demonstrate DART on various deep learning models, we conduct simulations for a 10 client FL setup with a fixed time budget of 1000s and we compare DART enhanced FedAvg clean, robust and average accuracy against FedAvg and FedAugMix for ResNet-18, MobileNet and VGG-16. Table 9 shows that DART enhances the robust accuracy of

Table 7. CIFAR-10-C corruption types and abbreviations used throughout the paper.

Corruption type	Abbreviation
Brightness	bright
Contrast	contr
Defocus blur	defoc
Elastic transform	elas
Fog	fog
Frost	frost
Gaussian blur	g_blur
Gaussian noise	g_noise
Glass blur	glass
Impulse noise	imp
JPEG compression	jpeg
Motion blur	motion
Pixelate	pix
Saturate	sat
Shot noise	shot
Snow	snow
Spatter	spatt
Speckle noise	speck
Zoom blur	zoom

FedAvg by 4.3% for a slight drop of 1.7% in clean accuracy on average. Additionally, DART enhanced FedAvg outperforms FedAugMix in both clean and robust accuracy by an average of 18.9% and 13.8%, respectively. Most notably, DART performs best with VGG-16 where the gains are most significant (29.2% in $\mathcal{A}_{\text{cln}}$, 24.3% in $\mathcal{A}_{\text{rob}}$). These results indicate that DART-enhanced FL generalizes consistently across model architectures while maintaining a favorable balance between clean and robust accuracy, making it well-suited for deployment in resource-constrained environments.

Fig. 4 shows that DART-enhanced FedAvg consistently outperforms both FedAvg and FedAugMix in robust accuracy ($\mathcal{A}_{\text{rob}}$) under constrained time and energy budgets. Across the full range of time and energy, DART-enhanced FL achieves higher robustness than FedAvg. When the per-client time budget is below 2200s or the energy budget is below 25kJ, DART also surpasses FedAugMix, highlighting its effectiveness in resource-constrained settings. Moreover, DART-enhanced FL reaches a target robust accuracy of 80% on ResNet-18 with the lowest time and energy cost among all methods. In contrast, FedAugMix requires substantially more computation to converge, limiting its practicality for deployment on resource-constrained devices. All methods are evaluated with 10 clients over 200 global rounds, with DART applied to the FedAvg model at each round. Similar trends are observed for MobileNet (Fig. 5) and VGG-16 (Fig. 6), demonstrating DART's generalization across model architectures.

11.2. CIFAR-100 Results

To showcase the benefit of using DART with different client datasets, we provide the following results that utilize CIFAR-100 on the clients and CIFAR-10 on the server. The number of clients used is 10.

Analyzing the CIFAR-100 training curves of ResNet-18 in Fig. 7, we notice that DART enhanced FedAvg achieves the best robust accuracy when time and energy are below 2500s and 30kJ, respectively. Within these intervals, DART also significantly outperforms FedAugMix in clean accuracy, while incurring a drop relative to FedAvg. As expected, FedAugMix requires substantially more time and energy to converge. DART provides the most robust model under resource constraints, even when CIFAR-100 is the client dataset.

These experiments demonstrate that the advantages of DART are not limited to a specific client dataset. By maintaining strong clean and robust performance across diverse data distributions, DART proves to be broadly applicable and adaptable, reinforcing its potential as a versatile solution for robust federated learning in scenarios constrained by compute, memory, time, and energy.

11.3. Impact of Robustification Period

We experimentally investigate the impact of the robustness-enhancement period T_{DART} on utility and robustness. Additional DART iterations can be executed since the server is not resource-limited and this imposes no further cost on clients.

Fig. 8 shows that reducing the robustification period T_{DART} has minimal impact on clean accuracy. In general, smaller values of T_{DART}, i.e., more frequent DART updates, leads to higher robust accuracy at no additional client training time and energy. For example, ResNet-18 robust accuracy improves from 81.98% to 83.39% when T_{DART} decreases from 200 to 25. Similar trends can be observed for MobileNet and VGG-16, where robust accuracy increases by $+1\%$ and $+0.72\%$ respectively. These findings suggest that robust accuracy can be increased by performing more frequent DART updates, however, in cases where the server cannot be overloaded, applying DART once at the final global update yields competitive performance.

11.4. DART Parameter Sweeps

In DART, performance is sensitive to the loss-weighting parameter α and the mixture width S. Fig 9 reports the effect of varying α and S on robust accuracy. To maximize robustness, we observe that α values in the range of 8–10 yield the strongest performance. Increasing S introduces more complex data augmentation but also increases server-side computation. Based on this trade-off, we select $\alpha = 10$ and $S = 3$ to achieve strong robustness while limiting server overhead.

