arXiv:2302.01428v2 [cs.LG] 9 Nov 2023

利用神经切线核和数据集蒸馏理解重建攻击

Noel Loo、Ramin Hasani、Mathias Lechner、AlexanderAmini 和 Daniela Rus麻省理工学院计算机科学与人工智能实验室，美国马萨诸塞州剑桥{loo,rhasani, mlechner, amini,rus}@mit.edu

摘要

现代深度学习需要大量数据，其中可能包含不得泄露的敏感信息或隐私信息。近期研究表明，对于齐次神经网络，仅凭访问训练后的网络参数，就可能重建很大一部分训练数据。尽管已有实验证明这种攻击有效，但对于其有效适用范围——哪些数据点容易被重建——目前仍缺乏充分的形式化理解。在这项工作中，我们首先构建了更强的数据集重构攻击，并证明在无限宽度区域中，该攻击能够恢复整个训练集。随后，我们通过实证研究两层网络上这种攻击的特性，发现其成功与否在很大程度上取决于网络偏离冻结的无限宽度神经切线核极限的程度。接下来，我们研究容易被重建的图像的特性。我们从理论和实证两方面表明，重建图像往往是数据集中的“离群点”，并且这些重建攻击可用于数据集蒸馏：也就是说，我们可以在重建图像上重新训练，并获得较高的预测准确率。

1 引言

研究表明，神经网络在各种任务上表现出色，甚至能够泛化，尽管它们在训练数据上的损失为零（Zhang 等人，2017；2021）。但如果神经网络因安全问题而无法在实践中使用，这种性能便毫无意义。神经网络安全领域的一个根本问题是：训练过程泄露 了多少信息？也就是说，能够访问训练后的模型或模型预测结果的攻击者，能否推断出哪些数据曾被用于训练该模型？理想情况下，我们希望确保模型能够抵御此类攻击。然而在实践中，这一理想往往无法实现。对这一原则的一种严重违背是记忆现象（Arpit 等人，2017；Feldman 和 Zhang，2020b；Feldman，2020；Carlini 等人，2019）：在生成模型中，研究者发现训练后的网络能够在测试时复现训练数据。Haim 等人（2022）展示了更极端的记忆案例：仅凭训练后的网络参数，就有可能恢复训练数据的一个较大子集。

这种攻击的存在引出了许多后续问题：“在什么情况下这种攻击能够成功？”；以及“这些恢复的图像具有什么性质？”本文探讨了 Haim 等人（2022）提出的攻击的一种更强变体，提供了关于数据集重建攻击的新理论见解和实证见解，并回答了上述问题。具体而言，我们提出了以下新**贡献**：

我们设计了一种更强版本的 Haim 等人（2022）的数据集重构攻击，在采用均方误差（MSE）损失训练时，该攻击能够以可证明的方式重构神经切线核（NTK）（Jacot 等，2018）机制下网络的整个训练集。该攻击也适用于有限网络，其成功与否取决于网络偏离 NTK 机制的程度。

我们发现，在我们的攻击下，离群数据点容易被重构，这印证了先前关于这一特性的研究。此外，我们还发现，移除易于重构的图像能够提高预测准确率。

我们从理论上证明并通过实验证明，数据集重构攻击是数据集蒸馏**的一种变体**。 在使用不同范数时，重构损失等于核诱导点（KIP）（Nguyen 等人，2021a;b）数据集蒸馏算法的损失，再加上一个受方差控制的项。此外，我们可以使用恢复的图像重新训练模型，并取得很高的性能。

2 背景与相关工作

机器学习隐私。 大量研究探讨了如何从训练后的模型中提取敏感信息。这会带来问题，因为 HIPAA 和 GDPR 等法规规定了哪些数据可以或不可以发布或使用（美国医疗保险和医疗补助服务中心，1996年；欧盟委员会，2016年）。量化训练样本的影响力，引出了影响函数（Koh 和 Liang，2017）以及成员推断攻击（Shokri 等人，2016；Carlini 等人，2021）这一研究主题；这些方法试图推断特定样本是否被用于训练。防御此类攻击的研究领域是差分隐私，它通过量化并限制模型对训练数据微小变化的敏感度来实现防护（Dwork 等人，2006年；Abadi 等人，2016）。同样，机器反学习 领域试图在训练后移除训练样本的影响力（Bourtoule 等人，2019年）。研究表明，若没有防御技术，训练后的网络会泄露信息（Rigaki 与 Garcia，2020年）。对于生成式语言模型，Carlini 等人（2019年）表明，大语言模型常会逐字复现训练语料库中的样本。模型反演 技术旨在通过查看模型激活来重建训练样本（Fredrikson 等人，2015年；Yang 等人，2019年；He 等人，2019年）。研究表明，在某些情况下，要取得高性能，记忆现象是不可或缺的（Feldman 和 Zhang，2020b；Brown 等人，2021）。

数据集重建。 近期研究（Haim 等人，2022 年）表明，利用神经网络的隐式偏置，可以从训练后的网络中重构出大量训练数据。他们指出，在逻辑损失下训练的齐次神经网络，其参数方向会收敛到以下最大间隔问题的解（Lyu 与 Li，2020年；Ji 与 Telgarsky，2020年）：

$$\arg\min_{\theta'} \frac{1}{2}\|\theta'\|_2^2 \quad \text{s.t.} \quad \forall i \in [n], y_i f_{\theta'}(x_i) \geq 1, \tag{1}$$

其中， $\{x_i, y_i\}$ 是包含图像 x_i 及标签 $y_i \in \{+1, -1\}$的训练集， $f_{\theta'}$（x）表示参数为 θ'的神经网络输出。（Haim 等人，2022 年）表明，通过获取一个训练后的神经网络，并优化图像（以及对偶参数），使其满足最大间隔问题的 Karush–Kuhn–Tucker（KKT）条件，就可以重构训练数据。这是一种会导致训练数据泄露的攻击。在这里，我们考虑一种更强的攻击变体，该变体要求使用均方误差（MSE）损失进行训练。

神经正切核。为了研究神经网络的泛化，我们可以使用神经正切核（NTK）理论（Jacot 等，2018；Arora 等人，2019）。神经正切核理论指出，随着网络宽度趋于无穷，网络的行为类似于在其初始化参数处对网络参数进行一阶泰勒展开（Lee 等人，2019）。此外，由此得到的特征映射和核会收敛至神经正切核，并且该核在整个训练过程中保持不变（Jacot 等，2018；Arora 等人，2019）。因此，宽神经网络类似于核机器；使用支持集 X_S 及标签 y_S 进行均方误差损失训练时，其测试预测为：

$$\hat{y}_T = K_{TS} K_{SS}^{-1} y_S,$$

其中 K 为神经正切核。对于全连接网络，这种核可以非常快速地精确计算（因为它们可化为反余弦核）；但对于更大型的卷积网络，精确计算会慢得多（Arora 等人，2019年；Zandieh 等人，2021）。实践中已发现，网络往往与固定核理论状态相去甚远：训练最初的几个轮次中，经验神经切线核会发生很大变化，之后才趋于稳定（Hanin 和 Nica，2020年；Aitken 和 Gur-Ari，2020年；Fort 等人，2020年；Loo 等人，2022年；Tsilivis 和 Kempe，2022年）。本文利用神经正切核理论，更好地理解这些重建攻击。

数据集蒸馏。数据集蒸馏旨在构建规模更小、能够准确表征较大数据集的合成数据集。具体而言，在规模小得多的蒸馏数据集上训练，所达到的性能可与完整数据集相当，并远高于对数据集进行随机采样的结果（Wang 等人，2018年；Zhao 等人，2021年；Zhao 和 Bilen，2021；Nguyen 等人，2021a;b；Zhou 等人，2022年；Loo 等人，2022a）。此类算法有很多，既包括直接展开计算的办法（Wang 等人，2018年），也包括尝试高效近似与蒸馏数据训练相关的内部展开计算的方法（Zhou 等人，2022年；Loo 等人，2022a；Nguyen 等人，2021b），以及其他启发式方法（Zhao 等人，2021年；Zhao 和 Bilen，2021）。其中一种算法是核诱导点（KIP）（Nguyen 等人，2021a;b），它利用神经正切核理论推导出如下损失：

$$\mathcal{L}_{KIP} = \frac{1}{2}\|y_t - K_{TS}K_{SS}^{-1}y_S\|_2^2.$$

损失表示无限宽度网络对蒸馏图像 X_S 和标签y_S的预测误差，随后会对这些图像和标签进行优化。我们在此讨论数据集蒸馏，因为本文表明，我们的数据集重构攻击是 KIP 的一种泛化形式，而且数据集蒸馏可以用于抵御该攻击。

3 神经切线核重构攻击

Haim 等人（2022）考虑了攻击者只能访问最终训练后的网络参数的情形。这种攻击要求网络是齐次的，并且训练许多轮次直至收敛，使网络在方向上收敛至方程 1 的最终 KKT 点。尽管这为此类攻击提供了很好的概念验证，但这种攻击在理论和实践上都存在若干局限。我们发现，Haim 等人（2022）提出的攻击很脆弱。具体而言，若不进行仔细的超参数调优和**谨慎的网络初始化策略**，我们无法可靠地复现他们的结果。他们的攻击还要求训练直至方向收敛，这**要求网络参数趋于无穷大**，并且要求网络是齐次的。此外，除少数排名靠前的重建图像外，绝大多数重建图像（$>70\%$）质量较差（更深入的讨论见附录A）。本文提出一种兼容提前停止、无需特殊初始化策略的攻击；在某些假设下，该攻击能够保证重建整个训练集，这是**在任何训练情形下首次提供重建保证**。不过，我们的攻击需要访问模型初始化或先前的训练检查点。在许多场景中，自然可以访问初始化或较早的训练检查点，例如从公开模型进行微调，或在联邦学习中，客户端会定期收到模型参数更新。

请注意，我们无法将本文的攻击与其他攻击进行比较，例如梯度泄露攻击（Zhu 等，2019）、成员推断攻击（Shokri 等人，2016）和生成模型攻击（Carlini 等人，2020），因为这些攻击分别需要梯度访问权限（在梯度泄露场景中，这也意味着需要访问参数）、成员推断场景中的特定查询数据点，或生成模型攻击所需的生成模型和查询点。我们的攻击不需要预先了解数据集。有鉴于此，本文所述的攻击为仅基于参数的攻击这一研究领域作出了贡献。

考虑一个在均方误差损失下训练的神经网络，$\mathcal{L} = \frac{1}{2}\sum_{i=0}^{N-1}(y_i - f_\theta(x_i))^2$, 针对 $x_i,\ y_i \in X_T,\ y_T$进行训练，其中这些符号表示训练集数据点及其标签。进一步假设该网络通过梯度流进行训练，且大致处于惰性/神经切线核机制下，也就是说，它的行为类似于网络输出的一阶泰勒展开（Chizat 等人，2019）：

$$f_\theta(x) \approx f_{lin,\theta}(x) = f_{\theta_0}(x) + (\theta - \theta_0)^\intercal \nabla_\theta f_{\theta_0}(x) \tag{2}$$

Lee 等人（2019年）表明，在此机制下，网络参数的时间演化为：

$$\theta(t) = \theta_0 - \nabla_\theta f_{\theta_0}(X_T)^\intercal K_0^{-1}\left(I - e^{-\eta K_0 t}\right)(f_{\theta_0}(X_T) - y_T)$$

其中，η 为学习率，K_0 为在 θ_0 处评估的有限宽度/经验神经切线核。也就是说，参数的最终变化为：

$$\Delta\theta = \theta_f - \theta_0 = \nabla_\theta f_{\theta_0}(X_T)^\intercal K_0^{-1}(y_T - f_{\theta_0}(X_T)) \tag{3}$$

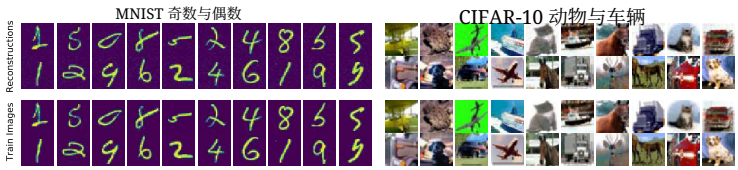


图 1：MNIST 奇数与偶数以及 CIFAR-10 动物与车辆分类任务的重建图像（上）与最接近的训练图像（下）。这些重建图像由采用标准动力学和较低学习率训练的 4096 宽度双隐藏层全连接网络生成。除少量噪声外，原始图像与其重建图像在视觉上难以区分。

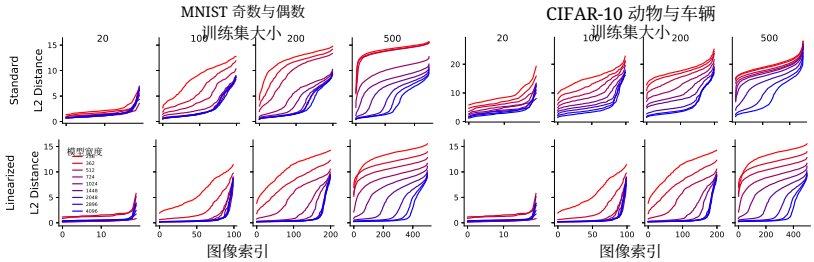


图 2：MNIST 奇数与偶数以及 CIFAR-10 动物与车辆分类任务的重建质量曲线。我们采用标准动力学（上）或线性化动力学（下），并改变训练集大小和模型宽度。较小的数据集更容易重构，而较宽的模型能够重构更多图像；线性化在这两种情况下都有帮助。

值得注意的是，这是以下优化问题的解：

$$\arg\min_{\Delta\theta} \frac{1}{2}\|\Delta\theta\|_2^2 \quad \text{s.t.} \quad \Delta\theta^\intercal \nabla_\theta f_{\theta_0}(X_T) = y_T - f_{\theta_0}(X_T) \tag{4}$$

相应的 KKT 条件为：

$$\Delta\theta = \alpha^\intercal \nabla_\theta f_{\theta_0}(X_T) \tag{5}$$

$$\Delta\theta^\intercal \nabla_\theta f_{\theta_0}(X_T) = y_T - f_{\theta_0}(X_T) \tag{6}$$

其中 α 是对偶参数集。在我们的表述中，式 (5) 确保我们处于驻点，而式 (6) 确保标签正确。与 Haim 等人（2022 年）类似，在给定网络的最终参数和初始化以获得 $\Delta\theta$ 的情况下，我们可以直接优化重建图像和对偶参数（分别为 X 和 α），使其满足这些 KKT 条件。实际上，出于下文将说明的原因，我们只需优化式 (5)，由此得到我们的重建损失：

$$\mathcal{L}_{\text{Reconstruction}} = \|\Delta\theta - \alpha^\intercal \nabla_\theta f_{\theta_0}(X_T)\|_2^2 \tag{7}$$

无限宽度下的重构。接下来，我们将说明，这种攻击表述能够恢复无限宽度模型的整个训练集。我们进一步假设训练数据位于单位超球面上。

定理 1。若 $\mathcal{L}_{reconstruction} = 0$（来自式 7），则在假设训练数据位于单位超球面上的情况下，我们可以在无限宽度极限下重构整个训练集。

证明。定义 $k_\theta(x, x') = \nabla_\theta f_\theta(x)^\intercal \nabla_\theta f_\theta(x')$，即有限宽度/经验神经正切核函数。我们知道，随着网络宽度 $w \to \infty$ 增大，$\Delta\theta = \sum_{\alpha_i,\ x_i \in \alpha^T,\ X_T} \alpha_i \nabla_{\theta_0} f_{\theta_0}(x_i)$，其中 $\alpha^T = K_{\theta_0}^{-1}{}_{,\ TT} y_T$，$X_T$ 为训练集，y_T 为训练标签，而 $K_{\theta_0,\ TT}$ 为在训练集上计算的有限宽度神经切线核。于是，我们的攻击变为：

$$\left\|\Delta\theta - \sum_{\alpha_j x_j \in \alpha^R, X_R} \alpha_j \nabla_{\theta_f} f_{\theta_f}(x_j)\right\|_2^2 = \left\|\sum_{\alpha_i, x_i \in \alpha^T, X_T} \alpha_i \nabla_{\theta_0} f(x_i) - \sum_{\alpha_j x_j \in \alpha^R, X_R} \alpha_j \nabla_{\theta_f} f_{\theta_f}(x_j)\right\|_2^2 \tag{8}$$

$$= \left\|\sum_{\alpha_i, x_i \in \alpha^T, X_T} \alpha_i k_{\theta_0}(x_i, \cdot) - \sum_{\alpha_j x_j \in \alpha^R, X_R} \alpha_j k_{\theta_f}(x_j, \cdot)\right\|_2^2 \tag{9}$$

表 1：将 500 张图像蒸馏为 20 张图像时，KIP、Recon-KIP（RKIP）以及基于训练后的网络的 RKIP（RKIP-finite）的性能。KIP 和 RKIP 在无限宽度下表现最佳，而 KIP 在有限模型下则失效。(n=7)

蒸馏算法	MNIST 奇偶分类			CIFAR-10 动物/交通工具		
	标准	线性化	无限宽度	标准	线性化	无限宽度
完整数据集（500 张图像）	92.85 ± 0.42	92.91 ± 0.33	93.18 ± 0.37	75.06 ± 0.21	74.60 ± 0.21	75.42 ± 0.28
KIP	57.42 ± 8.41	55.62 ± 7.48	$\mathbf{91.53 \pm 0.57}$	35.26 ± 5.67	32.37 ± 3.60	70.98 ± 0.43
RKIP	$\mathbf{89.61 \pm 1.18}$	$\mathbf{89.99 \pm 1.11}$	91.44 ± 0.48	$\mathbf{72.23 \pm 3.61}$	$\mathbf{72.76 \pm 3.74}$	$\mathbf{74.66 \pm 0.93}$
RKIP-finite	88.45 ± 0.89	86.15 ± 3.39	87.31 ± 3.24	71.96 ± 1.14	63.99 ± 4.02	62.05 ± 4.17
随机图像	73.52 ± 3.60	73.54 ± 3.61	74.12 ± 3.73	70.36 ± 2.53	70.18 ± 2.54	70.77 ± 2.04

其中 T 和 R 分别指训练集和重建集。由于 $w \to \infty$，我们知道 k_{θ_0}，$k_{\theta_f} \to k_{NTK}$。此外，定义

$$P_T = \sum_{\alpha_i, x_i \in \alpha^T, X_T} \alpha_i \delta(x_i), \qquad P_R = \sum_{\alpha_j, x_j \in \alpha^R, X_R} \alpha_j \delta(x_j)$$

分别表示与我们的训练后网络和重构相关联的测度，并且 $\mu_* = \int_\Omega k_{NTK}(x, \cdot) dP_*(x)$，其中 $\Omega = S^d$，d 为数据维度（假设数据位于单位超球面上）。μ_T 和 μ_R 现在分别是我们的训练后网络和重构的核嵌入。我们的重建损失变为：$\|\mu_T - \mu_R\|^2_{\mathcal{H}_{NTK}}$。这就是最大均值差异（MMD）（Gretton 等，2012）。注意，P_T, P_R 是有符号博雷尔测度（因为 α 是有限的，且我们的重构集/训练集位于单位球面上）。神经正切核在单位球面上是普适的（Jacot 等，2018），这意味着映射 μ：{有符号博雷尔测度族} $\to \mathcal{H}$ 是单射的（Sriperumbudur 等，2011），也就是说，只要 $\alpha_i \neq 0$ 成立（这几乎必然发生，见附录 C），我们就能够**恢复整个训练集**。 □

需要注意的是，实际应用中我们并不强制要求数据位于单位球面上，但仍能获得很高的重建质量，这一点将在第 4 节展示。从网络切空间到图像空间的映射也有助于理解梯度泄露攻击（Zhu 等，2019）为何能够成功：这类攻击利用梯度来寻找训练批次样本。

4 有限网络的数据集重建

虽然定理 1 中概述的攻击在无限宽度极限下具有重要的理论启示，但其实用性有限，因为计算核内积需要访问训练图像本身。对于有限宽度神经网络，这种攻击如何发挥作用？在什么条件下

这种攻击在什么情况下会成功?

