SCIENCE CHINA
Information Sciences

• LETTER •

July 2026, Vol. 69, Iss. 7, 179102:1–179102:2
https://doi.org/10.1007/s11432-024-4781-5

FDDME: feature divergence-directed data-free model extraction

Ruinan MA[1], Yu-An TAN[1], Yajie WANG[1*], Zuobin YING[2], Liehuang ZHU[1] & Jianfeng MA[3]

[1]*School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing 100081, China*
[2]*Faculty of Data Science, City University of Macau, Macao 999078, China*
[3]*School of Cyber Engineering, Xidian University, Xi'an 710071, China*

Received 18 March 2025/Revised 28 December 2025/Accepted 13 January 2026/Published online 6 May 2026

Citation Ma R N, Tan Y-A, Wang Y J, et al. FDDME: feature divergence-directed data-free model extraction. Sci China Inf Sci, 2026, 69(7): 179102, https://doi.org/10.1007/s11432-024-4781-5

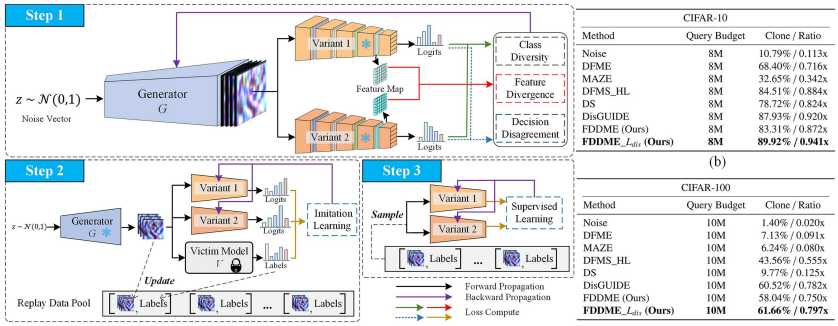
Deep neural networks (DNNs) have been widely deployed in practical applications. To lower the usage barrier and free users from heavy computing and data demands, service providers train DNNs on private datasets and offer machine learning as a service (MLaaS) via cloud APIs in a pay-per-query manner. The query-response mechanism fundamental to MLaaS inadvertently allows an adversary to perform a model extraction attack by issuing carefully crafted queries to systematically probe the victim's response behaviors and replicate its functionality in a clone model. Beyond a direct infringement of intellectual property, such a clone model can further facilitate downstream attacks against the victim model, such as adversarial attacks [1], thereby undermining model integrity. In this work, we focus on data-free model extraction (DFME) [2], where the adversary conducts an attack without relying on any accessible data from the victim's training distribution, and instead synthesizes queries and trains the clone using the victim's feedback. Under limited query budgets and restricted feedback such as hard-label responses, accurate and query-efficient model extraction is crucial for both risk assessment and the development of effective countermeasures.

DFME typically adopts a cooperative optimization paradigm between a generator and the clone model: the generator is optimized to synthesize informative queries that maximize the learning utility of each victim interaction, while the clone is trained to mimic the victim's feedback on the synthesized queries. To improve query efficiency, representative approaches such as DisGUIDE and DS [3,4] introduce two simultaneously learned clones and incorporate their inter-clone disagreement into the generator's optimization as an internal signal to steer query synthesis: disagreement on the same input indicates that at least one clone deviates from the victim's behavior, making such inputs more informative for accelerating imitation. Nevertheless, this strategy essentially performs a posteriori passive difference measurement at the output level and underutilizes the white-box advantages in DFME, where clone models are fully accessible to the adversary. In particular, relying solely on decision-space feedback can provide weak guidance when the victim returns restricted feedback such as hard-label responses. Some studies have shown that internal feature characteristics can affect model behavior [5]. These observations motivate us to ask: how can we leverage the white-box internal information of the clone model to construct a self-enhancing signal that guides query synthesis more effectively?

To answer this, we propose feature divergence-directed data-free model extraction (FDDME). FDDME builds a dynamic shadow model architecture by deriving two variant models from a single base architecture. It actively constructs model representation differences in the feature space and embeds them as a prior to encourage inconsistent decisions between the variants. Specifically, two variants are used to form differentiated feature patterns for the same batch of synthetic images, yielding a dual-difference coupling with output-layer disagreement to guide query synthesis, providing additional optimization incentives for the generator. We measure batch-level feature divergence with maximum mean discrepancy (MMD), a kernel-based distributional metric that captures discrepancies between the variants' same-layer feature distributions, and promote class balance via the variants' consensus entropy. The two variants are integrated via soft voting to form the final clone. Since synthesized queries are inexhaustible and nonrepetitive, we maintain a replay pool to intermittently revisit past queries and mitigate catastrophic forgetting. Complementarily, we propose a robustness reinforcement mechanism that leverages the extracted clone model and generator to improve the victim's adversarial robustness under black-box and data-privacy constraints.