Table 8. Impact of DART in enhancing the robustness of diverse standard FL methods. Clean accuracy is measured on CIFAR-10 ($\mathcal{A}_{\text{cln}}$), and robust accuracy on CIFAR-10-C ($\mathcal{A}^{\text{C}}_{\text{rob}}$) for ResNet-18. Accuracies on all CIFAR-10-C corruptions are shown. DART significantly enhances the robustness (avg. 4.3%) of standard FL methods while incurring a small drop (avg. 1.6%) in clean accuracy at no additional computational cost for the clients.

Method	$\mathcal{A}_{\text{cln}}$(%)	$\mathcal{A}^{\text{C}}_{\text{rob}}$(%)	bright (%)	contr (%)	defoc (%)	elas (%)	fog (%)	frost (%)	g_blur (%)	g_noise (%)	glass (%)	imp (%)	jpeg (%)	motion (%)	pix (%)	sat (%)	shot (%)	snow (%)	spatt (%)	speck (%)	zoom (%)
$\alpha_{\text{IID}} = 100$																					
FedAvg	90.3 ±0.9	70.5 ±0.5	88.9 ±0.9	68.2 ±2.3	75.8 ±2.4	79.2 ±1.3	82.5 ±1.4	72.2 ±1.7	67.6 ±3.2	47.8 ±6.6	46.0 ±3.3	55.6 ±4.2	77.7 ±2.1	71.8 ±3.7	73.2 ±2.5	87.1 ±1.1	58.3 ±4.8	75.2 ±1.3	79.6 ±1.1	60.6 ±4.1	71.6 ±2.4
FedDyn	85.0 ±1.0	67.7 ±1.3	83.1 ±1.1	63.7 ±2.4	73.7 ±2.9	74.6 ±2.1	76.9 ±2.0	68.1 ±2.7	66.9 ±3.1	53.5 ±6.2	43.3 ±6.9	56.9 ±2.6	75.8 ±1.1	65.9 ±4.2	66.0 ±2.0	81.8 ±1.2	61.3 ±5.0	68.4 ±2.0	75.5 ±1.1	62.5 ±4.6	68.3 ±3.4
FedNova	86.5 ±0.5	69.5 ±0.8	84.5 ±0.4	61.9 ±0.9	75.0 ±0.4	76.6 ±0.4	76.4 ±0.4	69.1 ±1.3	68.4 ±0.9	56.4 ±2.0	45.6 ±2.4	58.9 ±0.9	77.9 ±0.7	68.4 ±0.9	71.3 ±2.0	83.5 ±0.7	63.1 ±1.6	70.9 ±0.5	78.6 ±0.6	63.7 ±1.5	70.2 ±0.9
FedProx	**91.4** ±0.3	72.7 ±1.0	**89.9** ±0.4	73.5 ±1.6	79.1 ±1.0	81.0 ±1.0	**84.7** ±0.7	73.7 ±1.3	70.1 ±1.8	51.7 ±2.8	45.5 ±2.5	60.7 ±2.1	79.4 ±0.7	73.8 ±1.1	71.5 ±2.3	**88.5** ±0.4	61.8 ±2.5	77.2 ±1.1	82.4 ±1.0	64.0 ±2.3	73.5 ±1.5
FedAvg+DART	89.2 ±0.6	77.4 ±0.9	87.4 ±0.9	73.8 ±2.1	84.8 ±1.2	81.2 ±1.3	82.5 ±1.9	75.2 ±0.9	82.6 ±1.5	64.0 ±2.3	57.7 ±2.1	69.2 ±1.8	79.2 ±1.1	80.1 ±2.3	**77.8** ±1.8	86.2 ±0.6	71.7 ±1.5	77.2 ±1.2	82.8 ±1.1	73.8 ±1.5	82.5 ±1.4
FedDyn+DART	83.2 ±1.8	71.2 ±1.2	81.2 ±1.3	67.4 ±1.9	78.6 ±1.6	74.8 ±1.2	75.9 ±1.2	67.9 ±1.4	76.2 ±1.2	59.2 ±3.0	51.5 ±4.7	65.2 ±1.4	74.6 ±3.7	72.8 ±2.2	70.3 ±7.6	80.2 ±2.0	66.2 ±2.0	69.9 ±1.6	76.8 ±2.6	68.0 ±1.9	75.7 ±0.7