为回答这些问题，我们遵循 Haim 等人（2022）的实验方案，尝试从 MNIST 和 CIFAR-10 数据集中重构图像；对于 MNIST 和 CIFAR-10，任务分别是奇偶数字分类以及动物/车辆分类。我们将训练集大小从每类 10 张图像调整到每类 250 张图像（训练集总大小为 500）。我们考虑使用标准初始化的带偏置双隐藏层神经网络（而不是神经切线核参数化或文中提出的初始化方案）

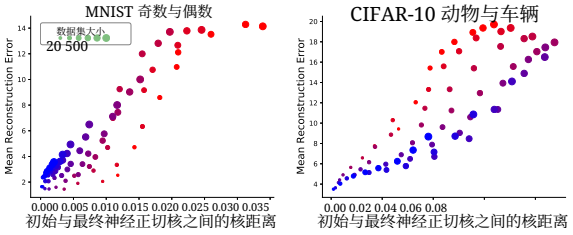


图3：平均重建误差与从初始化到最终核的核距离之间的关系。平均重建误差以重建曲线的平均值衡量，与有限宽度神经切线核在训练过程中的变化程度高度相关。数据集大小由点的大小表示，模型宽度则由颜色表示（见图2）。

Haim 等人（2022））。我们将神经网络的宽度设在 256 到 4096 之间，以观察偏离无限宽度状态会如何影响重建质量。此外，众所周知，对于有限宽度网络，有限宽度神经切线核会在训练过程中发生变化，从而偏离无限宽度状态。我们可以通过采用线性化训练来强制固定核；在这种训练方式中，我们对网络参数在其初始化值附近进行一阶泰勒展开并进行训练（见

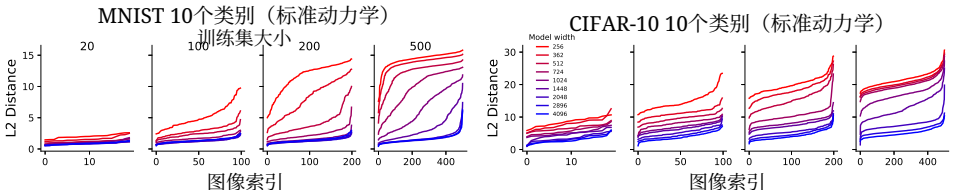


图 5：在 MNIST/CIFAR-10 多类别数据集上训练的网络的重建曲线。

公式（2））。我们分别在标准（未修改）动力学和线性化动力学下考察这两种网络。在附录 J.1 和附录 J.2 中，我们分别考察卷积架构和高分辨率数据集；但在正文中，我们将注意力限制在低分辨率图像上的多层感知机。

我们使用低学习率的全批量梯度下降训练这些网络 10^6 iteration，并在重构过程中生成 $M = 2N$ 重建图像，其中 N 为训练集大小。实验参数的完整说明见附录 H，算法运行时间的详细信息见附录 B。为衡量重建质量，我们采用以下指标。首先，测量每张训练图像与每个重建图像在像素空间中的平方 L_2 距离。选取距离最小的一对训练图像和重建图像，并将其从候选集合中移除，作为一对图像/重建图像。重复此过程，直至获得完整的 N 训练图像和重建图像集合（见图 1）。然后，将 L_2 距离按升序排列，并绘制该函数。我们将此称为与特定重建集相关联的**重建曲线**。我们在图 2 中绘制了不同数据集大小和模型宽度下的重建曲线。从图 2 中，我们得到以下三点观察：

训练集越小，越容易进行重构。 我们发现，对于所有模型宽度而言，较小数据集的重建曲线都具有较低的值。**更宽的模型能够解析更大的数据集。** 我们观察到，对于给定的模型宽度，存在一个阈值图像索引，超过该索引后，重建图像的质量会急剧下降。例如，对于 MNIST 奇偶分类，在包含200张图像且宽度为1024时，该阈值为80张图像。随着模型宽度增加，该阈值几乎单调递增。

线性化能够提升重建质量。 我们发现，与相同宽度的对应模型相比，线性化网络能够解析更多图像，并生成质量更高的图像。线性化的成功表明，偏离冻结 NTK 状态会影响重建质量。我们可以通过测量模型初始化的核距离 来衡量其偏离冻结核状态的程度

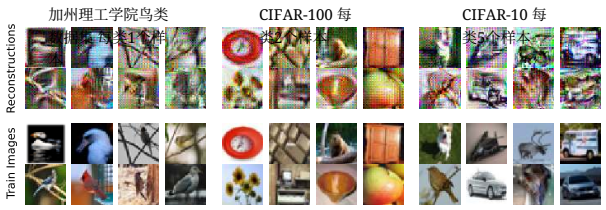


图4：在 ImageNet 上预训练的 ResNet-18 上，对加州理工学院鸟类数据集（每类1个样本）、CIFAR-100（每类2个样本）和 CIFAR-10（每类5个样本）进行少样本微调时，对训练数据的重建图像。

神经正切核（NTK）及其最终 NTK 定义如下：$d(K_0,\ K_f) = 1 - \frac{\text{迹}(K_0^\intercal K_f)}{\|K_0\|_F \|K_f\|_F}$。

直观来说，这个距离反映了初始化核与最终核的对齐程度。数值越大，表示核发生了显著变化，也意味着网络偏离 NTK 机制越远。我们在图 3 中绘制了这些核距离与重建曲线均值之间的关系。可以立即看出，重建质量与核距离密切相关，而且数据集越小、模型越宽，核距离越低。在附录 G 中，我们讨论了攻击如何**兼容提前停止和交叉熵损失**；这一点不同于 Haim 等人（2022 年）的方法，后者要求训练直至收敛。附录 G 中还对提前停止的影响进行了更详细的讨论。

多类别分类。 在前面的章节中，我们展示了该攻击在二分类中的有效性。这里我们验证该攻击在多类别情况下是否有效。我们按照第4节中的相同流程进行实验，但使用全部10个类别。详情见附录 B，附录 J.2 还提供了 Tiny-ImageNet 上进行200类分类的其他结果。结果通过图5中的重建曲线展示。我们观察到，类别越多，该攻击的**重建质量越高**。在附录J.3中，我们观察到，多类分类会带来更低的核距离，这表明该攻击的行为更接近核区间，从而解释了其更好的重建质量。

未来的工作可以对此作进一步研究。

5 微调中的数据集重建

攻击的一个关键前提是模型初始化。若从头开始训练，攻击者将无法获得这一信息，使得攻击毫无用处。然而，实践中通常不会从头开始训练，而是微调公开可用的大型预训练模型。此外，用户通常无法获得大量数据，因此任务实际上属于少样本场景。有证据表明，微调后期训练神经网络的过程可由冻结有限神经切线核理论很好地近似（Zancato 等人，2020年；Shon 等人，2022年；Zhou 等人，2021；Malladi 等人，2023年），因此，少样本微调场景很容易成为这种攻击的目标。为评估该场景下的攻击效果，我们使用公开可用、在 ImageNet 上预训练的 ResNet-18 模型，分别在加州理工学院鸟类数据集（200个类别）（Welinder 等人，2010年）、CIFAR-100 和 CIFAR-10 上进行少样本图像分类微调，每类样本数依次为 1、2 和 5。对于 θ_0，我们使用微调初始阶段的模型参数。最佳重建图像见图 4。可以看到，我们的攻击能够恢复部分训练图像，但质量有限。未来的工作可以探索如何改进这些攻击。

6 哪些数据点容易被重构?

先前的研究发现，并非所有数据点受到隐私攻击的影响都相同（Carlini 等人，2022年；Feldman 和 Zhang，2020年a；Carlini 等人，2021；Bagdasaryan 和 Shmatikov，2019年）。尤其是，离群值 图像往往比其他图像更容易泄露。在本节中，我们将展示这一现象也会出现在我们的攻击中，并为此提供理论依据。

6.1 难以拟合意味着易于重构

考虑到我们的重建损失 $\left\|\Delta\theta - \sum_j \alpha_j \phi(x_j)\right\|_2^2$ 以及 $\phi(x_j) = \nabla_{\theta_f} f_{\theta_f}(x_j)$，我们旨在学习一个基底来“解释” $\Delta\theta$，由此可见，这可以转化为一个稀疏编码问题。假设所有 $\phi(x_j)$的幅度大致相同，我们预计， α 参数较大的参数对 $\Delta\theta$的贡献更多，因此也更容易被重构。这与高影响力的数据点更有可能被记忆的现象密切相关（Feldman 和 Zhang，2020年a）。我们通过绘制图 7 中重建误差与相应的 $|\alpha|$ 值（在无限宽度下计算）之间的散点图，来验证这一经验判断在实证上是否成立。我们发现， α 值较小的图像.....

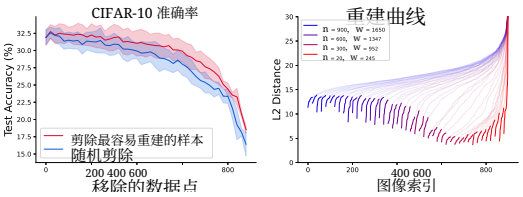


图 6：采用随机剪枝或基于易于重构的图像进行剪枝时，迭代剪枝的 CIFAR-10 的测试准确率（左），以及剪枝后的 CIFAR-10 的重建曲线（右）。我们看到，可以移除易于重构的数据点而不损害准确率。我们观察到一种隐私“洋葱”效应：移除易于重构的图像后，会显露出其他同样容易重构的图像。加粗线条表示剪枝后移除的 20 张图像。

由于它们对 $\Delta\theta$ 的贡献很小，可以被视为噪声，因此它们受到“保护”，不会被重构。根据附录 C，我们知道 $\alpha = 0$ 对应的图像/标签完全不会改变预测，这说明预测准确的数据点不会被重构。除了 $\alpha = K^{-1}y$的定义之外，我们还可以从另一个角度说明， α 与模型拟合该数据点的速度 密切相关。我们可以写成（推导见附录 C.1）： $\alpha_j = \int_0^\infty (y_j - f_{\theta_t}(x_j))dt$ 这意味着拟合较慢的数据点会有较大的 α 值，进一步支持了离群点更容易被重构这一观点。

6.2 重建隐私洋葱

Carlini 等人（2022）指出，移除“脆弱”的训练点后，会显露出另一组容易受到推理攻击的训练点。在这里，我们验证了我们的重建攻击也呈现出类似现象。具体而言，我们考虑在最初使用 $n = 900$ 对 n 个数据点进行训练后，重构 CIFAR-10 训练图像。我们使用 $w \propto \sqrt{n}$ 训练网络并实施攻击，然后根据重建曲线，迭代移除最容易重构的 20 个数据点。我们使用 n 缩放网络容量，使攻击无法重构整个训练

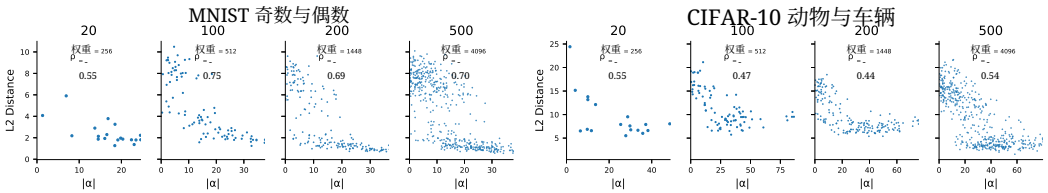


图 7：以 L2距离衡量的重建质量散点图，以及图像对应的 $|\alpha|$值。我们通过改变宽度并使用 n，以观察不同范围的重建质量。$|\alpha|$与重建误差呈负相关。

集。如图 6 所示，尽管随着移除的数据点增多，我们的网络容量和攻击能力都在下降，但我们仍能以不断提升的攻击质量重构数据点，重现“隐私洋葱”效应。未来的工作可以研究 α 参数之间的相互作用如何影响哪些项目在移除数据点后容易被重构。同样，我们在图 6 中评估了这些剪枝子集上的测试准确率。我们发现，这些容易重构的数据点往往是离群点，因此移除它们对测试准确率的影响较小。不过，正如 Sorscher 等人（2022年）所讨论的，剪枝时选择移除容易还是困难的数据点，取决于数据集大小和任务复杂度等其他因素。

7 统一重构与蒸馏

在前面的章节中，我们考虑了重构整个训练集的任务。为此，我们设置了重构图像数量 $M > N$。如果设置为 $M < N$，会发生什么？我们会重构出少量训练图像的子集，还是会形成平均了训练集的图像？

我们在 CIFAR-10 动物/车辆任务上进行这项实验，使用 500 张训练图像和一个宽度为 4096、具有线性化动力学的模型，目标是只重构 20 张图像。我们在附录 E 中恢复了所示图像。除少数例外，这些图像并非训练集中的样本，而是看起来像各类别的平均图像。

那么，如果我们在这些重建图像上重新训练一个网络，会怎样？注意到，完整重构的最优对偶参数集合由 $\alpha^T = K_{TT}^{-1} y_T$ 给出，因此这些图像的训练标签自然可以选为 $y_R = K_{RR}\alpha^R$；在这里，我们计算 K_{RR} 的经验神经正切核，并使用学习到的 α^R 参数。使用这些重建得到的 20 张图像，从头开始重新训练另一个网络，取得了较高的准确率，如表 1 中的 RKIP-finite 所示。这表明，通过进行这种重构，我们实现了数据集蒸馏，也就是说，我们构造了一个更小的图像集，能够准确近似完整数据集。这并非巧合；事实上，这两种算法是相同的。更正式地说：

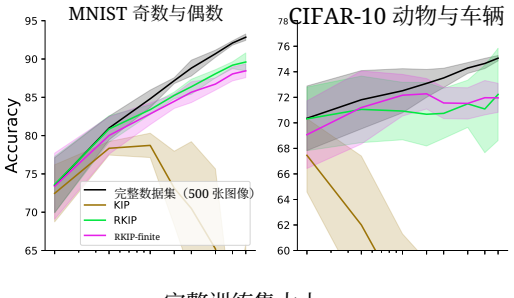


图 8：使用标准动力学，在宽度为 4096 的网络上训练，将 N 张图像蒸馏为 20 张图像时，KIP、RKIP 和 RKIP-有限网络的性能。KIP 无法迁移到有限网络，而 RKIP 的各种变体则能成功迁移。

定理 2. 公式 *4* 的重构方案，结合公式 *5* 和公式 *6* 的 *KKT* 点，在 $M \leq N$ 中，其中 M 是重构图像数量，N 是数据集大小，可以表示为采用不同范数的核诱导点蒸馏损失，并加上一个方差控制误差，如下所示：

$$\mathcal{L}_{Reconstruction} = \overbrace{\|y_T - K_{TR}K_{RR}^{-1}y_R\|^2_{K_{TT}^{-1}}}^{RKIP\ loss} + \lambda_{var\ of\ R|T}$$

完整证明见附录 D。$\lambda_{\text{var of } R|T}$与基于神经正切核（NTK）的、在给定训练数据条件下的重建数据点方差成正比（见附录 D）。直观而言，它确保训练图像能够为重建图像提供“信息”。与一种著名的数据集蒸馏算法 KIP 的损失相比

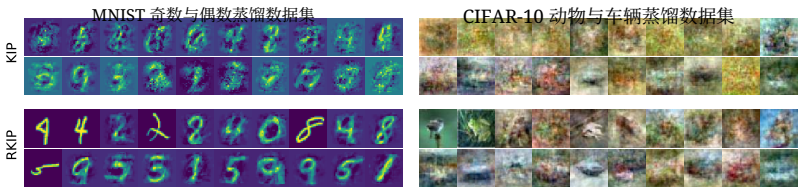


图 9：使用 KIP 和 RKIP 对 MNIST 奇偶分类及 CIFAR-10 动物/车辆分类所生成的蒸馏数据集可视化。我们将包含 500 张原始图像的数据集蒸馏为 20 张图像（如图所示）。KIP 不会复制原始训练图像，而 RKIP 偶尔会重现训练图像。

（其中 S 指蒸馏数据集）：

$$\mathcal{L}_{\text{KIP}} = \|y_T - K_{TS}K_{SS}^{-1}y_S\|_2^2$$

这种联系显而易见：重建损失等于采用不同范数的 KIP 数据集蒸馏损失；与对每个数据点赋予相同权重不同，我们根据 NTK 测得的逆相似度为训练图像加权，再加上 $\lambda_{\text{var 的 } r|T}$。由此得到 KIP 的一个变体，我们称之为 Recon-KIP（RKIP），它使用定理 2 中的重建损失。请注意，对于大型数据集，RKIP 在实践中并不可行，因为它需要计算 K_{TT}^{-1}，这通常是 $N \times N$。本研究处理的是小型数据集，因此仍然可以进行计算。