Methodology. We use V to represent the victim model. A generator $G(\cdot;\theta_G)$, and two variants $C_1(\cdot;\theta_1)$ and $C_2(\cdot;\theta_2)$ are obtained by separate random initialization from a single base architecture, where $C_c(\mathbf{x})$ denotes the logits output of variant $c \in \{1,2\}$. Given $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_d)$, the generator produces a query $\mathbf{x} = G(\mathbf{z})$. For a mini-batch $\{\mathbf{x}_i\}_{i=0}^{N-1}$, define the predicted probability of variant c on $\mathbf{x}_i$ as $A_{cik} = \left(\text{softmax}(C_c(\mathbf{x}_i))\right)_k$ for $k \in \{0,\ldots,K-1\}$. We encourage class balance using the variants' consensus entropy $L_{div} = -\sum_{k=0}^{K-1} w_k \log w_k$, where K is the number of classes and $w_k = \frac{1}{N}\sum_{i=0}^{N-1}\frac{1}{2}\sum_{c=1}^{2} A_{cik}$. We further promote decision disagreement by $L_{dis} = -\frac{1}{NK}\sum_{i=0}^{N-1}\sum_{k=0}^{K-1} S(A_{1ik}, A_{2ik})$, where

* Corresponding author (email: wangyajie19@bit.edu.cn)



CIFAR-10

Method	Query Budget	Clone / Ratio
Noise	8M	10.79% / 0.113x
DFME	8M	68.40% / 0.716x
MAZE	8M	32.65% / 0.342x
DFMS_HL	8M	84.51% / 0.884x
DS	8M	78.72% / 0.824x
DisGUIDE	8M	87.93% / 0.920x
FDDME (Ours)	8M	83.31% / 0.872x
FDDME_L_{dis} (Ours)	**8M**	**89.92% / 0.941x**

(b)

CIFAR-100

Method	Query Budget	Clone / Ratio
Noise	10M	1.40% / 0.020x
DFME	10M	7.13% / 0.091x
MAZE	10M	6.24% / 0.080x
DFMS_HL	10M	43.56% / 0.555x
DS	10M	9.77% / 0.125x
DisGUIDE	10M	60.52% / 0.782x
FDDME (Ours)	10M	58.04% / 0.750x
FDDME_L_{dis} (Ours)	**10M**	**61.66% / 0.797x**

(c)

Figure 1 (Color online) (a) The method pipeline involves three sequential steps in each iteration. Step 1: Optimize generator G by calculating the loss between Variant 1 and Variant 2. Step 2: Optimize variant models via imitation learning. Step 3: Update variant models using past data sampled from the replay data pool. The process ends when the query budget reaches the upper limit. We show quantitative comparisons on CIFAR-10 (b) and CIFAR-100 (c) when the victim model only returns hard labels. Each entry is reported as "clone accuracy (%)/accuracy ratio (×)", where the ratio is computed as $\mathrm{Acc}_{\mathrm{clone}}/\mathrm{Acc}_{\mathrm{victim}}$. Best results are in bold.

$S(\cdot,\cdot)$ denotes the standard deviation across the two variants. To exploit internal representations, let $\mathbf{h}_1^{(i)}$ and $\mathbf{h}_2^{(i)}$ be same-depth features of the i-th synthetic image from C_1 and C_2, respectively. We maximize feature-space discrepancy via the Gaussian-kernel MMD-based feature divergence loss:

$$L_{FD} = -\left[\frac{1}{N(N-1)}\sum_{i\neq j} k(\mathbf{h}_1^{(i)}, \mathbf{h}_1^{(j)}) + \frac{1}{N(N-1)}\sum_{i\neq j} k(\mathbf{h}_2^{(i)}, \mathbf{h}_2^{(j)}) - \frac{2}{N^2}\sum_{i,j} k(\mathbf{h}_1^{(i)}, \mathbf{h}_2^{(j)})\right], \tag{1}$$

where N is the batch size and $k(\cdot,\cdot)$ denotes the Gaussian kernel. The generator is optimized by $L_G = \beta L_{FD} - \lambda L_{div} + L_{dis}$.