FedNova+DART	85.1 ±0.6	72.6 ±0.8	82.8 ±0.5	66.7 ±1.4	81.0 ±0.8	77.5 ±0.8	76.1 ±0.9	68.1 ±1.1	78.7 ±0.9	60.0 ±1.0	51.0 ±1.6	64.5 ±1.0	77.2 ±0.8	74.3 ±1.4	76.2 ±0.7	82.3 ±0.6	67.0 ±0.8	70.9 ±0.8	78.8 ±0.6	68.6 ±0.9	78.1 ±0.9
FedProx+DART	90.2 ±0.4	**79.0** ±0.8	88.6 ±0.6	**77.7** ±2.0	**86.1** ±0.7	**82.4** ±0.6	84.2 ±1.4	**77.3** ±1.4	**83.9** ±0.7	**65.6** ±2.7	**57.9** ±3.6	**72.6** ±1.9	**80.9** ±0.4	**81.8** ±1.0	77.7 ±1.3	87.3 ±0.8	**73.7** ±1.9	**79.2** ±1.0	**84.8** ±0.8	**76.0** ±1.5	**83.5** ±0.8
$\alpha_{\text{IID}} = 10$																					
FedAvg	89.5 ±0.9	69.6 ±1.8	88.1 ±0.9	68.6 ±1.7	76.0 ±1.0	78.3 ±1.5	82.4 ±0.8	69.7 ±2.5	68.0 ±1.1	46.6 ±7.4	43.6 ±3.2	54.3 ±5.9	77.2 ±2.0	72.6 ±1.1	72.2 ±4.1	86.4 ±1.3	56.6 ±6.3	74.1 ±1.9	78.8 ±1.5	58.4 ±6.0	71.1 ±2.2
FedDyn	83.9 ±1.7	66.7 ±1.1	82.0 ±1.8	60.9 ±4.0	72.6 ±1.8	72.6 ±2.0	75.1 ±2.3	65.9 ±1.9	66.5 ±1.7	52.5 ±6.6	43.5 ±6.8	57.2 ±2.4	73.9 ±1.8	66.6 ±0.6	67.1 ±5.1	80.9 ±1.5	59.7 ±4.9	67.4 ±1.5	74.9 ±1.7	60.6 ±4.2	67.1 ±1.7
FedNova	85.7 ±0.3	67.9 ±0.5	83.6 ±0.4	60.1 ±1.3	73.6 ±0.9	75.1 ±0.7	74.7 ±1.1	67.1 ±0.5	66.5 ±1.2	54.2 ±1.6	44.6 ±1.0	57.3 ±1.8	76.8 ±0.6	65.6 ±0.7	70.0 ±1.7	82.7 ±0.6	61.1 ±1.6	69.2 ±0.3	77.7 ±0.7	61.7 ±1.5	68.4 ±1.1
FedProx	**90.6** ±0.4	71.9 ±1.0	**89.2** ±0.5	71.9 ±2.1	79.4 ±1.0	80.6 ±0.5	**83.6** ±1.1	71.7 ±0.9	71.2 ±2.1	50.0 ±4.2	45.6 ±3.1	59.1 ±2.4	77.8 ±0.8	73.9 ±1.1	71.2 ±1.0	**87.9** ±0.2	59.7 ±3.1	75.6 ±0.9	82.0 ±1.0	61.6 ±2.6	74.9 ±1.5
FedAvg+DART	88.3 ±0.7	75.8 ±1.5	86.4 ±1.2	71.8 ±1.0	83.4 ±1.2	79.5 ±1.3	81.6 ±0.9	73.2 ±2.6	81.1 ±1.5	62.1 ±3.9	54.7 ±4.1	67.4 ±2.4	78.0 ±0.8	78.9 ±1.7	**77.6** ±1.0	85.7 ±0.8	70.0 ±3.1	75.2 ±1.7	81.2 ±1.4	72.1 ±2.8	81.1 ±1.3
FedDyn+DART	82.7 ±1.4	69.5 ±2.4	80.0 ±2.1	64.1 ±3.1	77.9 ±1.8	73.8 ±1.8	74.1 ±2.6	64.7 ±3.8	75.6 ±2.0	56.8 ±4.4	47.8 ±4.1	63.0 ±3.5	73.6 ±2.3	71.4 ±3.2	70.8 ±1.9	80.1 ±1.4	63.7 ±3.9	67.6 ±2.9	76.0 ±1.9	65.5 ±3.7	74.6 ±1.8
FedNova+DART	84.7 ±0.3	71.8 ±0.4	82.3 ±0.2	66.7 ±0.5	80.4 ±0.5	76.5 ±0.5	76.0 ±0.4	67.3 ±0.7	78.1 ±0.6	58.2 ±1.6	48.9 ±1.4	64.0 ±0.4	76.3 ±0.6	74.2 ±0.6	75.1 ±1.3	82.0 ±0.4	65.5 ±1.2	69.8 ±0.5	78.1 ±0.4	67.3 ±1.2	77.2 ±0.6