我们在表 1 中总结了 KIP 和 RKIP 在 MNIST 奇偶分类以及 CIFAR-10 动物/车辆任务上的表现，将 500 张图像蒸馏为 20 张。我们评估了采用标准动力学或线性化动力学的宽度为 4096 的网络，以及使用神经正切核的无限宽度网络。此外，我们还考虑使用在完整数据集上训练的有限网络重建数据集样本点，从而得到图像/标签，并将这种方法称为 RKIP-finite。请注意，在这种情况下，标签不一定是 $\{+1, -1\}$，因为 $K_{0,\ RR}\alpha^R$ 不一定是独热标签。图 9 展示了类似的结果：将数量较少的图像（20–500 张训练图像）蒸馏为 20 张蒸馏图像。

从表 1 可以看出，KIP 和 RKIP 在无限宽度下的准确率都很高，但 KIP 迁移到有限网络时性能显著下降。例如，KIP 在 MNIST 奇偶分类任务上的无限宽度测试准确率达到 91.53%，但其有限宽度下的表现为 55。62%，并未显著优于随机猜测。有趣的是，如图 8 所示，蒸馏的图像越多，这种性能差距就越大。对于较小的蒸馏数据集，性能几乎没有下降，但对于较大的数据集，差异则十分显著。与之相反，RKIP 在有限宽度设置下的性能下降出奇地小。我们推测，KIP 与 RKIP 在有限宽度下的迁移性能差异可能源于 $\lambda_{\text{var of } r|T}$的贡献，相关讨论见附录 D。我们将进一步探索这一问题留待未来工作。此外，尽管 RKIP-finite 的蒸馏仅使用了单个有限宽度神经网络的信息，其表现仍几乎与 RKIP 一样好。

8 讨论、局限性与结论

在这项工作中，我们证明了 Haim 等人（2022）提出的一种更强的攻击变体能够重构整个训练集。该攻击要求使用均方误差损失训练宽神经网络，其有效性可由神经正切核测度嵌入的单射性严格证明。我们还表明，该攻击在有限宽度网络上实践中同样有效，但偏离无限宽度状态会削弱攻击效果。我们探讨了在我们的攻击下，离群数据点更容易被重构这一现象，并发现这些易于重构的图像可能不利于学习。最后，我们在这种重建攻击与数据集蒸馏之间建立了新的联系。尽管这些发现有助于我们理解数据集重建攻击及其理论基础，但仍有许多方向值得探索。

在这项工作中，我们主要研究了两层全连接网络；众所周知，这类神经网络的表现与其无限宽度对应模型相似。与此同时，深层卷积网络已知会出现显著偏离，因此尚不清楚本文中的攻击在这些情形下是否同样适用，以及需要做出哪些调整。其次，虽然我们观察到，模型宽度增加会提升网络的“分辨能力”（即它能够重构的图像数量），但未来的工作可以探讨这一数量如何源于有限宽度神经切线核与无限宽度神经切线核之间的偏离。最后，我们仍需厘清数据集

重构这一隐私概念如何与差分隐私等更成熟的概念相联系；这将是未来工作的研究主题。

我们相信，这项工作为理解数据集重建攻击的优势与劣势迈出了重要一步，并与神经切线核和数据集蒸馏等现有研究建立了新的联系。

8 可复现性声明

本工作使用开源数据集和模型。实验细节（如超参数）见附录 H。我们还在补充材料中提供了运行实验的代码。

参考文献

Martin Abadi、Andy Chu、Ian Goodfellow、H. Brendan McMahan、Ilya Mironov、Kunal Talwar 和 Li Zhang。差分隐私深度学习。载于*2016*年 *ACM SIGSAC*计算机与通信安全会议论文集，CCS '16，第 308–318 页，纽约，纽约州，美国，2016年。计算机协会。ISBN 9781450341394。数字对象标识符：10.1145/2976749.2978318。URLhttps://doi.org/10.1145/2976749.2978318。

Kyle Aitken 和 Guy Gur-Ari。《多项式激活函数下宽网络的渐近性质》。*ArXiv*，abs/2006.06687，2020年。

Sanjeev Arora、Simon S Du、Wei Hu、Zhiyuan Li、Russ R Salakhutdinov 和 Ruosong Wang。《关于无限宽神经网络的精确计算》。收录于神经信息处理系统进展，第 8141–8150 页。Curran Associates, Inc., 2019。

Devansh Arpit、Stanisław Jastrzebski、Nicolas Ballas、David Krueger、Emmanuel Bengio、Maxin- der S Kanwal、Tegan Maharaj、Asja Fischer、Aaron Courville、Yoshua Bengio 等。《深入考察深度网络中的记忆》。收录于国际机器学习会议，第 233–242 页。PMLR，2017。

Igor Babuschkin、Kate Baumli、Alison Bell、Surya Bhupatiraju、Jake Bruce、Peter Buchlovsky、David Budden、Trevor Cai、Aidan Clark、Ivo Danihelka、Claudio Fantacci、Jonathan Godwin、Chris Jones、Ross Hemsley、Tom Hennigan、Matteo Hessel、Shaobo Hou、Steven Kapturowski、Thomas Keck、Iurii Kemaev、Michael King、Markus Kunesch、Lena Martens、Hamza Merzic、Vladimir Mikulik、Tamara Norman、John Quan、George Papamakarios、Roman Ring、Francisco Ruiz、Alvaro Sanchez、Rosalia Schneider、Eren Sezener、Stephen Spencer、Srivatsan Srinivasan、Luyu Wang、Wojciech Stokowiec 和 Fabio Viola。《DeepMind JAX生态系统》，2020年。URL http://github.com/deepmind。

Eugene Bagdasaryan 和 Vitaly Shmatikov。差分隐私对模型准确率的影响并不均衡。*CoRR*，abs/1905.12101，2019年。网址：http://arxiv.org/abs/1905.12101。

Lucas Bourtoule、Varun Chandrasekaran、Christopher A. Choquette-Choo、Hengrui Jia、Adelin Travers、Baiwu Zhang、David Lie 和 Nicolas Papernot。机器反学习。*CoRR*，abs/1912.03817，2019年。网址：http://arxiv.org/abs/1912.03817。

James Bradbury、Roy Frostig、Peter Hawkins、Matthew James Johnson、Chris Leary、Dougal Maclaurin、George Necula、Adam Paszke、Jake VanderPlas、SkyeWanderman-Milne 和 Qiao Zhang。JAX：可组合的 Python+NumPy 程序变换，2018年。网址：http: //github.com/google/jax。

Gavin Brown、Mark Bun、Vitaly Feldman、Adam Smith 和 Kunal Talwar。何时需要记忆无关训练数据才能实现高准确率学习？载于第*53*届 *ACM SIGACT* 计算理论年会论文集，第 123–132 页，2021。

´Nicholas Carlini、Chang Liu、Ulfar Erlingsson、Jernej Kos 和 Dawn Song。《秘密共享者：评估和测试神经网络中的非预期记忆》。载于第*28*届 *USENIX* 安全研讨会论文集，SEC'19，第 267–284 页，美国，2019年。USENIX 协会。ISBN 9781939133069。

Nicholas Carlini、Florian Tram`er、Eric Wallace、Matthew Jagielski、Ariel Herbert-Voss、Katherine´ Lee、Adam Roberts、Tom B. Brown、Dawn Song、Ulfar Erlingsson、Alina Oprea 和 Colin Raf-fel。《从大语言模型中提取训练数据》。*CoRR*，abs/2012.07805，2020年。网址：https://arxiv.org/abs/2012.07805。

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, AndreasTerzis 和 Florian Tram`er。《从第一原理出发的成员推理攻击》。*CoRR*，abs/2112.03570，2021。网址：https: //arxiv.org/abs/2112.03570。

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis 和 Florian Tramer。《隐私洋葱效应：记忆是相对的》。载于 S. Koyejo、S. Mohamed、A. Agarwal、D. Belgrave、K. Cho 和 A. Oh（编），神经信息处理系统进展，第 35 卷，第 13263–13276 页。Curran Associates, Inc.，2022年。网址：https://proceedings.neurips.cc/paper_files/paper/2022/ file/564b5f8289ba846ebc498417e834c253-Paper-Conference.pdf。

美国医疗保险和医疗补助服务中心。《1996年健康保险流通与责任法案》（HIPAA）。在线网址：http://www.cms.hhs.gov/hipaa/，1996年。

L´ena ¨ıc Chizat、Edouard Oyallon 和 Francis R. Bach。关于可微编程中的惰性训练 m-m中*NeurIPS*，2019年。

Rodrigo de Azevedo（https://math.stackexchange.com/users/339790/rodrigo-de azevedo）。梯度下降在最小二乘问题中会收敛到最小范数解吗？数学 Stack Exchange。网址：https://math.stackexchange.com/q/3499305。网址：https://math.stackexchange.com/q/3499305（版本：2022-02-18）。

Cynthia Dwork、Frank McSherry、Kobbi Nissim 和 Adam Smith。隐私数据分析中将噪声校准至敏感度。载于第三届密码学理论会议论文集，TCC'06，第265–284页，柏林、海德堡，2006年。Springer-Verlag。ISBN 3540327312。数字对象标识符：10. 1007/11681878 14。网址：https://doi.org/10.1007/11681878_14。

欧盟委员会。《欧洲议会和理事会于2016年4月27日颁布的关于在处理个人数据时保护自然人及此类数据自由流动，并废除指令95/46/EC的条例（EU）2016/679》（《通用数据保护条例》）（与欧洲经济区相关的文本），2016年。网址：https://eur-lex.europa.eu/eli/reg/2016/679/oj。

Vitaly Feldman。学习是否需要记忆？关于长尾的短篇故事。载于第*52*届 *ACM SIGACT* 年度计算理论研讨会论文集，STOC 2020，第954–959页，纽约，纽约州，美国，2020年。计算机协会。ISBN 9781450369794。数字对象标识符：10.1145/3357713.3384290。网址：https://doi.org/10.1145/3357713.3384290。

Vitaly Feldman 和 Chiyuan Zhang。《神经网络记忆什么以及为何记忆：通过影响力估计发现长尾》。*CoRR*，abs/2008.03703，2020a。网址：https://arxiv. org/abs/2008.03703。

Vitaly Feldman 和 Chiyuan Zhang。《神经网络记忆什么以及为何记忆：通过影响力估计发现长尾》。载于第*34*届神经信息处理系统国际会议论文集，NIPS'20，纽约州雷德胡克，美国，2020b。Curran Associates Inc. ISBN 9781713829546。

Stanislav Fort、Gintare Karolina Dziugaite、Mansheej Paul、Sepideh Kharaghani、Daniel M. Roy 和 Surya Ganguli。《深度学习与核学习：对损失景观几何和神经切线核随时间演化的实证研究》。载于*NeurIPS*，2020年。网址：https://proceedings.neurips.cc/paper/2020/hash/405075699f065e43581f27d67bb68478-Abstract.html。

Matt Fredrikson、Somesh Jha 和 Thomas Ristenpart。利用置信度信息的模型反演攻击及基本对策。载于第 *22* 届 *ACM SIGSAC* 计算机与通信安全会议论文集，第 1322–1333 页，2015年。

Arthur Gretton、Karsten M. Borgwardt、Malte J. Rasch、Bernhard Sch¨olkopf 和 Alexander Smola。核双样本检验。机器学习研究杂志，13(null):723–773，2012 年 3 月。国际标准连续出版物号 1532-4435。

Niv Haim、Gal Vardi、Gilad Yehudai、michal Irani 和 Ohad Shamir。从训练好的神经网络中重建训练数据。载于 Alice H. Oh、Alekh Agarwal、Danielle Belgrave 和 Kyunghyun Cho（编），神经信息处理系统进展，2022年。网址：https://openreview.net/forum?id=Sxk8Bse3RKO。

Boris Hanin 和 Mihai Nica。神经切线核的有限深度和宽度修正。载于国际学习表征会议，2020年。网址：https://openreview. net/forum?id=SJgndT4KwB。

Zecheng He、Tianwei Zhang 和 Ruby B. Lee。针对协同推理的模型反演攻击。载于第*35*届年度计算机安全应用会议论文集，ACSAC '19，第148–162页，纽约，纽约州，美国，2019年。计算机协会。ISBN 9781450376280。数字对象标识符：10.1145/3359789.3359824。网址：https://doi.org/10.1145/ 3359789.3359824。

Jonathan Heek、Anselm Levskaya、Avital Oliver、Marvin Ritter、Bertrand Rondepierre、Andreas Steiner 和 Marc van Zee。Flax：面向 JAX 的神经网络库及其生态系统，2020年。网址：http://github.com/google/flax。

Arthur Jacot、Franck Gabriel 和 Clement Hongler。神经切线核：神经网络中的收敛与泛化。载于 S. Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 和 R. Garnett（编），神经信息处理系统进展，第 31 卷。Curran Associates, Inc.，2018年。网址：https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf。

Ziwei Ji 和 Matus Telgarsky。深度学习中的方向收敛与对齐。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin（编），神经信息处理系统进展，第 33 卷，第 17176–17186 页。Curran Associates, Inc.，2020年。网址：https://proceedings.neurips.cc/paper/2020/file/c76e4b2fa54f8506719a5c0dc14c2eb9-Paper.pdf。

Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于 Yoshua Bengio 和 Yann LeCun（编），第 *3* 届国际学习表征会议，*ICLR 2015*，美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集，2015年。网址http://arxiv.org/abs/1412.6980。

Pang Wei Koh 和 Percy Liang。通过影响函数理解黑盒预测。载于第 *34* 届国际机器学习会议论文集——第 *70* 卷，ICML'17，第 1885–1894 页。JMLR.org，2017。

Ya Le 和 Xuan S. Yang。Tiny ImageNet 视觉识别挑战赛。2015年。网址 https: // api.semanticscholar.org/CorpusID:16664790。

Jaehoon Lee、Lechao Xiao、Samuel Schoenholz、Yasaman Bahri、Roman Novak、Jascha Sohl-Dickstein 和 Jeffrey Pennington。宽神经网络在梯度下降下，无论深度如何，都会演化为线性模型。神经信息处理系统进展，32，2019年。

Noel Loo、Ramin Hasani、Alexander Amini 和 Daniela Rus。使用随机特征近似进行高效的数据集蒸馏。神经信息处理系统进展，2022年a。

Noel Loo、Ramin Hasani、Alexander Amini 和 Daniela Rus。良性训练与对抗训练下神经切线核的演化。载于神经信息处理系统进展，2022年b。

Kaifeng Lyu 和 Jian Li。梯度下降使齐次神经网络的间隔最大化。载于国际学习表征会议，2020年。网址：https:// openreview.net/forum?id=SJeLIgBKPS。

Sadhika Malladi、Alexander Wettig、Dingli Yu、Danqi Chen 和 Sanjeev Arora。《基于核的语言模型微调视角》，2023年。网址：https://openreview.net/forum?id=erHaiO9gz3m。

Timothy Nguyen、Zhourong Chen 和 Jaehoon Lee。基于核岭回归的数据集元学习。载于国际学习表征会议，2021a。网址：https: //openreview.net/forum?id=l-Prr QrK0QR。Timothy Nguyen、Roman Novak、Lechao Xiao 和 Jaehoon Lee。使用无限宽卷积网络进行数据集蒸馏。载于第三十五届神经信息处理系统进展会议，2021b。网址：https://openreview.net/forum?id=hXWPpJedrVP。Roman Novak、Lechao Xiao、Jiri Hron、Jaehoon Lee、Alexander A. Alemi、Jascha Sohl-Dickstein 和 Samuel S. Schoenholz。神经切线：在 Python 中快速、轻松地使用无限神经网络。载于国际学习表征会议，2020年。网址：https://openreview. net/forum?id=SklD9yrFPS。Roman Novak、Jascha Sohl-Dickstein 和 Samuel S. Schoenholz。快速有限宽度神经切线核。载于 Kamalika Chaudhuri、Stefanie Jegelka、Le Song、Csaba Szepesv´ari、Gang Niu 和 Sivan Sabato（编），国际机器学习会议（*ICML 2022*），*2022*年*7*月*17*日至*23*日，美国马里兰州巴尔的摩，第162卷机器学习研究论文集，第17018–17044页。PMLR，2022年。网址：https://proceedings.mlr.press/ v162/novak22a.html。Maria Rigaki 和 Sebastian Garcia。机器学习隐私攻击综述。*CoRR*，abs/2007.07646，2020 年。网址：https://arxiv.org/abs/2007.07646。Reza Shokri、Marco Stronati 和 Vitaly Shmatikov。针对机器学习模型的成员推断攻击。*CoRR*，abs/1610.05820，2016年。网址：http://arxiv.org/abs/1610. 05820。Hyounguk Shon、Janghyeon Lee、Seung Hwan Kim 和 Junmo Kim。DLCFT：用于通用增量学习的深度线性持续微调。*CoRR*，abs/2208.08112，2022年。数字对象标识符：10.48550/arXiv. 2208.08112。网址：https://doi.org/10.48550/arXiv.2208.08112。Ben Sorscher、Robert Geirhos、Shashank Shekhar、Surya Ganguli 和 Ari S. Morcos。超越神经缩放定律：通过数据剪枝打破幂律缩放。*ArXiv*，abs/2206.14486，2022年。网址：https://api.semanticscholar.org/CorpusID:250113273。Bharath K. Sriperumbudur、Kenji Fukumizu 和 Gert R.G. Lanckriet。通用性、特征核与测度的 RKHS 嵌入。机器学习研究期刊，12(70):2389–2410，2011年。网址：http://jmlr.org/papers/v12/sriperumbudur11a.html。Nikolaos Tsilivis 和 Julia Kempe。神经切线核能告诉我们关于对抗鲁棒性的哪些信息？载于 Alice H. Oh、Alekh Agarwal、Danielle Belgrave 和 Kyunghyun Cho（编），神经信息处理系统进展，2022年。网址：https://openreview. net/forum?id=KBUgVv8z7OA。Tongzhou Wang、Jun-Yan Zhu、Antonio Torralba 和 Alexei A Efros。数据集蒸馏。*arXiv*预印本 *arXiv:1811.10959*，2018 年。P. Welinder 、 S. Branson 、 T. Mita 、 C. Wah 、 F. Schroff、S. Belongie 和 P. Perona。Caltech-UCSD鸟类200。技术报告 CNS-TR-2010-001，加州理工学院，2010年。Matthias Wright。Flax模型。https://github.com/matthias-wright/flaxmodels，2022年。