The variants imitate V on synthesized queries by querying $\mathbf{y} = V(\mathbf{x})$ and minimizing $L_{\mathrm{imitate}}(C_j(\mathbf{x}), \mathbf{y})$ for $j \in \{1,2\}$: with probability feedback, we use $\|C_j(\mathbf{x}) - \hat{\ell}_V(\mathbf{x})\|_1$, where $\hat{\ell}_V(\mathbf{x})$ denotes estimated victim logits inferred from the returned softmax probabilities; with one-hot feedback, we use $\mathrm{CE}(C_j(\mathbf{x}), \mathbf{y})$, where $\mathrm{CE}(\cdot,\cdot)$ is the cross-entropy. To mitigate catastrophic forgetting under streaming synthetic data, we maintain a replay pool $\mathcal{R}$ (FIFO queue) that stores queried pairs $(\mathbf{x}, \mathbf{y})$ and intermittently updates C_1, C_2 on mini-batches sampled from $\mathcal{R}$. The final clone uses soft voting $p(\mathbf{x}) = \frac{1}{2}\sum_{c=1}^{2} \mathrm{softmax}(C_c(\mathbf{x}))$ and predicts $\arg\max_k p_k(\mathbf{x})$. The overall procedure follows Figure 1(a).

For robustness reinforcement, we construct a synthetic adversarial set D_{que} using the trained G and clone: sample $\mathbf{x}' = G(\mathbf{z})$, set pseudo-label $y' = \arg\max_k p_k(\mathbf{x}')$ (optionally keep only samples with $c' = \max_k p_k(\mathbf{x}') \geqslant \tau$), and generate $\mathbf{x}' + \delta$ by a white-box attack on the clone with $\|\delta\| \leqslant \epsilon$. The victim owner then fine-tunes V on private data D_v and D_{que} by optimizing

$$\min_{\theta_V}\Big(\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim D_v}\left[L(\theta_V, \mathbf{x}, \mathbf{y})\right] + \gamma \cdot \mathbb{E}_{(\mathbf{x}',y')\sim D_{que}}\left[\max_{\|\delta\|\leqslant\epsilon} L(\theta_V, \mathbf{x}'+\delta, y')\right]\Big), \tag{2}$$

where γ controls the robustness term and ϵ bounds the perturbation. Details and settings are in Appendixes C and D.

Experiments. Figures 1(b) and (c) summarize CIFAR-10/100 comparisons with representative baselines. Our method achieves the strongest cloning accuracy and improves query efficiency: on CIFAR-10, it reaches DisGUIDE's final accuracy with 18.80% fewer queries (6.50M vs. 8M) and attains 89.92 ± 0.20 final accuracy; on CIFAR-100, it requires 12.00% fewer queries (8.80M vs. 10M) and achieves 61.66 ± 1.72 final accuracy. Adding decision disagreement on top of feature divergence (FDDME_L_{dis} vs. FDDME) further improves performance, confirming their complementarity. Comprehensive experiments, including in-depth mechanism analysis, attack-defense evaluation, and robustness reinforcement studies, are provided in Appendix D.

Conclusion and discussion. Our method advances practical risk assessment for MLaaS by showing that accurate, query-efficient extraction remains feasible under limited feedback and budgets, and by turning the extracted surrogate into a data-private robustness reinforcement tool for deployed models. More broadly, it highlights a general principle: leveraging white-box signals from evolving surrogates to synthesize higher-information queries. This idea is promising for large visual models, where queries are expensive: one can parameterize the query generator with a generative prior and drive search using surrogate-side representation divergence together with output disagreement on embedding/similarity scores, thereby improving accuracy per query under API constraints.

Acknowledgements This work was supported by Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62402040), National Natural Science Foundation of China (Grant No. U2336201), Funds for International Cooperation and Exchange of the National Natural Science Foundation of China (Grant No. 62561160099), Yunnan Provincial Major Science and Technology Special Plan Projects (Grant No. 202502AD080008), Yunnan Provincial New R&D Institution Cultivation Project (Grant No. 202404BQ040148), and Yunnan Science and Technology Program (Grant No. 202605AK340003).

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Wang K, He X, Wang W, et al. Boosting adversarial transferability by block shuffle and rotation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 24336–24346
2 Truong J, Maini P, Walls R J, et al. Data-free model extraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 4771–4780
3 Rosenthal J, Enouen E, Pham H V, et al. DisGUIDE: disagreement-guided data-free model extraction. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2023. 9614–9622
4 Beetham J, Kardan N, Mian A S, et al. Dual student networks for data-free model stealing. In: Proceedings of the 11th International Conference on Learning Representations, 2023
5 Ahmad O, Béreux N, Baret L, et al. Causal analysis for robust interpretability of neural networks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024. 4673–4682