FedProx+DART	88.7 ±0.6	**77.2** ±2.0	87.0 ±1.3	**74.7** ±3.7	**84.6** ±0.8	80.5 ±1.0	82.1 ±2.5	**74.5** ±3.4	**82.4** ±1.0	**64.5** ±4.3	**56.8** ±4.3	**70.7** ±2.5	**78.0** ±2.6	**80.0** ±2.0	76.6 ±2.4	86.3 ±0.7	**72.0** ±3.7	**76.7** ±2.7	**83.1** ±1.0	**74.1** ±3.6	**82.1** ±1.2
$\alpha_{\text{IID}} = 1$																					
FedAvg	87.8 ±1.4	65.6 ±1.9	85.7 ±1.0	63.8 ±1.8	73.0 ±1.4	73.5 ±2.1	77.3 ±2.4	65.4 ±4.0	65.2 ±1.8	42.1 ±4.8	39.2 ±4.3	50.1 ±3.1	71.9 ±2.4	65.4 ±2.7	70.7 ±4.8	84.4 ±1.3	52.3 ±5.0	69.6 ±2.6	75.1 ±1.5	54.5 ±5.0	66.4 ±2.5
FedDyn	80.7 ±1.0	64.3 ±1.1	78.4 ±2.1	58.9 ±2.0	69.9 ±2.3	70.7 ±2.2	71.8 ±1.7	62.6 ±2.8	63.9 ±3.1	52.7 ±4.1	44.1 ±5.6	53.9 ±2.5	72.7 ±1.2	62.6 ±3.9	63.0 ±3.0	77.0 ±1.5	58.8 ±3.3	64.4 ±1.7	71.8 ±1.4	59.2 ±2.9	65.2 ±3.6
FedNova	82.7 ±0.6	64.9 ±0.8	80.3 ±0.5	55.3 ±0.8	69.3 ±0.9	70.4 ±0.8	69.0 ±1.2	64.3 ±0.4	62.8 ±1.0	54.1 ±1.6	45.6 ±1.7	55.3 ±1.3	74.2 ±0.9	60.1 ±1.3	68.3 ±1.0	79.5 ±0.5	60.2 ±1.4	66.4 ±0.9	74.9 ±1.0	60.4 ±1.4	63.4 ±0.7
FedProx	**88.9** ±1.1	70.2 ±2.0	**86.8** ±0.8	68.1 ±1.5	77.1 ±2.1	**77.4** ±1.5	**80.2** ±1.2	69.4 ±2.9	69.3 ±3.0	51.7 ±5.9	42.9 ±5.7	58.7 ±5.0	76.6 ±1.7	70.0 ±3.4	71.4 ±2.2	**85.9** ±1.1	60.8 ±4.7	**72.9** ±1.9	79.1 ±2.1	62.7 ±3.9	72.2 ±2.3
FedAvg+DART	83.2 ±3.2	68.4 ±5.5	80.1 ±3.5	63.9 ±5.7	76.1 ±4.4	70.5 ±5.5	72.7 ±5.8	64.8 ±5.4	72.8 ±5.0	56.2 ±7.0	46.6 ±9.2	60.8 ±6.8	71.6 ±5.3	67.8 ±6.3	73.0 ±3.7	79.7 ±3.5	63.0 ±6.9	67.6 ±6.8	75.1 ±4.9	64.8 ±6.9	71.8 ±5.3
FedDyn+DART	79.1 ±1.5	67.2 ±1.6	77.1 ±1.7	63.1 ±3.3	74.4 ±2.5	71.1 ±2.3	70.6 ±3.2	63.5 ±3.2	72.2 ±2.8	56.2 ±3.7	48.8 ±3.6	59.5 ±0.9	71.8 ±1.4	67.9 ±3.0	69.7 ±3.2	75.9 ±2.3	62.2 ±3.1	65.0 ±1.8	72.3 ±1.8	63.2 ±2.8	72.4 ±2.5
FedNova+DART	80.3 ±0.7	67.6 ±1.1	77.4 ±1.0	59.4 ±1.9	74.6 ±1.1	71.1 ±1.5	69.1 ±1.4	63.6 ±1.6	71.9 ±1.5	58.2 ±1.5	**50.1** ±1.6	60.3 ±1.6	72.7 ±1.0	67.1 ±1.8	72.1 ±1.6	77.4 ±0.6	63.8 ±1.2	65.8 ±1.1	73.8 ±0.9	64.7 ±1.2	71.3 ±1.9