Ziqi Yang、Jiyi Zhang、Ee-Chien Chang 和 Zhenkai Liang。通过背景知识对齐在对抗环境下进行神经网络反演。载于 *2019*年 *ACM SIGSAC*计算机与通信安全会议论文集，CCS '19，第225–240页，纽约，纽约州，美国，2019年。计算机协会。ISBN 9781450367479。数字对象标识符：10.1145/3319535.3354261。URLhttps://doi.org/10.1145/3319535.3354261。

Luca Zancato、Alessandro Achille、Avinash Ravichandran、Rahul Bhotika 和 Stefano Soatto。无需训练即可预测训练时间。载于 H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin（编），神经信息处理系统进展，第 33 卷，第 6136–6146 页。Curran Associates, Inc.，2020年。网址：https://proceedings.neurips.cc/paper_files/paper/2020/file/440e7c3eb9bbcd4c33c3535354a51605-Paper.pdf。

Amir Zandieh、Insu Han、Haim Avron、Neta Shoham、Chaewon Kim 和 Jinwoo Shin。通过草图化和随机特征扩展神经切线核。载于 A. Beygelzimer、Y. Dauphin、P. Liang 和 J. Wortman Vaughan（编），神经信息处理系统进展，2021。网址：https://openreview.net/forum?id=vIRFiA658rh。

Chiyuan Zhang、Samy Bengio、Moritz Hardt、Benjamin Recht 和 Oriol Vinyals。理解深度学习需要重新思考泛化。在国际学习表征会议，2017年。URL https://openreview.net/forum?id=Sy8gdB9xx。

Chiyuan Zhang、Samy Bengio、Moritz Hardt、Benjamin Recht 和 Oriol Vinyals。理解深度学习（仍然）需要重新思考泛化。《style id='1'>《ACM通讯》，64(3):107–115，2021年2月。国际标准连续出版物号0001-0782。数字对象标识符：10.1145/3446776。URL https://doi.org/10.1145/3446776。

Bo Zhao 和 Hakan Bilen。通过可微孪生增强进行数据集凝缩。*arXiv* 预印本 *arXiv:2102.08259*，2021年。

Bo Zhao、Konda Reddy Mopuri 和 Hakan Bilen。通过梯度匹配进行数据集凝缩。在国际学习表征会议，2021年。URL https://openreview. net/forum?id=mSAKhLYLSsl。

Yongchao Zhou、Ehsan Nezhadarya 和 Jimmy Ba。使用神经特征回归进行数据集蒸馏。载于神经信息处理系统进展论文集（*NeurIPS*），2022年。

Yufan Zhou、Zhenyi Wang、Jiayi Xian、Changyou Chen 和 Jinhui Xu。基于神经切线核的元学习。*CoRR*，abs/2102.03909，2021年。网址：https://arxiv.org/abs/2102. 03909。

Ligeng Zhu、Zhijian Liu 和 Song Han。梯度中的深度泄露。载于神经信息处理系统年会（*NeurIPS*），2019年。

附录

与 Haim 等人（2022年）的比较

In s在第 3 节中，我们提到过，复现 Haim 等人（2022年）的攻击时遇到了问题。这里，w e co比较这两种攻击，并讨论我们发现的他们方法中的问题。

首先，我们比较这两种攻击的质量。Haim 等人（2022）公开了他们的代码，并提供了针对 CIFAR-10 和 MNIST-10 的两组最佳重建图像。我们在此绘制这些重建图像的重建曲线：

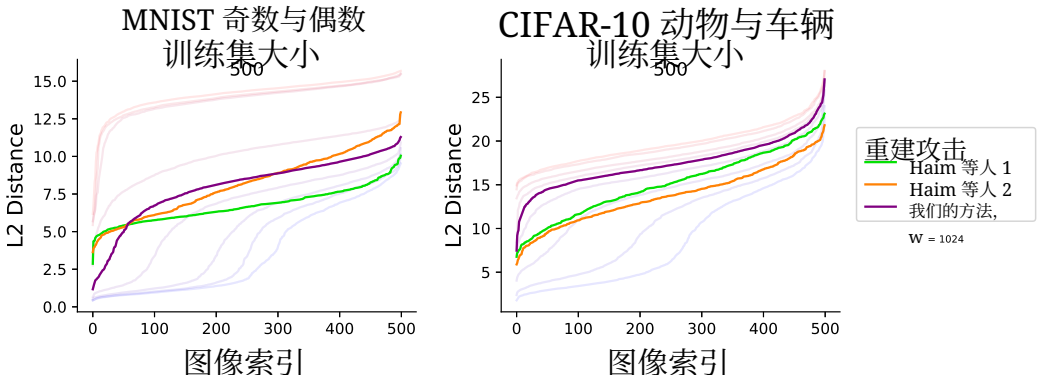


图 10：Haim 等人（2022）所述攻击的重建曲线，与我们的重建攻击进行比较；两者的宽度相当，均为1024。

他们的论文考察了宽度为 1000 的网络，因此，为了公平比较，我们展示了在标准动力学下宽度相当（为 1024）的重建结果。可以看到，对于 MNIST，在图像索引达到 50 之前，我们的攻击质量显著更好；到索引 50 时，两种攻击的表现都很差。对于 CIFAR-10，我们的攻击表现较差，但增大宽度或进行线性化可以提升其表现。

请注意，Haim 等人（2022）展示的两条重建曲线对应于经过精心选择的超参数所得到的两种最佳 重建图像。这些超参数的选择旨在最大限度地提高重建质量，而衡量质量需要访问训练数据。事先而言，攻击者无法评估重建质量，因此也无法进行这样的超参数调优。相比之下，我们没有对攻击参数（例如学习率/初始化）进行微调，并且对每次攻击都使用相同的超参数。经过更细致的调优，可能还会取得更好的效果。我们想强调的是，本研究的目标未必是打造最强的攻击，而更多是探索攻击的特性及其失败或成功的条件。

他们的攻击还有一个局限：要求使用齐次神经网络（即没有偏置），而我们的方法使用偏置，因此更贴近实际。其攻击最大的局限在于要求**采用谨慎的初始化方案**，即第一层的权重以显著较小的方差进行初始化。Haim 等人（2022年）指出，这对于他们攻击的成功**至关重要**。相比之下，我们采用 Flax 神经网络库（Heek 等人，2020年）提供的默认初始化方案。

我们还观察到，式 (1) 中给出的 KKT 条件在实践中往往无法达到，而这是他们的攻击奏效所必需的。重申一下方向收敛的定义：对于训练后的网络参数 θ，以及 θ'（式 (1) 的解），它要求 $\lim_{t\to\infty} \frac{\theta}{\|\theta\|_2} \to \frac{\theta'}{\|\theta'\|_2}$。显然，如果 $\frac{\|\theta-\theta_0\|_2^2}{\|\theta_0\|_2^2} < 1$，也就是说参数并未偏离其初始化值太远，那么我们就无法指望达到 KKT 点（当然，也存在一些不太可能的例外，例如 θ_0 已经接近 KKT 点）。实践中，我们发现运行他们的攻击时，$\|\theta_0\|_2^2 \approx 333$ 和 $\|\theta\|_2^2 \approx 377$，这表明初始化仍然对最终参数有显著影响，也说明尚未达到 KKT 条件。当然，我们的攻击也有其局限性，

其中最显著的一点是，我们需要网络初始化。我们将把消除这一要求留待未来工作。

B 重建攻击算法详解

算法1标准重建攻击

需要： 初始参数 θ_0、最终参数 θ_f、网络函数 f_θ、随机初始化的重建图像和对偶参数 $\{X_R, \alpha_R\}$、优化器 Optim(params, gradients、步数 T

$\Delta\theta = \theta_f - \theta_0$

$t \leftarrow 1$

当 $t < T$ **时执行**

$G \leftarrow \sum_i \alpha_i \nabla_\theta f_{\theta_f}(x_i)$ 对于 $\alpha_i,\ x_i \in \alpha_R,\ X_R$ ▷ 计算重建梯度

$\mathcal{L}_{\text{recon}} = ||\Delta\theta - G||_2^2$ ▷ 计算重建损失

$\alpha_R,\ X_R \leftarrow$ 优化 $\left(\{\alpha_R, X_R\}, \frac{\partial L_{recon}}{\partial\{\alpha_R, X_R\}}\right)$ ▷ 更新重建图像

$t \leftarrow t + 1$

结束循环

算法 2批量重建攻击

需要：初始参数 θ_0、最终参数 θ_f、网络函数 f_θ、随机初始化的重建图像和对偶参数 $\{X_R, \alpha_R\}$、优化器 Optim(params, gradients, 步数 T、批量大小 $|B|$$\Delta\theta = \theta_f - \theta_0$ $G_R = \sum_i \alpha_i \nabla_\theta f_{\theta_f}(x_i)$ 对于 $\alpha_i, x_i \in \alpha_R, X_R$ ▷计算总重建梯度（此步骤也可以分批处理）当 $t < T$时执行从批次中均匀采样 $\alpha_B, X_B \subset \alpha_R, X_R$，批量大小为 $|B|$ $G_B \leftarrow \sum_i \alpha_i \nabla_\theta f_{\theta_f}(x_i)$ 对于 $\alpha_i, x_i \in \alpha_B, X_B$ ▷计算该批次的重建梯度$G_{B,old} \leftarrow$ 分离(G_B)▷存储旧批次梯度 $\mathcal{L}_{\text{recon}} \leftarrow ||\Delta\theta - (G_R - G_{B,old} + G_B)||_2^2$ ▷计算重建损失$\alpha_B, X_B \leftarrow$ Optim$(\{\alpha_B, X_B\}, \partial L_{recon}/\partial\{\alpha_B, X_B\})$▷优化批次图像$G_{B,new} \leftarrow \sum_i \alpha_i \nabla_\theta f_{\theta_f}(x_i)$ 对于 $\alpha_i, x_i \in \alpha_B, X_B$ ▷计算新的批次梯度$G_R \leftarrow G_R - G_{B,old} +$分离$(G_{B,new})$▷更新总重建梯度$t \leftarrow t + 1$结束循环

这里我们讨论攻击的运行时间，假设模型有 P 个参数、 M 张重建图像，并进行 T 次迭代。我们介绍攻击的两个版本：算法1给出的标准版本，以及算法 2 中的小批量版本。这两个版本在数学上等价，但批量版本能够处理更大的重建集，而这些重建集可能无法一次全部放入内存。

对于算法1中介绍的标准攻击版本，需要 $O(MPT)$ 的时间和 $O(MP)$ 的内存，因为朴素做法需要存储每张重建图像对应的梯度（并通过这些梯度进行反向传播）。显然，对于大型数据集，无法一次将整个数据集输入模型并进行反向传播，而我们需要 $M > N$ 重构完整数据集，因此这似乎会带来问题。对于论文中的数据集，这并不是问题；但对于更大的数据集，情况就不同了。

为缓解这一问题，我们在算法 2 中提出了该攻击的小批量版本，其时间复杂度为$O(BPT)$，内存复杂度为 $O(BP)$。该版本的前提是存储重构样本的梯度总和的缓冲值 $G_R = \sum_i^M g_i$，并通过谨慎使用自动微分，仅更新一部分梯度（使用 G_R 的缓冲值并减去批次梯度）。在算法 2 中，detach 函数指的是大多数自动微分库中常见的自动微分计算图分离函数。该方法已实现，且

其表现与原始攻击完全相同，不过本文中的实验结果不需要使用该方法，因为我们处理的是较小的 N。

C 完全恢复的标签条件

如第 3 节所述，为了恢复训练图像，我们要求 $\alpha_i \neq 0$。这里我们讨论实现这一点的条件。我们知道 $\alpha = K^{-1}y$，并且不失一般性，考虑重建索引为 N 的最终图像。聚焦于 α_N 的值，我们的 α 方程变为:

$$\begin{bmatrix}\alpha_{:N-1}\\ \alpha_N\end{bmatrix} = \begin{bmatrix}K_{:N-1,:N-1} & K_{:N-1,N}\\ K_{:N-1,N}^\intercal & K_{N,N}\end{bmatrix}^{-1}\begin{bmatrix}y_{:N-1}\\ y_N\end{bmatrix}$$

$$\begin{bmatrix}\alpha_{:N-1}\\ \alpha_N\end{bmatrix} = \begin{bmatrix}K_{:N-1,:N-1}^{-1} + K_{:N-1,:N-1}^{-1}K_{:N-1,N}QK_{:N-1,N}^\intercal K_{:N-1,:N-1}^{-1} & -K_{:N-1,:N-1}^{-1}K_{:N-1,N}Q\\ -QK_{:N-1,N}^T K_{:N-1,:N-1}^{-1} & Q\end{bmatrix}\begin{bmatrix}y_{:N-1}\\ y_N\end{bmatrix}$$

对于 $Q = (K_{N,\,N} - K_{:N-1,\,N}K^{-1}{}_{:\,N-1,\,:\,N-1}K^\intercal{}_{:\,N-1,\,N})^{-1}$。令 $\alpha_N = 0$:

$$\begin{aligned}0 &= -QK_{:N-1,N}^T K_{:N-1,:N-1}^{-1}y_{:N-1} + Qy_N\\ y_N &= K_{:N-1,N}^T K_{:N-1,:N-1}^{-1}y_{:N-1}\end{aligned}$$

注意到$K^T_{:N-1,\,N}K^{-1}{}_{:\,N-1,\,:\,N-1}y_{:N-1}$ 对应于给定$y_{:N-1}$时对 y_N 的核回归预测，我们可知，$\alpha_N = 0$ 当且仅当剩余数据点已能完美预测 x_N。当然，标签与预测完全一致的概率为 0。

C.1 α的另一种解释

在第 6 节中，我们讨论了如何将 α 参数视为“难以拟合”或“容易拟合”的对象。这里，我们推导出对于均方误差损失，$\alpha_j = \int_0^\infty (y_j - f_{\theta_t}(x_j))dt$ 的形式（相差一个缩放参数）。

$$\begin{aligned}\mathcal{L}(\theta_t) &= \frac{1}{2}\sum_i (y_i - f_{\theta_t}(x_i))^2\\ \frac{\partial\theta_t}{\partial t} &= -\eta\frac{\partial\mathcal{L}(\theta_t)}{\partial\theta_t}\\ \frac{\partial\theta_t}{\partial t} &= -\eta\sum_i (y_i - f_{\theta_t}(x_i))\frac{\partial f_{\theta_t}(x_i)}{\partial\theta_t}\\ \Delta\theta &= \int_0^\infty \frac{\partial\theta_t}{\partial t}dt\\ &= \int_0^\infty -\eta\sum_i (y_i - f_{\theta_t}(x_i))\frac{\partial f_{\theta_t}(x_i)}{\partial\theta_t}dt\\ &= -\eta\sum_i\left[\int_0^\infty (y_i - f_{\theta_t}(x_i))\frac{\partial f_{\theta_t}(x_i)}{\partial\theta_t}dt\right]\\ &\approx -\eta\sum_i\left[\int_0^\infty (y_i - f_{\theta_t}(x_i))dt\frac{\partial f_{\theta_0}(x_i)}{\partial\theta_0}\right] \quad \text{(Frozen Kernel approximation)}\end{aligned}$$

注意到 $\Delta\theta = \sum_i \alpha_i{}^{\partial f_{\theta_0}(x_i)\,\partial\theta_0}$（等式 (5)）并匹配各项，同时注意到 $P > T$（参数多于训练点），可知当 $\alpha_i = \overline{-\eta\int_0^\infty (y_j - f_{\theta_t}(x_j))dt}$ 时，该方程组有唯一解。我们可以通过绘制计算得到的 α_i 与 $\int_0^\infty (y_j - f_{\theta_t}(x_j))dt$ 的关系来实验验证这一点，见图 11。

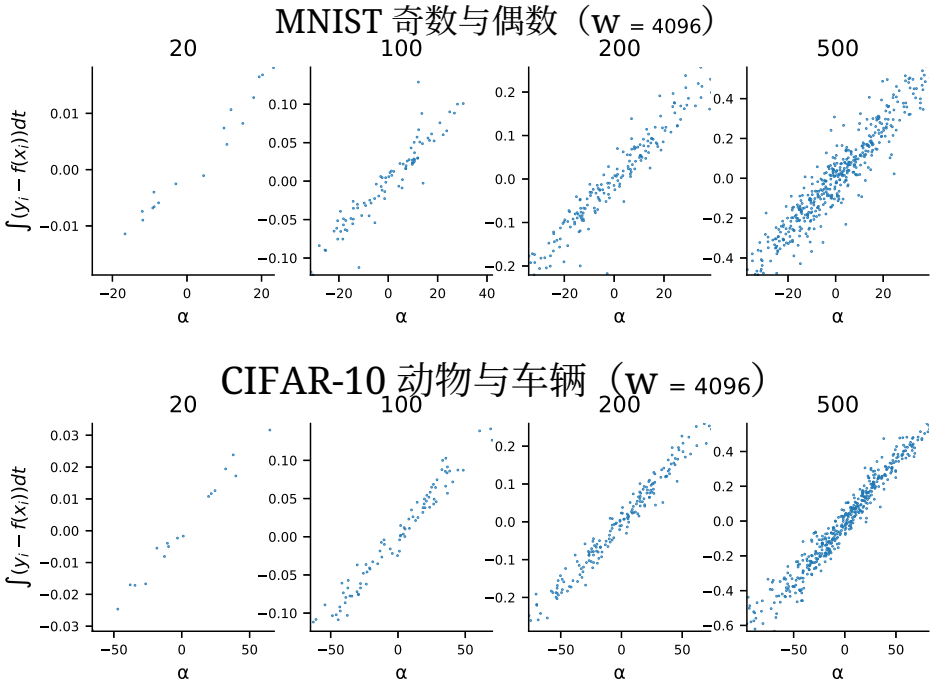


图 11： α 的无限宽度值与由 4096 宽度的有限网络计算出的误差积分形式误差之间的关系。两者呈现强相关性。

D 定理 2 的证明：从重建损失推导 RKIP

证明。这里我们推导如何从式（7）给出的重建损失得到定理 2，即 RKIP 损失。我们在此重述这些损失：

$$\mathcal{L}_{\text{Reconstruction}} = \|\Delta\theta - \alpha^{\intercal}\nabla_{\theta} f_{\theta_0}(X_T)\|_2^2$$
$$\mathcal{L}_{\text{RKIP}} = \|y_T - K_{TR}K_{RR}^{-1}y_R\|_{K_{TT}^{-1}}^2$$

首先注意到 $\alpha^T = K_{TT}^{-1}y_T$；如果我们使用标签为 y_R 的重建图像进行训练，那么 $\alpha^R = K_{RR}^{-1}y_R$。为简洁起见，我们记 $s_i^* = \{\alpha_i^*, x_i^*\}$ 和 $S^* = \{\alpha^*, X_*\}$。令 S^T 和 S^R 分别表示训练集和重建集

$$\begin{aligned}
&\left\|\Delta\theta - \sum_{s_j^R \in S^R} \alpha_j^R \nabla_{\theta_f} f_{\theta_f}(x_j^R)\right\|_2^2 \\
&= \left\|\sum_{s_i^T \in S^T} \alpha_i^T \nabla_{\theta_0} f(x_i^T) - \sum_{s_j^R \in S_R} \alpha_j^R \nabla_{\theta_f} f_{\theta_f}(x_j^R)\right\|_2^2 \\
&= \left\|\sum_{s_i^T \in S^T} \alpha_i^T k_{\theta_0}(x_i^T, \cdot) - \sum_{s_j^R \in S_R} \alpha_j^R k_{\theta_f}(x_j^R, \cdot)\right\|_2^2
\end{aligned}$$

再次取无限宽度极限，使得 $k_{\theta_0}, k_{\theta_f} \to k_{NTK}$，我们为简洁起见将其记作 k。

$$
\begin{aligned}
\Big\| \sum_{s_i^T \in S^T} \alpha_i^T k_{\theta_0}(x_i^T, \cdot) - \sum_{s_j^R \in S_R} \alpha_j^R k_{\theta_f}(x_j^R, \cdot) \Big\|_2^2 &\to \Big\| \sum_{s_i^T \in S^T} \alpha_i^T k(x_i^T, \cdot) - \sum_{s_j^R \in S_R} \alpha_j^R k(x_j^R, \cdot) \Big\|_2^2 \\
&= \sum_{s_i^T \in S^T} \sum_{s_j^T \in S^T} \alpha_i^T \alpha_j^T k(x_i^T, x_j^T) - 2 \sum_{s_i^T \in S^T} \sum_{s_j^R \in S^R} \alpha_i^T \alpha_j^R k(x_i^T, x_j^R) \\
&\quad + \sum_{s_i^R \in S^R} \sum_{s_j^R \in S^R} \alpha_i^R \alpha_j^R k(x_i^R, x_j^R) \\
&= \alpha^{T\intercal} K_{TT} \alpha^T - 2\alpha^{T\intercal} K_{TR} \alpha^R + \alpha^{R\intercal} K_{RR} \alpha^R \\
&= y_T^\intercal K_{TT}^{-1} K_{TT} K_{TT}^{-1} y_T - 2 y_T^\intercal K_{TT}^{-1} K_{TR} K_{RR}^{-1} y_R + y_R^\intercal K_{RR}^{-1} K_{RR} K_{RR}^{-1} y_R \\
&= y_T^\intercal K_{TT}^{-1} y_T - 2 y_T^\intercal K_{TT}^{-1} K_{TR} K_{RR}^{-1} y_R + y_R^\intercal K_{RR}^{-1} y_R \\
&= y_T^\intercal K_{TT}^{-1} y_T - 2 y_T^\intercal K_{TT}^{-1} K_{TR} K_{RR}^{-1} y_R + y_R^\intercal K_{RR}^{-1} K_{RT} K_{TT}^{-1} K_{TR} K_{RR}^{-1} y_R \\
&\quad - y_R^\intercal K_{RR}^{-1} K_{RT} K_{TT}^{-1} K_{TR} K_{RR}^{-1} y_R + y_R^\intercal K_{RR}^{-1} y_R \\
&= \| y_T - K_{TR} K_{RR}^{-1} y_R \|_{K_{TT}^{-1}}^2 + y_R^\intercal (K_{RR}^{-1} - K_{RR}^{-1} K_{RT} K_{TT}^{-1} K_{TR} K_{RR}^{-1}) y_R \\
&= \underbrace{\| y_T - K_{TR} K_{RR}^{-1} y_R \|_{K_{TT}^{-1}}^2}_{\text{label matching}} + \underbrace{y_R^\intercal K_{RR}^{-1} (K_{RR} - K_{RT} K_{TT}^{-1} K_{TR}) K_{RR}^{-1} y_R}_{\lambda_{\text{var of } R|T}}
\end{aligned}
$$

□

$\lambda_{\text{var of } R|T}$与 $K_{RR} - K_{RT} K_{TT}^{-1} K_{TR}$成正比，即 $K_{[T,R],[T,R]}/K_{TT}$，它是 $K_{[T,R],[T,R]}$在 K_{TT} 条件下的舒尔补。注意，这就是高斯条件方差公式，假设我们是基于 T对 R 进行预测。这个正则项确保重建图像不仅能产生正确的预测（由标签匹配项保证），而且蒸馏数据集图像也不会明显偏离训练分布（以神经正切核衡量）。我们推测，正是这一项使 RKIP 在有限宽度设置下优于 KIP。由于没有任何机制直接确保 KIP 蒸馏得到的数据点与训练图像保持“相似”（只要求它们具有预测能力），这些蒸馏图像可能更容易受到领域偏移的影响，例如从无限宽度设置转向有限宽度设置。这种有趣的现象或可成为未来研究的课题。

E FINITERKIP 图像



图12：从一个在 500 张 CIFAR-10 图像上训练的网络中重构 20 张图像。重构图像往往与实际训练图像不匹配，并且包含严重损坏。使用这些图像重新训练可获得较高的准确率。

图13和图14展示了从一个最多可能包含500张图像的数据集中重构20张图像所得到的重建图像。这些重建图像由4096个具有线性化动力学的网络生成。

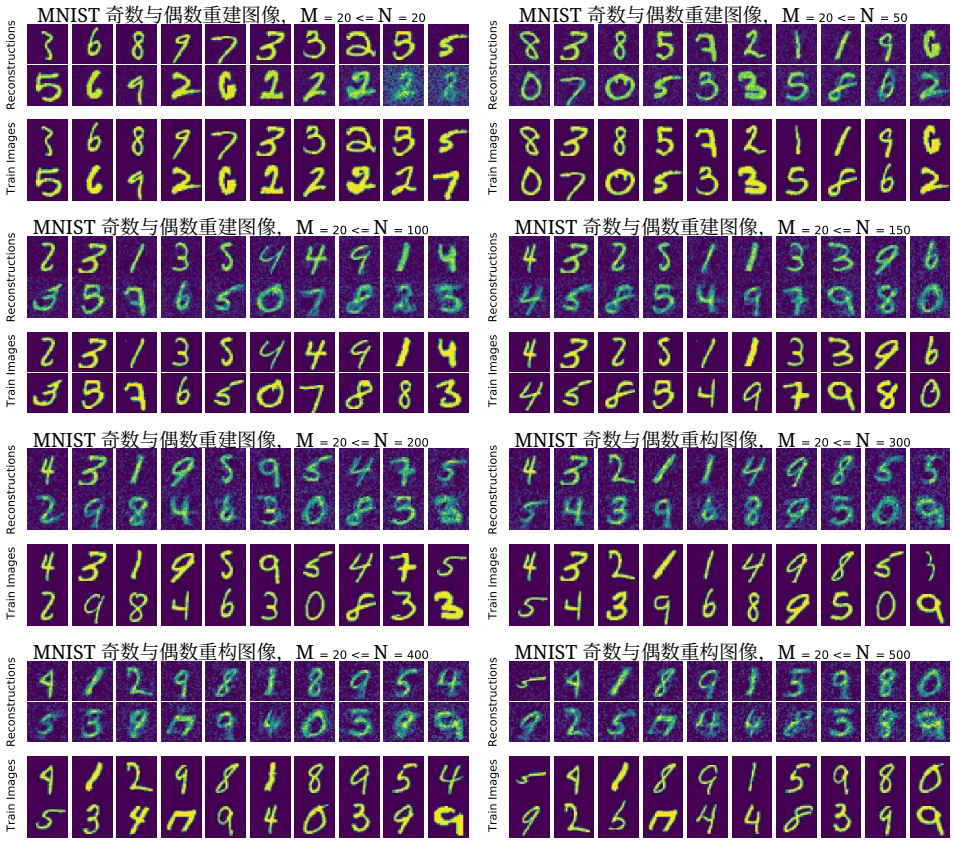


图13：从一个可能包含超过 20 张图像的数据集中重构 20 张图像（MNIST 奇数与偶数）



图14：从可能大于 20 张图像的数据集（CIFAR-10 动物与车辆）中重建 20 张图像

F 多类别重建损失

Here 我们推导出多类别重建损失，并在第4节中使用该损失。我们的最小范数 pr
预测器满足的条件（假设网络处于线性区间 ）：

$$\arg\min_{\Delta\theta} \frac{1}{2}\|\Delta\theta\|_2^2 \quad s.t.$$
$$\forall c \in [C], \quad \Delta\theta^\intercal \nabla_\theta f^c_{\theta_0}(X_T) = y^c_T - f^c_{\theta_0}(X_T). \tag{10}$$

其中， $f^c_{\theta_0}$ 指网络在第 c类上的网络输出， y^c_T 指第 c类的训练标签。由于网络在梯度流下表现为惰性区间，网络将收敛到最小范数解（与初始化之差的范数）（de Azevedo, https://math.stackexchange.com/users/339790/rodrigo-de azevedo）。使用对偶变量将该方程写成完整的拉格朗日函数形式为

$$\mathcal{L}(\Delta\theta, \alpha) = \frac{1}{2}\|\Delta\theta\|_2^2 + \sum_{c=1}^{C} \alpha^{c\intercal}\left(\Delta\theta^\intercal \nabla_\theta f^c_{\theta_0}(X_T) - (y^c_T - f^c_{\theta_0}(X_T))\right). \tag{11}$$

有了 α 我们这组对偶参数 $\in R^{C\times M}$，也就是说，每个类别都有一组 M 对偶参数。对 $\Delta\theta$求导:

$$0 = \nabla_{\Delta\theta}\mathcal{L}(\Delta\theta, \alpha) = \Delta\theta + \sum_{c=1}^{C} \alpha^{c\intercal}\left(\nabla_\theta f^c_{\theta_0}(X_T)\right). \tag{12}$$

因此，我们的多类别重建损失为

$$\mathcal{L}_{\text{reconstrution}} = \left\| \Delta\theta - \sum_{c=1}^{C} \alpha^{c\intercal} \nabla_\theta f_{\theta_0}^c(X_T) \right\|_2^2 . \tag{13}$$

We c可以使用与第 3 节相同的论证，证明这种攻击在无限宽度极限下是精确的 it.

G 提前停止与交叉熵

G.1 提前停止

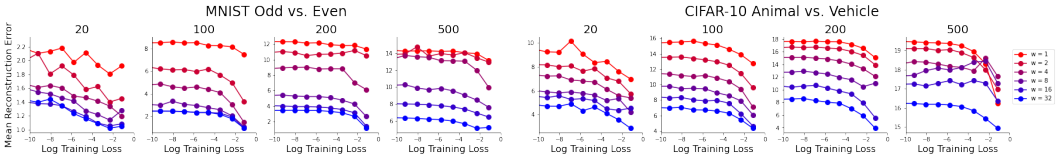


图 15：训练至不同提前停止损失的网络的平均重构误差，范围从10^{-1} 到 10^{-6}

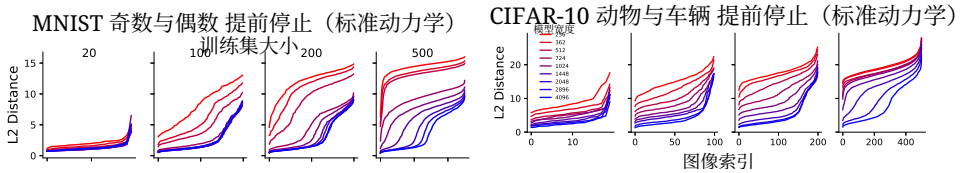


图 16：训练至损失为 1e-2（即明显欠拟合）的网络的重建曲线（采用标准动力学）。与图 2 相比，我们发现提前停止不会影响重建质量，这与理论一致。

正文中，我们考虑了以较低学习率训练 10^6 训练轮次的模型，以达到第 3 节所述的 KKT 条件。在实践中，这种做法计算开销很大，而且由于容易过拟合，往往并不可取；提前停止是一种常见的正则化技术。Haim 等人（2022）提出的攻击有一个主要局限：它要求网络达到 KKT 点才能实施攻击。我们的方法**不要求模型达到收敛**即可成功实施攻击。再次指出，网络参数的时间演化由下式给出：

$$\Delta\theta(t) = \nabla_\theta f_{\theta_0}(X_T)^\intercal \underbrace{K_0^{-1}\left(I - e^{-\eta K_0 t}\right)\left(y_T - f_{\theta_0}(X_T)\right)}_{\text{time-dependent weights, } \alpha(t)}$$

值得注意的是，即使在有限时间内，网络参数的变化仍然是训练集的有限宽度神经切线核特征映射的线性组合，$\nabla_\theta f_{\theta_0}(X_T)$，其权重随时间变化，$\alpha(t)$。请注意，无限宽度下攻击能够成功，依赖于核测度嵌入的单射性，而非 $\alpha(t)$达到收敛值，这意味着提前停止时攻击仍然有效。提前停止攻击的一个限制是，我们未必能够恢复原始训练标签。

我们通过在提前停止的情况下重复第4节中的攻击流程，验证该攻击在实践中有效。我们对平均训练损失从 10^{-1} 到10^{-8} 的网络实施攻击（请注意，这意味着对于 $\{+1,-1\}$ 个标签，输出值的偏差约为 0.45，即在 $\mathcal{L} = 10^{-1}$的情况下严重欠拟合），结果如图 15 所示，$\mathcal{L} = 10^{-2}$ 的具体重建曲线见图 16。我们观察到，提前停止通常会改善 重建质量。我们认为这可能有两个原因：首先，在训练早期，核演化的时间较短，因此网络表现出的网络动力学更接近惰性区间。其次，我们假设，这可能是因为在训练早期，所有数据点的权重都大致相等。

对参数变化的贡献，而在训练后期，某些数据点的影响力更强（见第 6 节和附录 C.1）。当某些数据点的影响力远强于其他数据点时，影响力较弱的数据点可能会被影响力较强的数据点的信号“淹没”。未来的研究可以更深入地探讨提前停止、异常值影响与网络脆弱性之间的复杂关系。我们还指出，在仅进行一次梯度步的极限情况下，我们的攻击对应于梯度泄漏 攻击（Zhu 等，2019）。

G.2 交叉熵损失

Foll根据 Lee 等人（2019年）的研究，我们知道，使用交叉熵损失训练的宽网络也 o exhi会呈现惰性动力学，前提是其宽度足够大。相应的常微分方程为：

$$\begin{aligned}\mathcal{L}(\theta_t) &= -\sum_i y_i \log \sigma(f_{\theta_t}(x_i)) \\ \frac{\partial \theta_t}{\partial t} &= -\eta \frac{\partial \mathcal{L}(\theta_t)}{\partial \theta_t} \\ \frac{\partial \theta_t}{\partial t} &= -\eta \sum_i (\sigma(f_{\theta_t}(x_i) - y_i) \frac{\partial f_{\theta_t}(x_i)}{\partial \theta_t}\end{aligned}$$

与公式（3）不同，这里没有闭式解，不过关键在于 $\Delta\theta$ 仍然是 $\nabla_\theta f_\theta$（x_i）的线性组合。

H 实验细节

H.1 库和硬件

我们使用 JAX、Optax、Flax 和 neural-tangents 库（Bradbury 等人，2018年；Babuschkin 等人，2020年；Heek 等人，2020年；Novak 等人，2020年；2022年）。所有实验均在配备 24Gb 显存的 Nvidia Titan RTX 显卡上运行。

H.2 网络训练

除非另有说明，在真实数据上训练的网络均采用随机梯度下降，动量为 0.9，进行 10^6 次全批量梯度下降迭代。学习率设为 $\eta = N \times 2\text{e}{-7}$，其中 N 为训练图像的数量。对于蒸馏数据，我们使用 $\eta = N \times 6\text{e}{-6}$ 的学习率，其中 N 现在表示蒸馏数据集大小。我们发现，结果并不十分依赖训练期间使用的学习率。此外，如果训练损失小于 $1\text{e}{-10}$，我们就提前终止训练。正文中的每条重建曲线，都是 3 个不同网络在 3 个不同训练数据划分上训练所得结果的平均值。

对于二分类，我们使用 $\{+1, -2\}$ 中的标签；对于 10 类多类别分类，所选类别对应的标签为 0.9，其他类别的标签为 -0.1。

H.3 重建图像

为了生成重建图像，我们将重建图像初始化为标准差为 0.2 的分布，并将对偶参数初始化为在 $[-0.5, 0.5]$ 内均匀随机分布。我们使用 Adam 优化器（Kingma & Ba，2015年），对所有重建图像采用 0.02 的学习率。如附录A所述，对这些设置进行微调可以提升性能。我们对图像进行 80k 次迭代优化。与 Haim 等人（2022年）类似，我们发现，由于平滑梯度损失，在重构过程中使用 softplus 而非 ReLU 很有帮助。我们在训练过程中将 softplus 温度从 10 逐渐升至 200，因此训练结束时实际上使用的是 ReLU。除非另有说明，我们的目标是重建 $M = 2N$ 重建图像，其中 N 为训练集大小。

H.4 蒸馏

我们将蒸馏图像初始化为标准差为 0.2 的分布，并使用 Adam 优化器以 0.001 的学习率进行 50k 次迭代蒸馏。

H.5 微调实验

对于第 5 节中的微调实验，我们使用来自 flaxmodels 的预训练 ResNet-18（Wright, 2022年）。在这些实验中，我们使用 64 位训练，因为我们发现这能提高重构质量。我们以 $1e-5$ 的学习率训练 30000 次迭代。我们使用不带动量的随机梯度下降，并将批归一化层冻结为初始模型所用的参数。我们使用附录 I 中描述的混合损失。