FedProx+DART	87.1 ±0.9	**73.9** ±1.5	84.4 ±0.7	**71.0** ±1.8	**82.0** ±1.2	**77.4** ±1.3	78.4 ±1.6	**70.6** ±1.7	**79.1** ±1.3	**60.8** ±3.3	49.7 ±2.4	**67.4** ±2.0	**77.2** ±2.7	**74.8** ±2.4	**75.9** ±3.8	83.8 ±0.9	**68.5** ±2.8	**72.9** ±1.4	**80.7** ±1.4	**70.5** ±2.8	**78.6** ±1.3
$\alpha_{\text{IID}} = 0.1$																					
FedAvg	49.0 ±14.6	37.1 ±10.7	48.1 ±14.5	34.7 ±10.0	36.4 ±14.4	35.3 ±15.8	40.4 ±14.0	38.6 ±10.3	31.9 ±12.7	32.7 ±5.7	26.6 ±6.3	36.4 ±7.6	37.8 ±12.8	33.8 ±13.1	40.0 ±5.0	47.5 ±14.0	37.2 ±7.4	36.9 ±10.9	42.3 ±12.6	38.6 ±7.8	29.7 ±14.8
FedDyn	34.7 ±7.5	31.6 ±6.4	30.7 ±6.0	19.4 ±1.4	33.0 ±6.9	32.6 ±6.8	24.4 ±4.8	25.9 ±4.5	32.3 ±6.6	34.6 ±7.4	33.4 ±7.0	34.1 ±6.8	34.5 ±7.4	32.6 ±6.0	34.2 ±7.4	31.6 ±6.6	34.6 ±7.5	31.1 ±6.9	34.3 ±7.3	34.6 ±7.6	32.0 ±6.5
FedNova	54.3 ±3.8	45.0 ±3.1	50.5 ±4.5	26.1 ±5.8	44.0 ±3.5	44.3 ±4.0	34.7 ±7.5	43.4 ±4.8	40.6 ±3.4	48.4 ±1.3	43.9 ±2.4	45.2 ±0.9	52.5 ±3.3	37.9 ±3.1	52.1 ±3.7	51.1 ±3.7	50.2 ±1.9	48.8 ±2.8	51.6 ±2.8	50.3 ±2.2	40.1 ±3.2
FedProx	**68.0** ±1.3	56.7 ±0.8	**64.6** ±1.0	43.9 ±4.2	56.4 ±1.5	56.7 ±1.6	54.6 ±2.5	56.7 ±2.1	52.4 ±1.5	55.3 ±2.1	**50.6** ±2.7	52.9 ±2.2	**63.7** ±1.1	51.4 ±3.9	**63.9** ±2.7	**63.9** ±1.9	59.1 ±2.1	**58.0** ±1.5	**63.2** ±1.8	59.3 ±2.3	51.4 ±1.7
FedAvg+DART	43.0 ±15.8	32.9 ±14.7	41.2 ±16.0	29.9 ±12.5	36.9 ±15.9	33.3 ±16.4	35.2 ±15.4	30.1 ±14.6	34.8 ±16.0	26.7 ±14.8	20.1 ±13.4	29.6 ±15.5	34.4 ±13.3	34.3 ±16.7	36.2 ±7.4	41.6 ±15.4	30.2 ±15.1	30.3 ±15.2	36.2 ±15.8	31.3 ±14.9	33.0 ±16.1
FedDyn+DART	23.8 ±15.3	21.9 ±13.6	21.1 ±13.9	16.0 ±7.6	22.9 ±14.7	22.6 ±14.3	18.8 ±10.8	18.2 ±11.4	22.5 ±14.4	23.1 ±14.4	22.6 ±14.1	22.7 ±13.8	23.6 ±15.2	22.2 ±13.5	23.3 ±15.0	22.0 ±13.9	23.3 ±14.7	21.4 ±14.1	23.2 ±14.7	23.3 ±14.7	22.6 ±14.6
FedNova+DART	48.0 ±6.0	40.5 ±5.1	43.4 ±6.7	26.2 ±5.9	42.0 ±5.7	40.8 ±5.9	32.7 ±7.3	35.7 ±5.8	40.0 ±5.3	42.7 ±4.0	37.3 ±4.0	40.3 ±4.0	46.0 ±5.6	37.7 ±4.7	45.6 ±5.8	44.8 ±6.0	44.0 ±4.5	40.5 ±5.0	44.9 ±5.3	44.0 ±4.7	40.0 ±5.6
FedProx+DART	66.9 ±4.3	**58.7** ±2.1	63.9 ±2.6	**51.0** ±0.0	**61.4** ±5.3	**59.3** ±4.8	**58.3** ±0.7	**57.1** ±2.0	**59.6** ±5.2	**55.9** ±1.6	49.8 ±0.0	**53.5** ±0.2	62.7 ±4.2	**59.1** ±1.4	63.3 ±7.4	62.8 ±3.0	**59.4** ±0.2	56.7 ±1.6	61.4 ±4.2	**60.1** ±0.2	**59.4** ±4.2