H.6 剪枝实验

对于剪枝实验，每次训练运行训练 10^5 个训练轮次，学习率为$n \times 3\text{e}-6$，并使用动量为 0.9 的随机梯度下降。对于重构，我们使用正文中的相同重建攻击，进行 30000 次迭代。每次训练迭代中，我们将网络宽度设为 $w = 55\sqrt{n}$，因为我们发现这样通常能使不同 n 值下的攻击强度保持一致。如果攻击过强，就会有太多训练点被重构，我们也就无法再区分哪些示例“易于重构”，因为所有示例都会被同样好地重构。

H.7 重建后处理

我们不会对重建图像进行后处理，重建曲线和可视化均基于未经修改的重建图像。请注意，Haim 等人（2022）采用了更复杂的重构方案，包括对重构结果进行重新缩放和取平均，具体内容详见 Haim 等人（2022）的附录。

I 核的选择

在重构过程中，我们可以进行一个小的选择，具体而言，是使用初始化梯度切向空间 $\nabla_{\theta_0} f_{\theta_0}(x_i)$，还是使用最终梯度切向量空间 $\nabla_{\theta_f} f_{\theta_f}(x_i)$。具体来说，我们可以选择优化 $\mathcal{L}_f$ 或 $\mathcal{L}_0$:

$$\mathcal{L}_f = \left\| \Delta\theta - \sum_{s_j^R \in S^R} \alpha_j^R \nabla_{\theta_f} f_{\theta_f}(x_j^R) \right\|_2^2$$

$$\mathcal{L}_0 = \left\| \Delta\theta - \sum_{s_j^R \in S^R} \alpha_j^R \nabla_{\theta_0} f_{\theta_0}(x_j^R) \right\|_2^2$$

在线性化动力学下，这两者相等，因为 $\nabla_\theta f_\theta(x)$不会发生变化。然而，对于标准动力学，两者之间存在微小差异，而且核变化越大，差异也越大。对于正文中的结果，我们使用 $\mathcal{L}_f$。我们发现，与$\mathcal{L}_0$相比，这种方法的性能略好一些，但差异相当小。这也带来了第三种重构损失的选择，我们称之为“混合”损失 $\mathcal{L}_h$，它将参数的变化视为最终核和初始核的混合：

$$\mathcal{L}_h = \left\| \Delta\theta - \sum_{s_j^R \in S^R} \alpha_{j,0}^R \nabla_{\theta_f} f_{\theta_f}(x_j^R) - \sum_{s_j^R \in S^R} \alpha_{j,f}^R \nabla_{\theta_0} f_{\theta_0}(x_j^R) \right\|_2^2$$

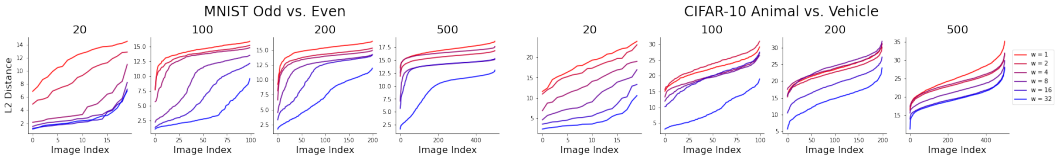
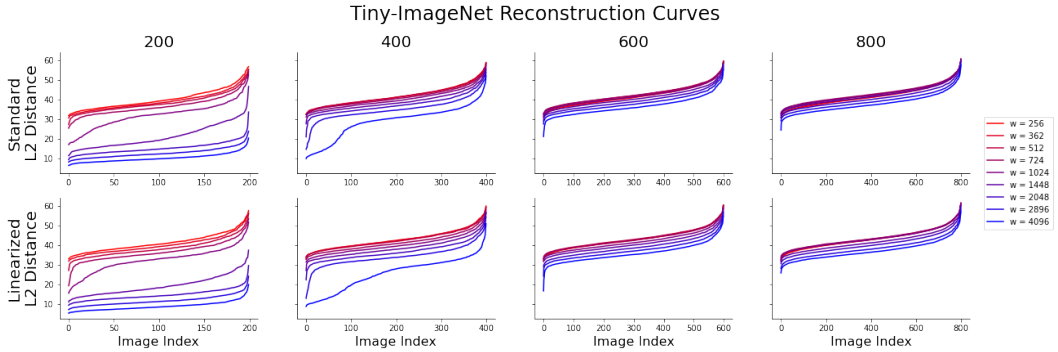


图 17：在 MNIST 奇数与偶数和 CIFAR-10 动物与车辆分类任务上训练的卷积架构的重建曲线，采用不同的宽度乘数，范围为 $1-32$



Fi图 18：在 Tiny-ImageNet 200类分类任务上训练的网络的重建曲线

其中我们有两组对偶参数 α_0 和 α_f。这进一步提升了性能，但通常比优化 $\mathcal{L}_0$ 或 $\mathcal{L}_f$慢一倍，因此我们选择在 ResNet-18 微调实验中使用该损失，因为这一设置更具挑战性。请注意，这种攻击可以推广为沿轨迹包含多个检查点 θ_t 以及更多对偶参数；不过，当然，这需要访问更多训练检查点。

J 附加结果

J.1 卷积架构

这里我们展示了将攻击应用于卷积架构的结果。与正文中一样，我们在二分类 MNIST 和 CIFAR-10 分类任务上训练网络，但使用卷积架构从头开始训练。我们采用常见的 LeNet-5 架构，宽度乘数范围为 1–32。图 17 展示了结果。在两层网络上观察到的结论在这种设置下仍然成立，不过我们的攻击在更深的架构上效果较差。



图 19：Tiny-ImageNet 重构的重建图像及其最近的训练图像对应项，使用宽度为 $n = 400$ 的 $w = 4096$网络

J.2 TINY-IMAGENET

正文主要研究在 MNIST 和 CIFAR-10 等小型低分辨率数据集上的分类。这里，我们通过将攻击应用于 Tiny-ImageNet 分类，考察分辨率更高、复杂度更高的数据集。Tiny-ImageNet 包含 200个类别，图像分辨率为64×64 (Le & Yang, 2015)。由于图像分辨率会对神经正切核近似的质量产生负面影响（Jacot 等，2018），这项实验为检验该攻击在高分辨率图像上的可行性提供了重要的测试场景。我们展示了少样本分类的结果：图 18 中考虑$1 - 4$ 每类图像数，图 19 中展示重建图像。重建更高分辨率的图像更具挑战性，如何改进这一攻击以处理高分辨率图像，是未来研究的一个有趣方向。

J.3 多类分类的核距离与重建质量散点图

图 20 展示了多类分类对应的图 3。正如我们所观察到的，多类分类具有更好的重建质量。从图 20 可以看出，与二分类相比，多类分类的核距离显著更低（MNIST 最大为 1e-2，CIFAR-10 最大为 2e-2；二分类分别最大为 MNIST 的 3e-2 和 CIFAR-10 的 9e-2，见图 3），这或许可以解释为什么重建图像的质量更好。

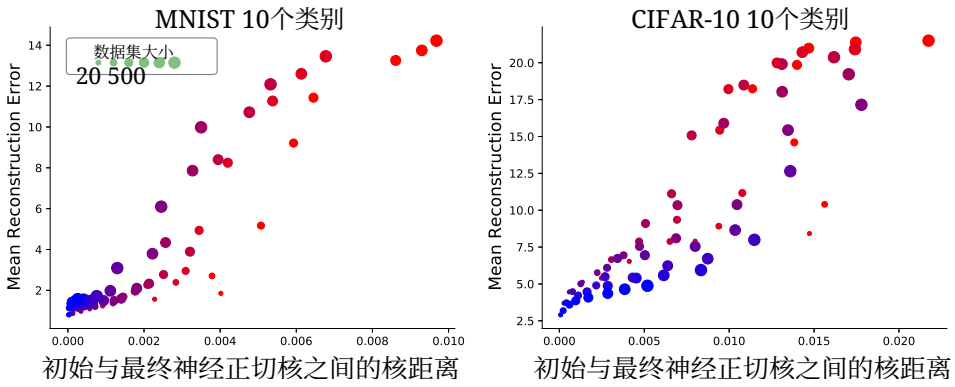


图 20：多类分类中，平均重建误差与从初始化核到最终核的核距离之间的关系。平均重建误差以重建曲线的平均值进行测量，与有限宽度神经切线核在训练过程中的变化程度高度相关。数据集大小由点的大小表示，而颜色表示模型宽度。与二分类相比，多类分类的核距离显著更低（MNIST 最大为 1e-2，CIFAR-10 最大为 2e-2；二分类分别最大为 MNIST 的 3e-2 和 CIFAR-10 的 9e-2，见图 3），这可能是重建质量更好的原因。

J.4 额外剪枝实验

在第 6 节中，我们考察了在每次训练迭代中移除训练数据的类别均衡子集，移除方式要么是随机的，要么是依据重构难易程度。如果允许类别不平衡，我们会看到与第 6 节中相同的初始表现；但随着移除的数据点增多，我们从图 21 中看到，移除容易重构的数据会导致类别严重不平衡，从而使测试准确率较低。要理解为什么某些类别更容易被重构，可能需要参考第 6 节中的讨论。此外，我们发现，如果数据欠拟合，那么无论是剪枝随机数据还是重建图像，测试准确率都没有差异。这表明，只有在训练后期，当部分数据点已得到良好拟合而其他数据点仍欠拟合时，大 α 值的影响才会显现。

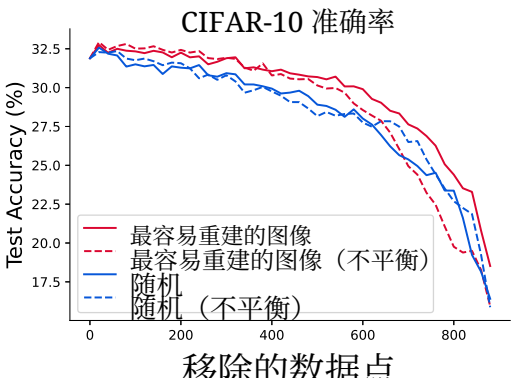


图 21：使用随机剪枝或基于易于重构图像的剪枝，并采用类别均衡子集或非均衡子集时，迭代剪枝的 CIFAR-10 的测试准确率。

J.5 额外重构曲线

我们在图 22 中展示了 MNIST 奇数与偶数以及 CIFAR-10 动物与车辆在 $[20, 50, 100, 150, 200, 300, 400, 500]$ 所有数据集大小下的更多重建曲线。我们在**??**中展示了相同的重建曲线及蒸馏重建。图 23 展示了提前停止时相同的重建曲线。最后，图 24 展示了多分类时相同的曲线。

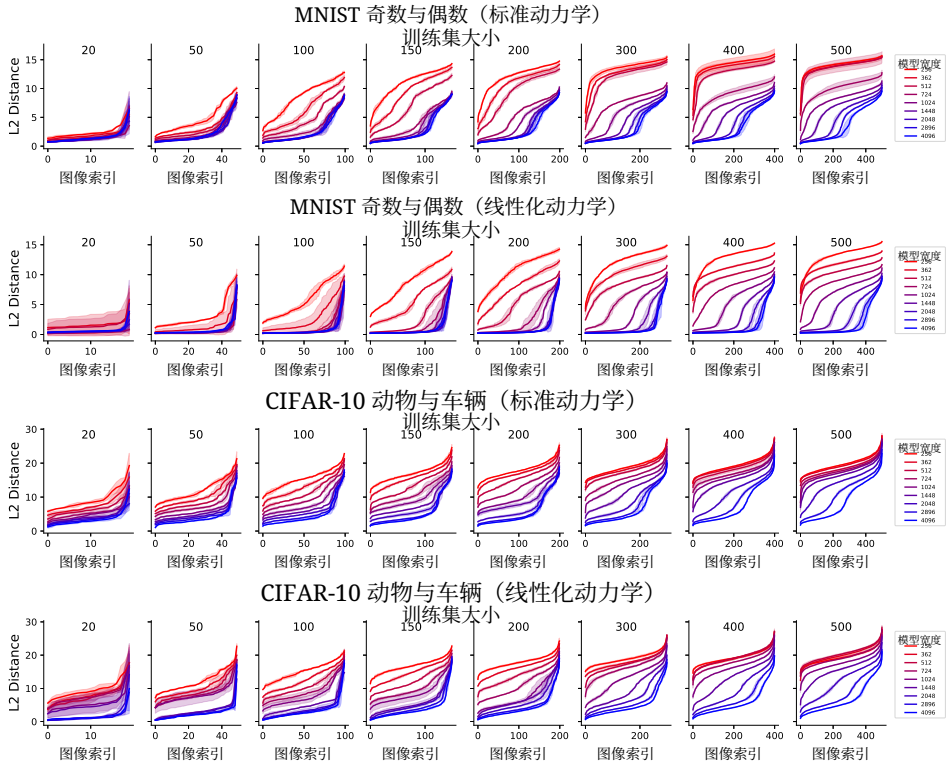


图 22：二分类任务的重建曲线

J.6 重建图像

此处我们展示所有重构图像及其在 L_2 距离意义下最接近的训练图像。图像根据其在重建曲线中的排名排序。

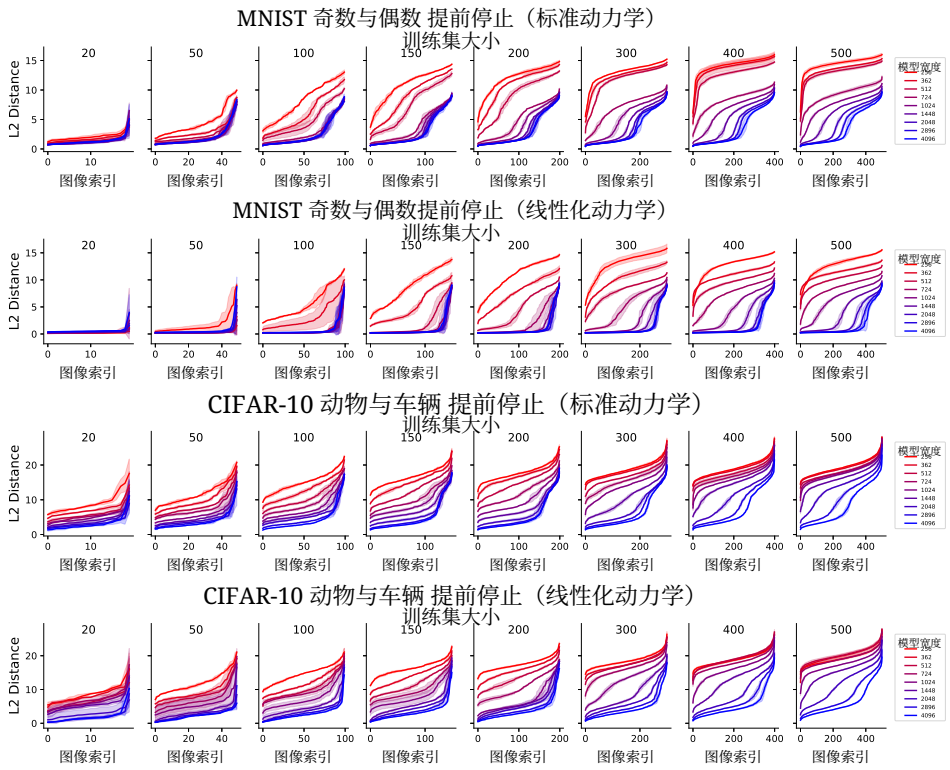


图 23：提前停止条件下二分类任务的重建曲线

J.6.1 二分类

我们在图 25 至图 32 中展示了宽度为 4096 和 1024 的网络在采用线性化或标准动力学时，MNIST 奇数与偶数及 CIFAR-10 动物与车辆任务的重建曲线。

J.6.2 多分类

我们在图 33 至图 40 中展示了宽度为 4096 和 1024 的网络在 MNIST 和 CIFAR-10 十分类任务中的重建曲线，采用线性化动力学或标准动力学。