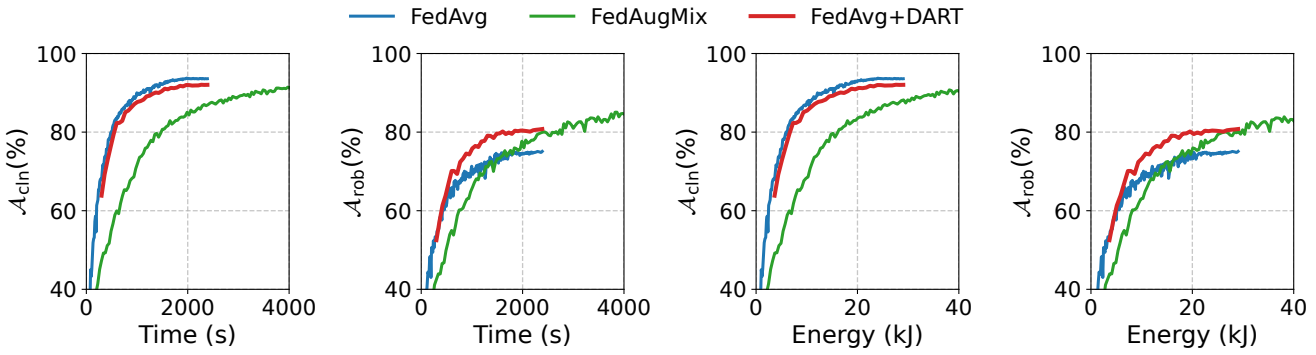


Figure 4. CIFAR-10 clean and CIFAR-10-C robust accuracy, $\mathcal{A}_{\text{cln}}$ and $\mathcal{A}_{\text{rob}}$, under FedAvg, FedAugMix, and FedAvg+DART as a function of time and energy. The model used is ResNet-18, the server dataset is CIFAR-100, and 10 clients are deployed. FedAvg+DART takes the least energy and time to reach $\mathcal{A}_{\text{rob}} = 80\%$.

11.5. Loss Component Ablations

Table 10 presents an ablation study on the components $\mathcal{L}_{\text{c}}$ and $\mathcal{L}_{\text{d}}$ of the DART loss (1) as compared to standard clean training. Applying DART without the consistency loss $\mathcal{L}_{\text{c}}$ results in a model with clean and robust accuracy comparable to the original pre-trained model. In contrast, removing the distillation loss $\mathcal{L}_{\text{d}}$ leads to significant drops in both utility and robustness. Only when both loss components are present does DART maintain clean accuracy (with a minor 1.5% drop) while improving robustness by 5.7%.

These findings are consistent with those in Section 4: $\mathcal{L}_{\text{d}}$ preserves clean accuracy, while $\mathcal{L}_{\text{c}}$ promotes consistency, which translates to improved robust accuracy with minimal loss in clean accuracy, when both loss components are jointly optimized.

12. Using Synthetic Data on the Server

When high-quality server data is unavailable, generative models can be used to supply the server dataset for DART.

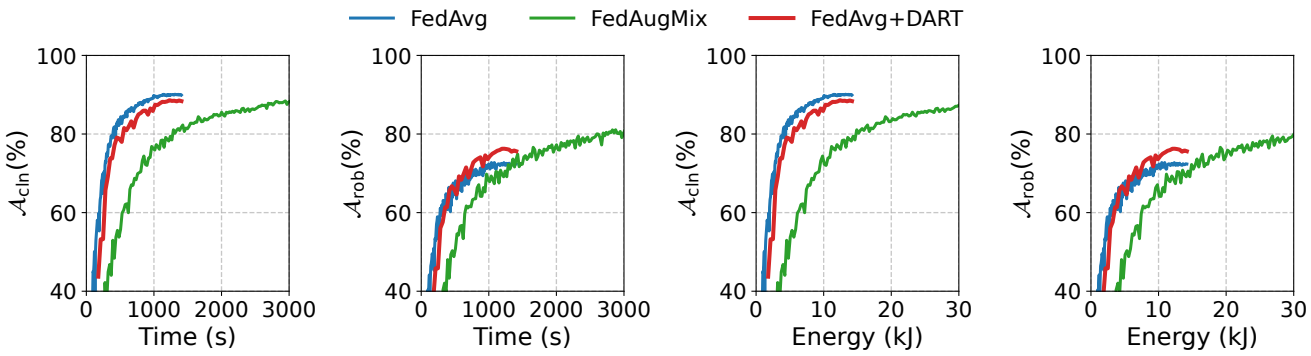


Figure 5. CIFAR-10 clean and CIFAR-10-C robust accuracy, $\mathcal{A}_{\text{cln}}$ and $\mathcal{A}_{\text{rob}}$, under FedAvg, FedAugMix, and FedAvg+DART as a function of time and energy. The model used is MobileNet, the server dataset is CIFAR-100, and 10 clients are deployed.