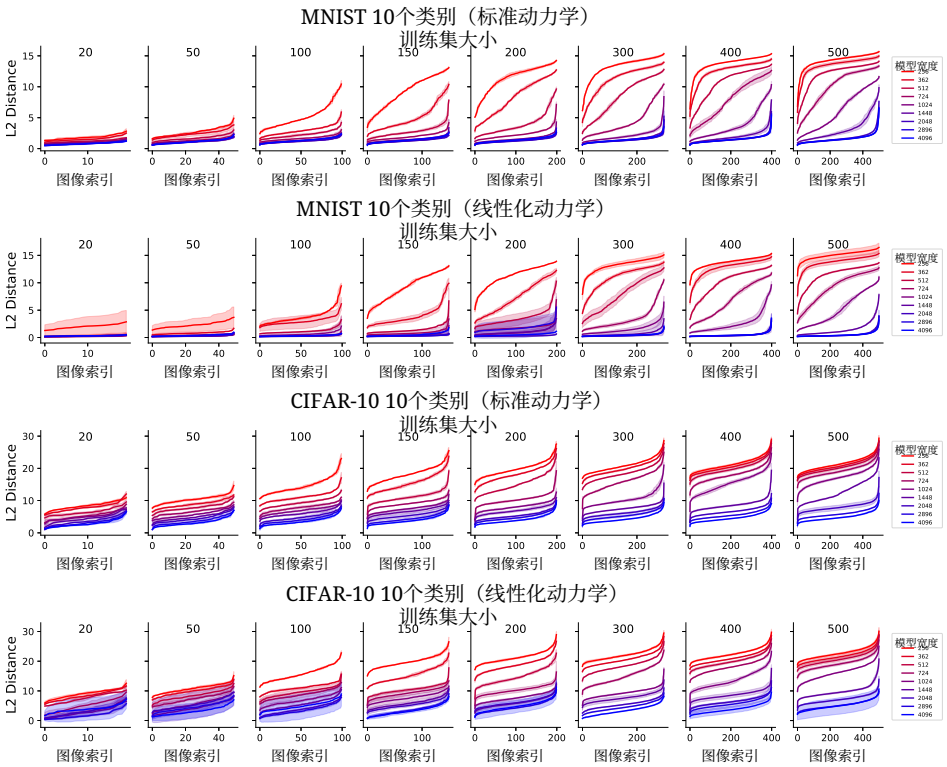


图 24：多类分类的重建曲线

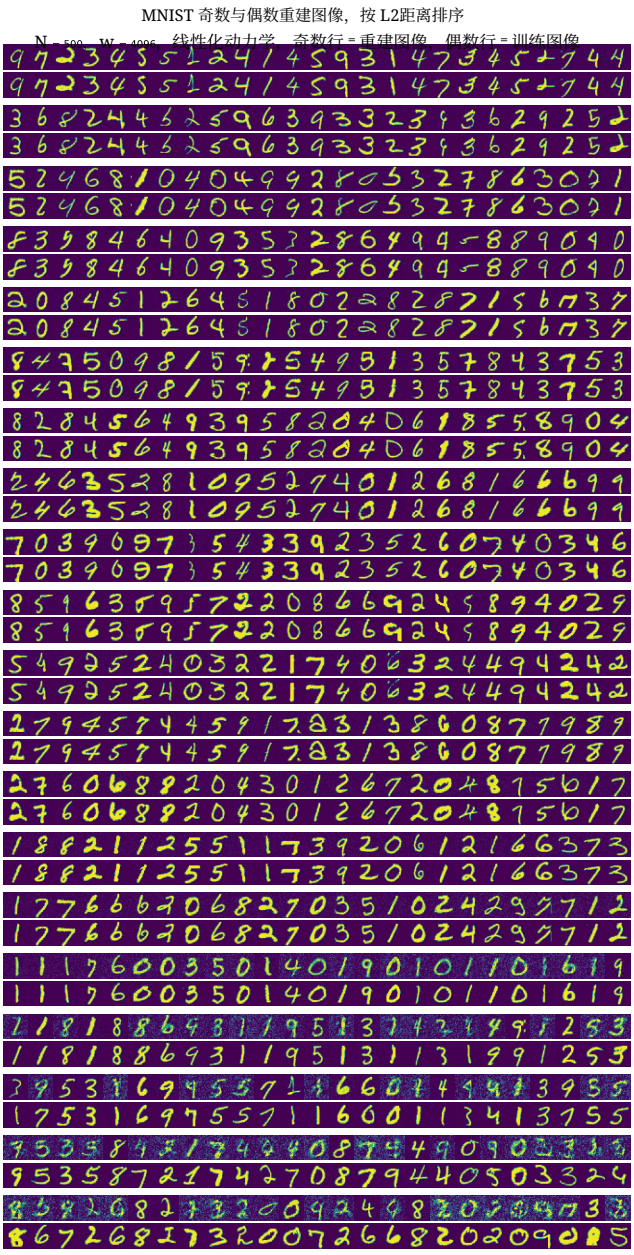


图25：MNIST 奇数与偶数的重建图像，线性化动力学，4096宽度。

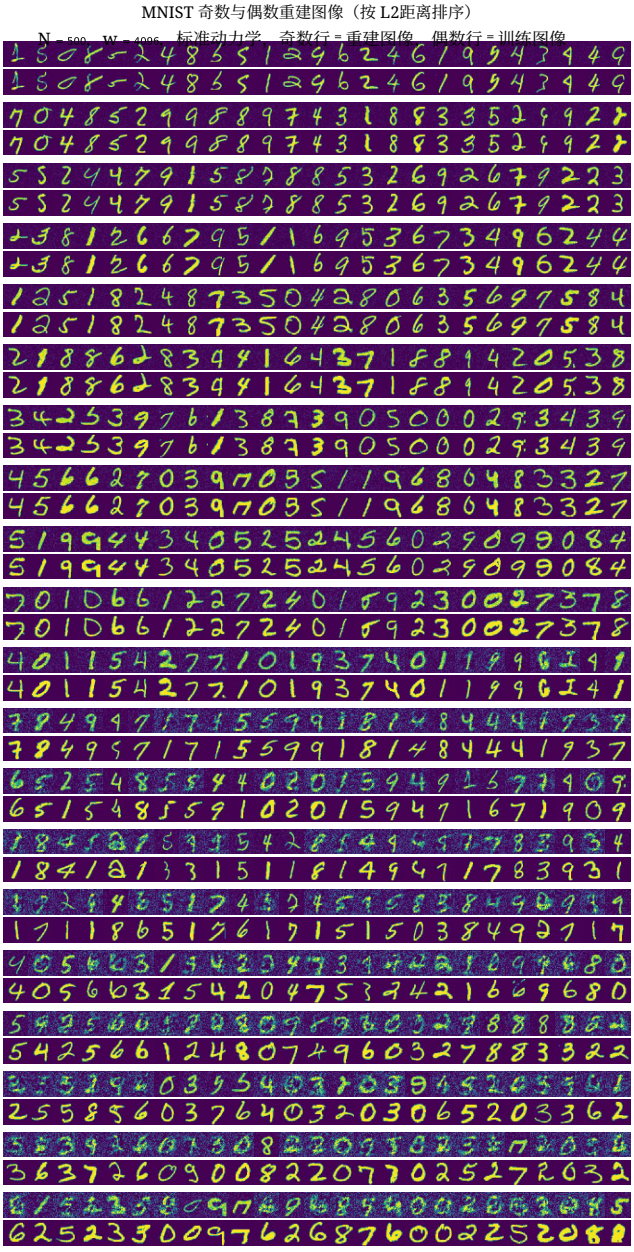


图 26：MNIST 奇数与偶数的重建图像，标准动力学，4096宽度。

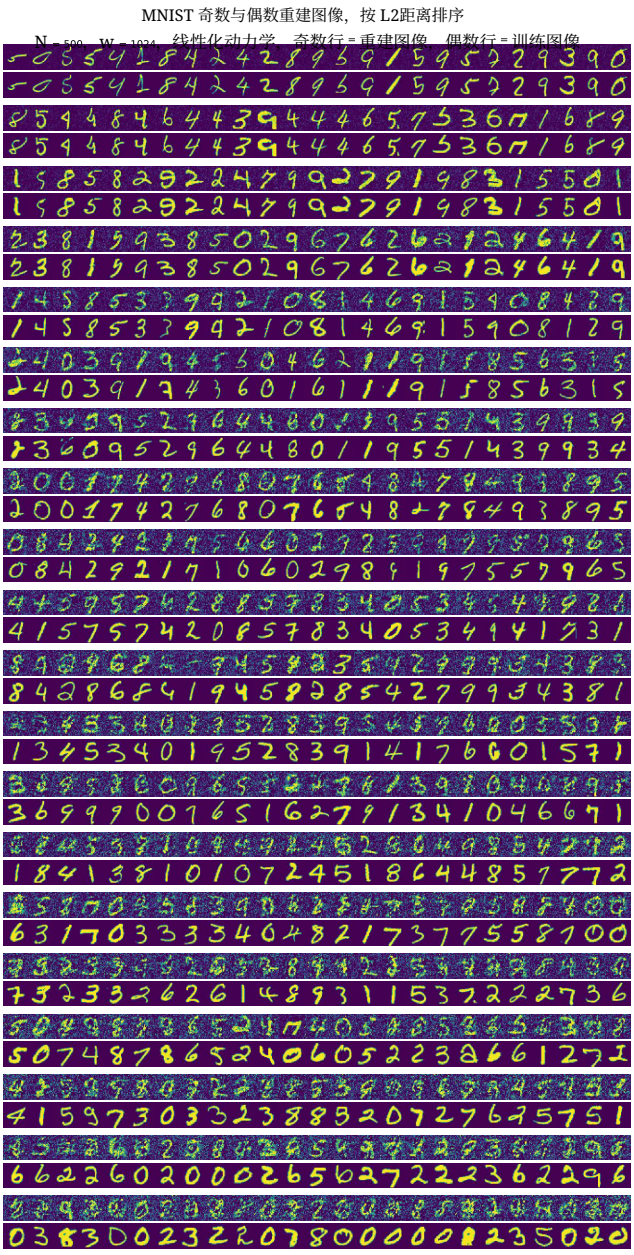


图 27：MNIST 奇数与偶数的重建图像，线性化动力学，宽度为 1024。

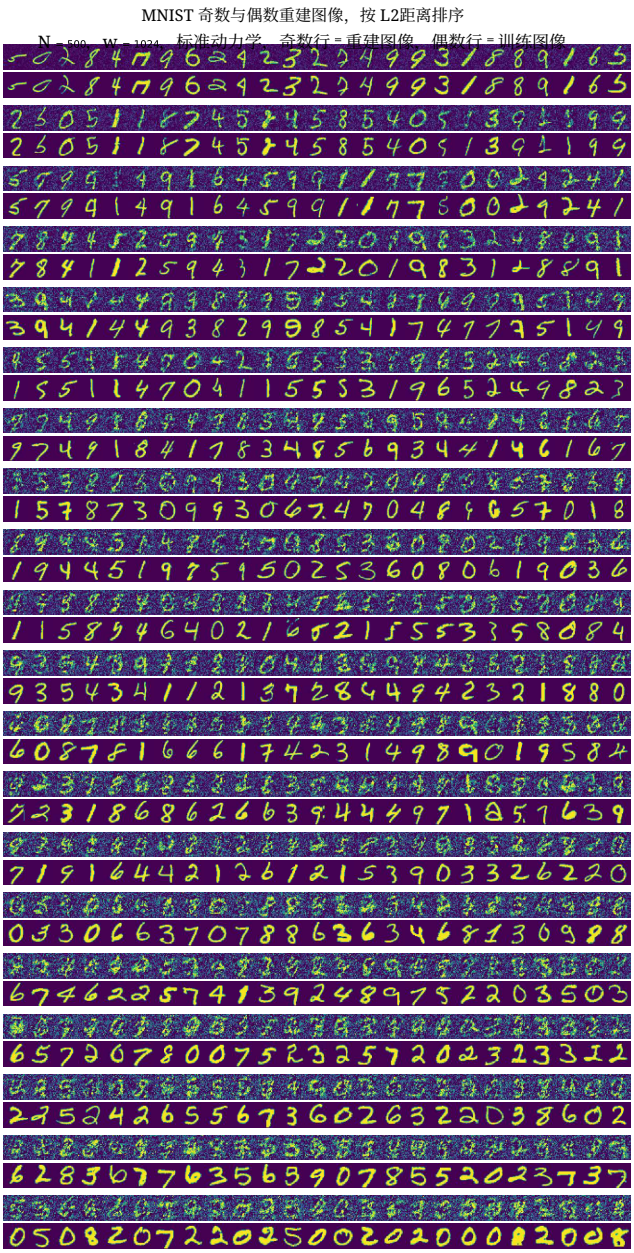


图 28：MNIST 奇数与偶数的重建图像，标准动力学，宽度为 1024。

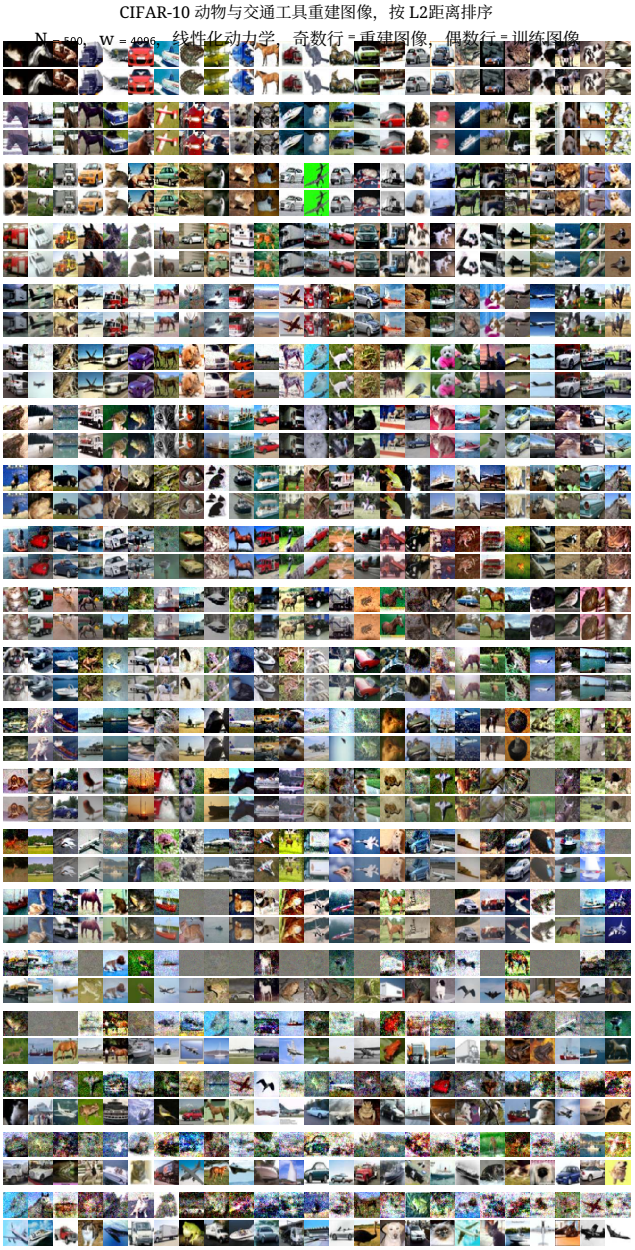


图 29：CIFAR-10 动物与车辆的重建图像，线性化动力学，4096宽度。

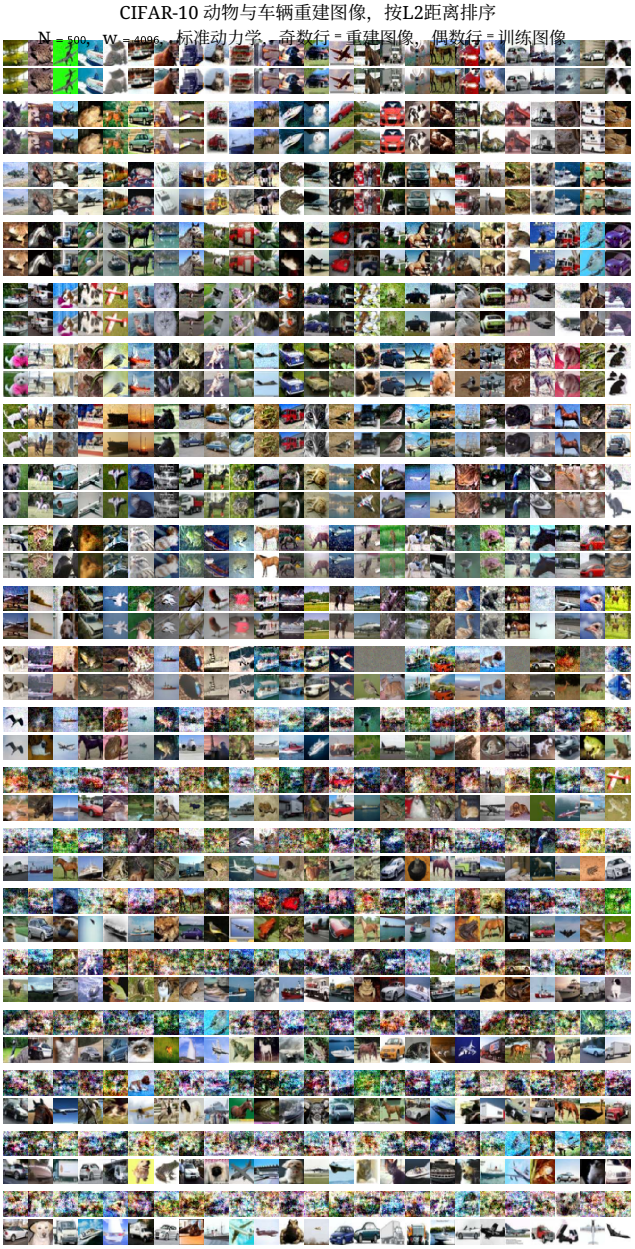


图30：CIFAR-10 动物与车辆的重建图像，标准动力学，4096宽度。

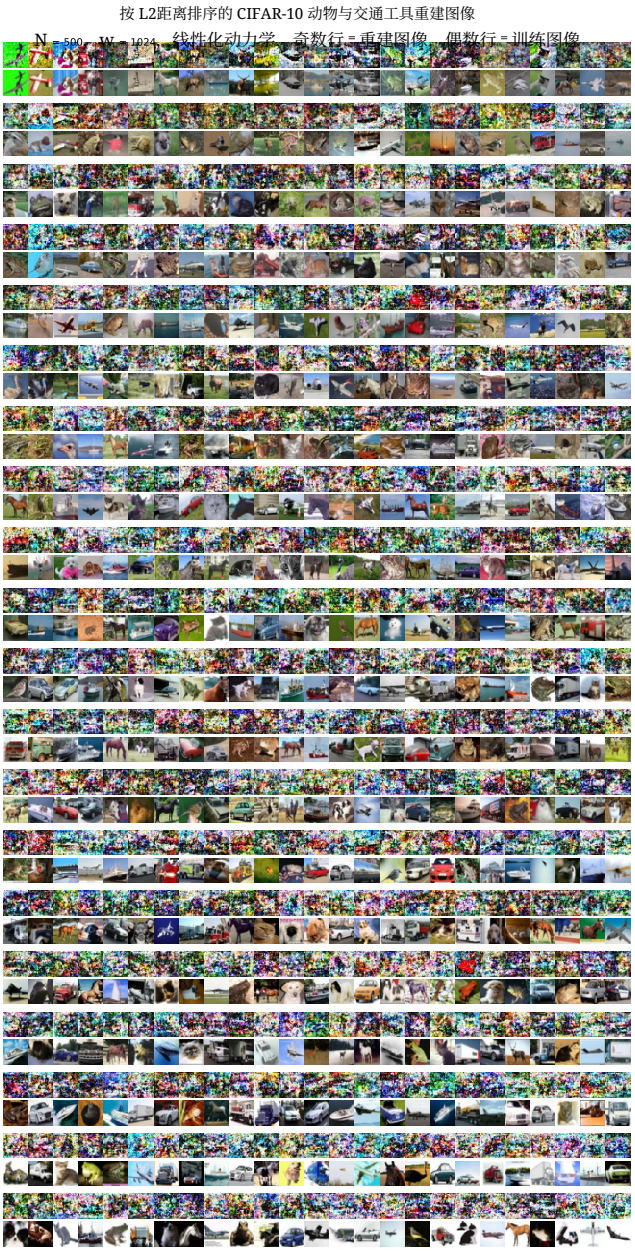