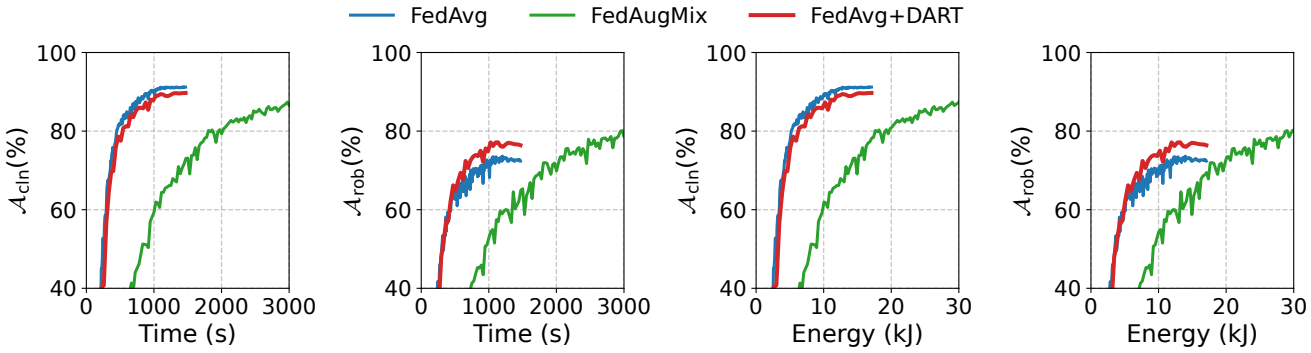


Figure 6. CIFAR-10 clean and CIFAR-10-C robust accuracy, $\mathcal{A}_{\text{cln}}$ and $\mathcal{A}_{\text{rob}}$, under FedAvg, FedAugMix, and FedAvg+DART as a function of time and energy. The model used is VGG-16, the server dataset is CIFAR-100, and 10 clients are deployed.

Table 9. Accuracy at iso-time of $\approx$ 1000s for different models. The value of G is set to meet the time budget. FedAvg+DART achieves the highest robustness at zero client overhead.

Model	Method	G	$\mathcal{A}_{\text{cln}}$ (%)	$\mathcal{A}_{\text{rob}}$ (%)	$\mathcal{A}_{\text{avg}}$ (%)
ResNet-18	FedAvg	83	**90.00**	71.32	80.66
	FedAugMix	27	72.22	65.97	69.10
	FedAvg+DART	83	87.62	**75.85**	**81.74**
MobileNet	FedAvg	141	**89.83**	72.38	81.11
	FedAugMix	43	76.30	68.41	72.36
	FedAvg+DART	141	88.07	**75.73**	**81.90**
VGG-16	FedAvg	135	**89.90**	72.08	80.99
	FedAugMix	36	59.31	52.83	56.07
	FedAvg+DART	135	88.81	**77.14**	**82.98**

Table 11 shows that using BigGAN [5]-generated data with FedAvg+DART still yields substantial robustness gains, improving $\mathcal{A}_{\text{rob}}$ by 4.8

Table 10. CIFAR-10 clean and CIFAR-10-C robust accuracy results for FedAvgand FedAvg+DART with custom DART. The model used is ResNet-18. To improve robustness, DART must include both the consistency loss $\mathcal{L}_{\text{c}}$ and the distillation loss $\mathcal{L}_{\text{d}}$.

Method	$\mathcal{A}_{\text{cln}}$ (%)	$\mathcal{A}_{\text{rob}}$ (%)	$\mathcal{A}_{\text{avg}}$ (%)
Clean training	**93.58**	75.09	84.34
DART w/o $\mathcal{L}_{\text{c}}$	93.46	75.36	84.41
DART w/o $\mathcal{L}_{\text{d}}$	19.24	17.53	18.39
DART	92.08	**80.79**	**86.44**

13. Accounting for Server-side Costs in DART-enhanced FL

DART offloads the resource intensive robust training to the server. Fig. 10 shows that the large server side resources and parallelism reduce overall training time when DART is applied to FedAvg compared to other robust FL works. Crucially, client-side time and energy remain comparable

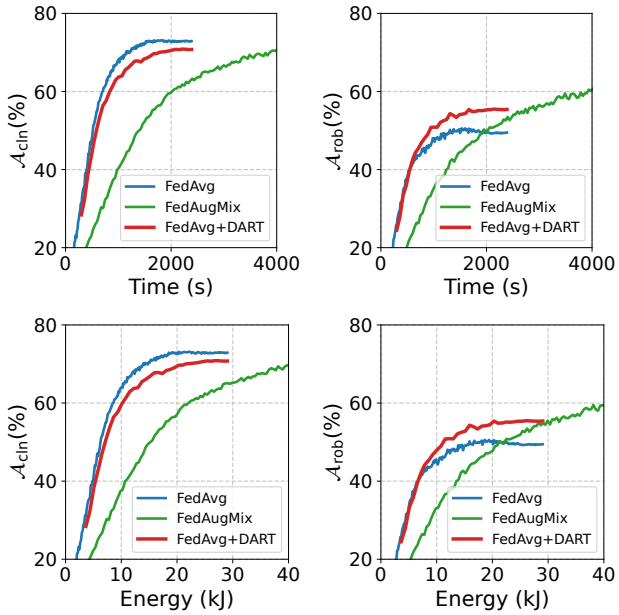


Figure 7. CIFAR-100 clean and CIFAR-100-C robust accuracy, $\mathcal{A}_{\text{cln}}$ and $\mathcal{A}_{\text{rob}}$, under FedAvg, FedAugMix, and FedAvg+DART as a function of time and energy. The model used is ResNet-18, the server dataset is CIFAR-10, and 10 clients are deployed.

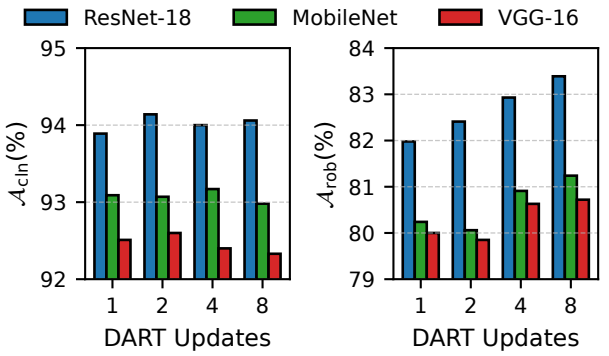


Figure 8. FedAvg+DART CIFAR-10 clean and CIFAR-10-C robust accuracy for different numbers of DART updates. The number of global updates, E, is set to 200. The number of clients is 4 and the server dataset is CIFAR-100.

to non-robust methods as shown in 4.

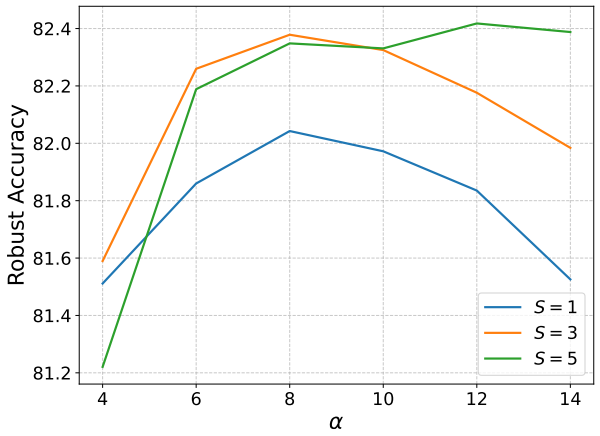
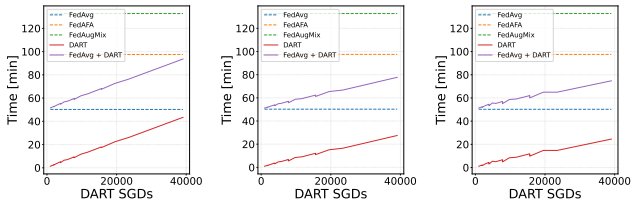


Figure 9. Impact of DART loss weighing parameter α and data augmentation mixture width S on CIFAR-10-C robust accuracy. We choose $\alpha = 10$ and $S = 3$ to maximize robustness and limit the server workload.



(a) 1 GPU (b) 2 GPUs (c) 4 GPUs

Figure 10. Total time versus number of DART SGD updates. DART-enhanced FedAvg maintains lower training time compared to robust FL methods.

Table 11. Impact of DART on CIFAR-10 clean and robust accuracy with BigGAN-generated server data.

Method	α_{iid}	$\mathcal{A}_{\text{cln}}$ (%)	$\mathcal{A}_{\text{rob}}$ (%)
FedAvg	100	89.8	70.8
FedAvg	10	89.5	70.1
DART+FedAvg	100	86.4	73.7
DART+FedAvg	10	86.3	74.9