图31：CIFAR-10 动物与车辆的重建图像，线性化动力学，宽度为1024。

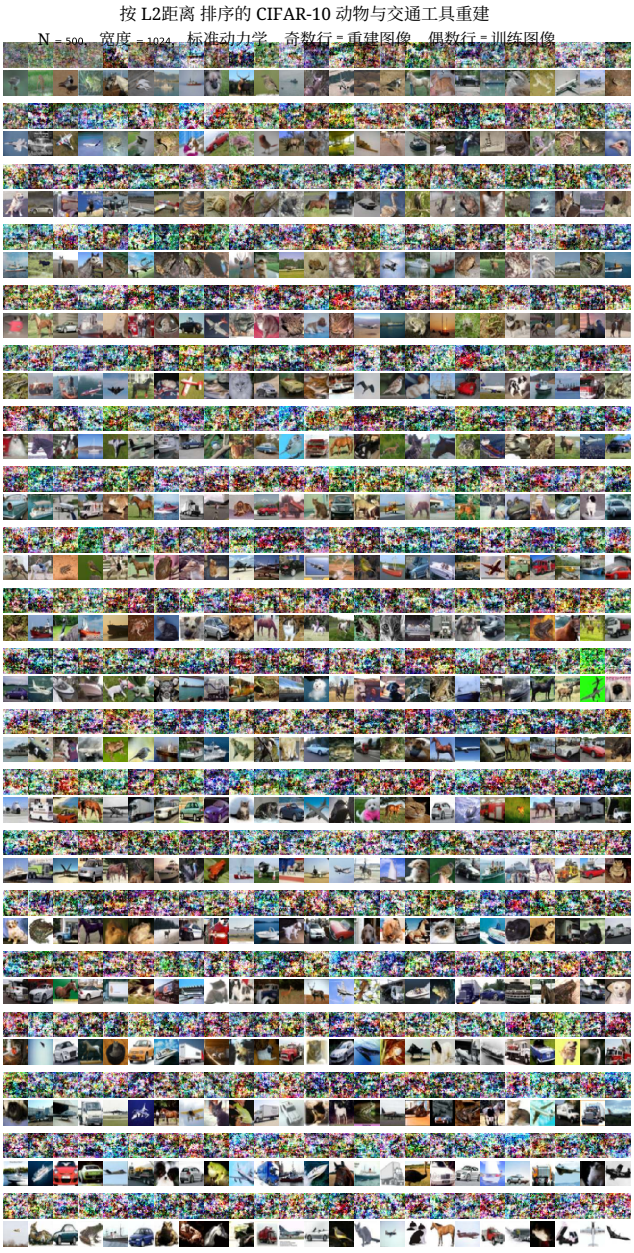


图32：CIFAR-10 动物与车辆的重建图像，标准动力学，宽度为 1024。

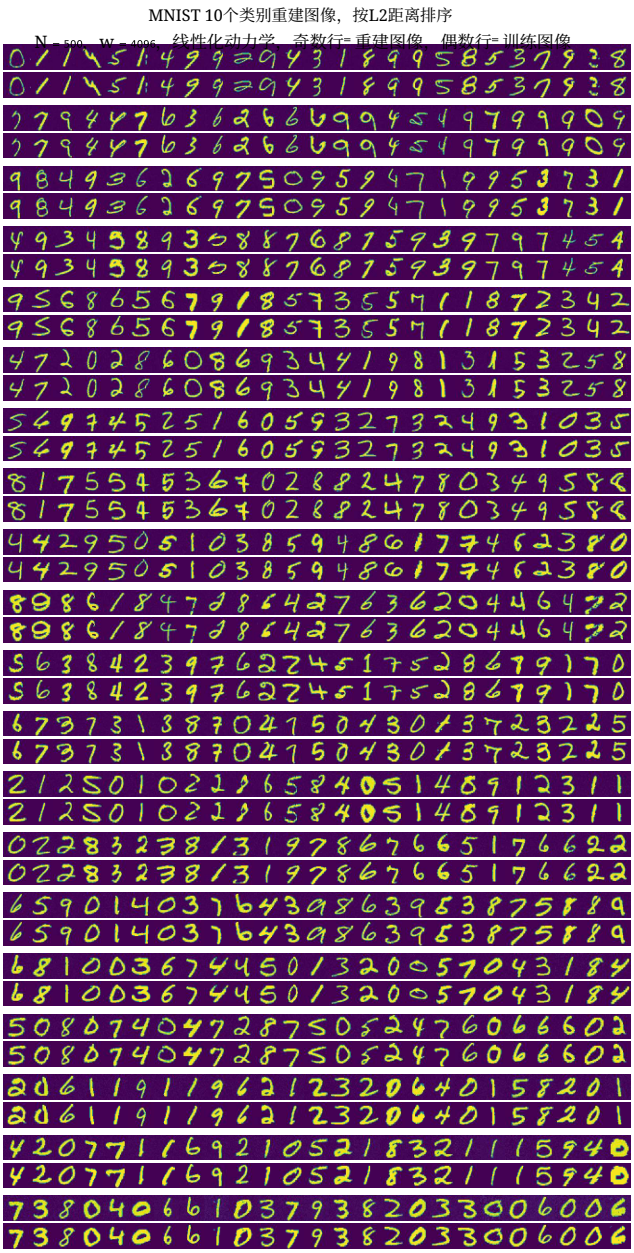


图33：MNIST 10个类别的重建图像，线性化动力学，4096宽度。

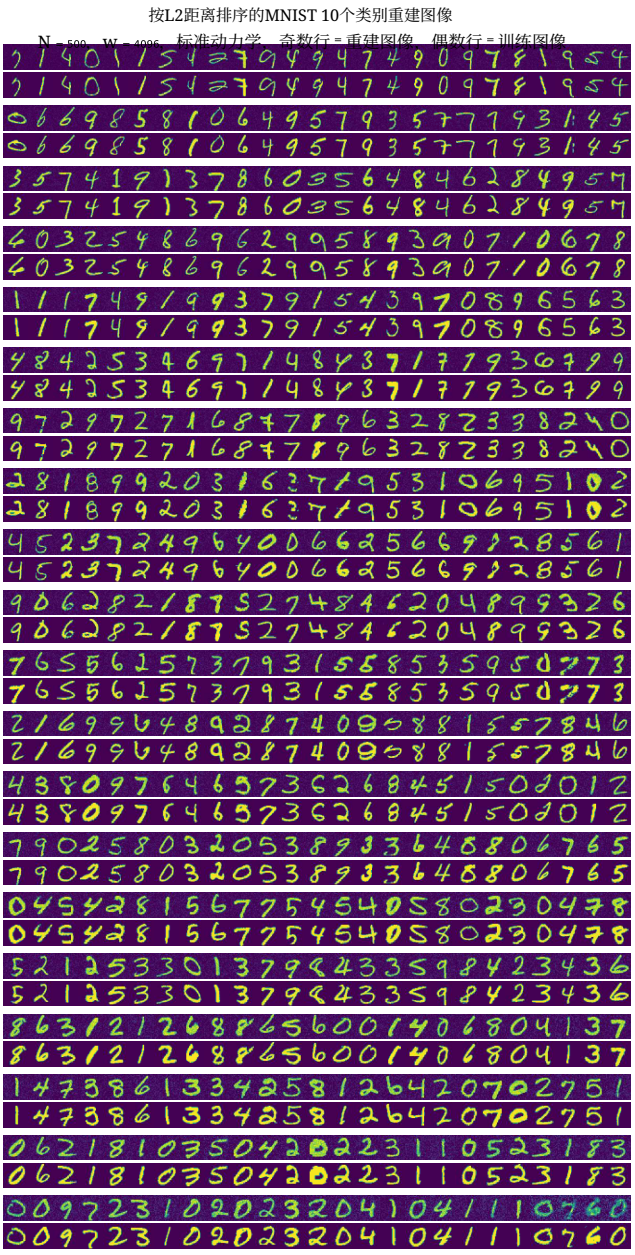


图34：MNIST 10个类别的重建图像，标准动力学，4096宽度。

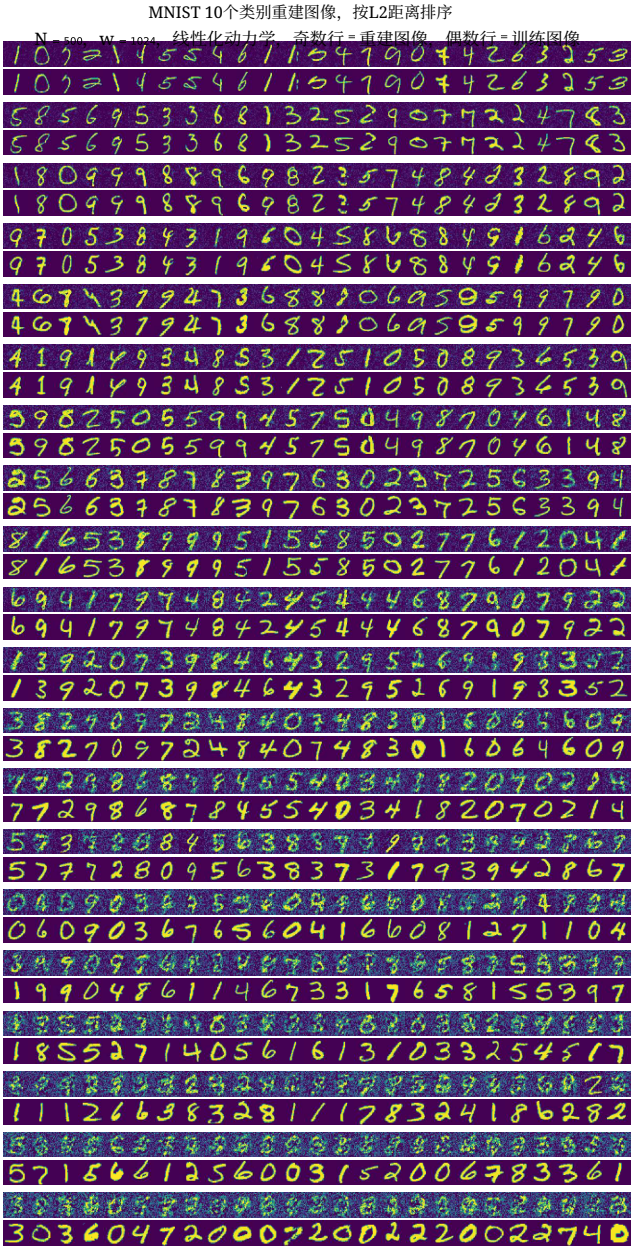


图35：MNIST 10个类别的重建图像，线性化动力学，宽度为1024。

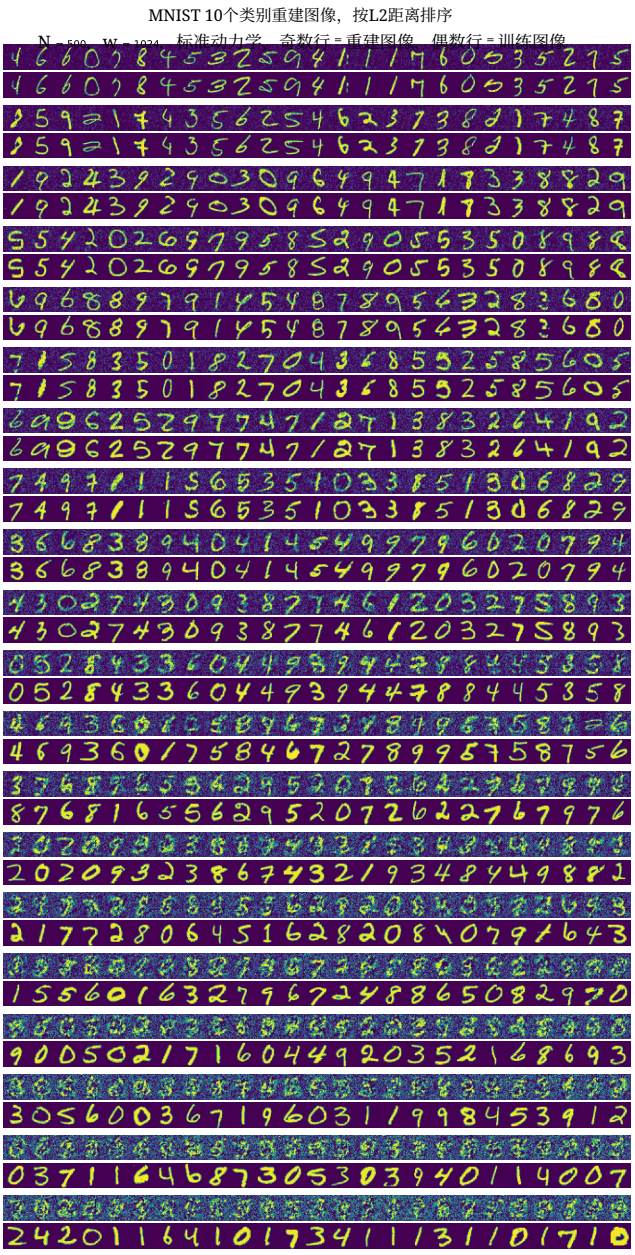


图36：MNIST 10个类别的重建图像，标准动力学，1024宽度。

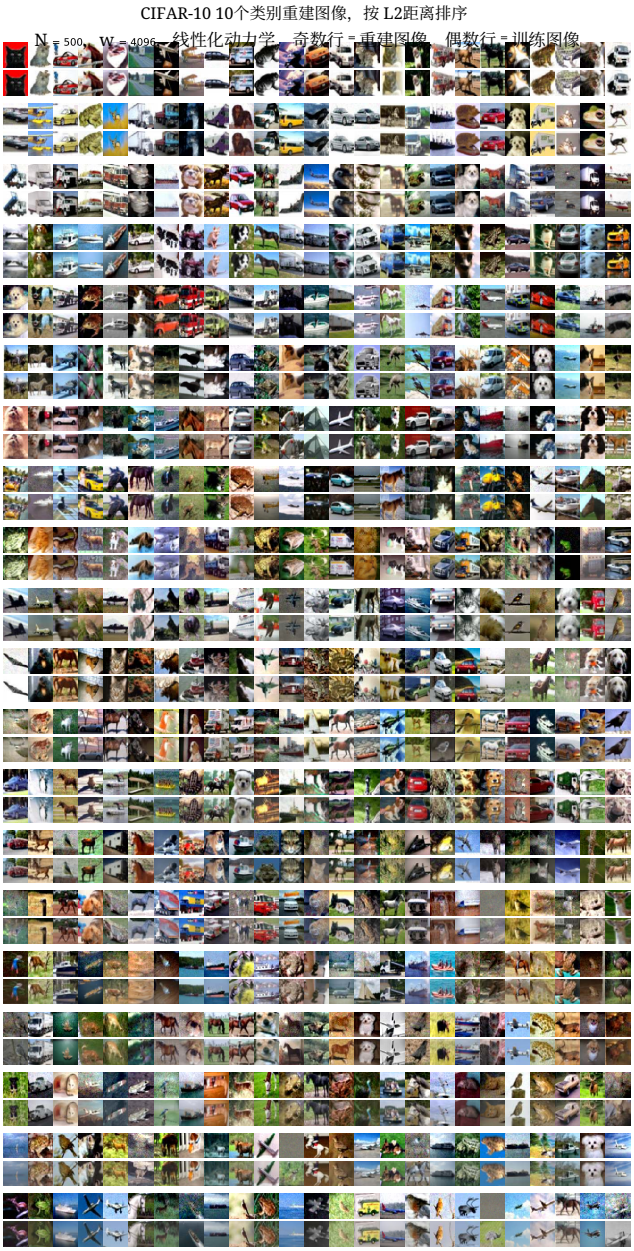


图37：CIFAR-10 10个类别的重建图像，线性化动力学，4096宽度。

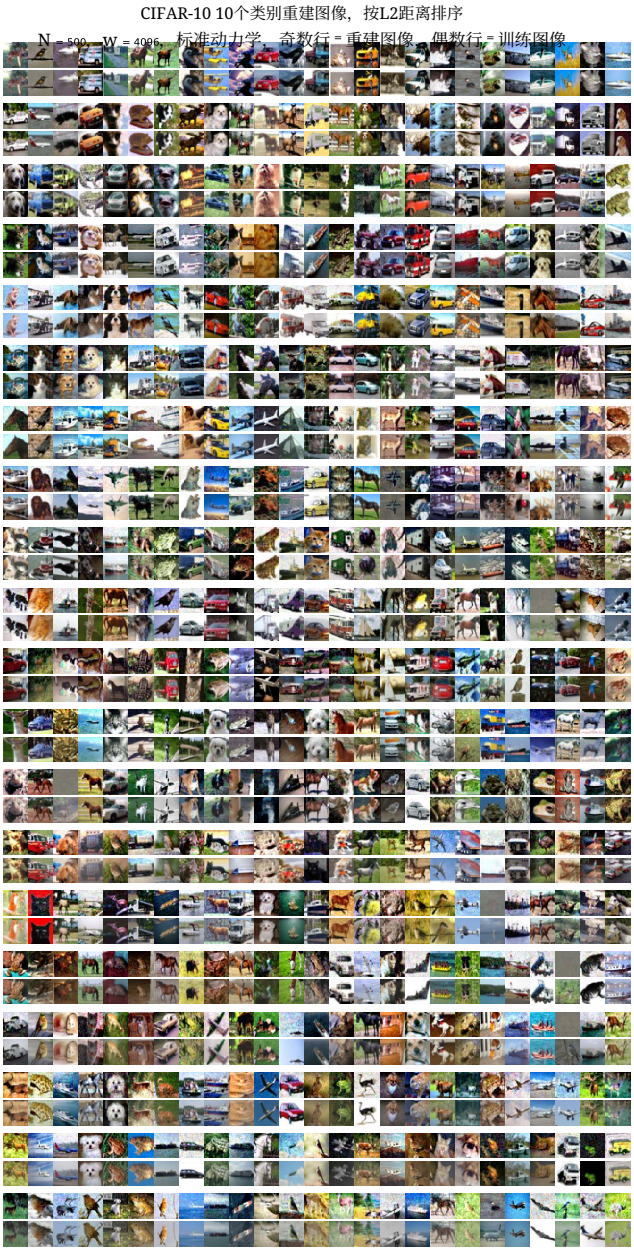
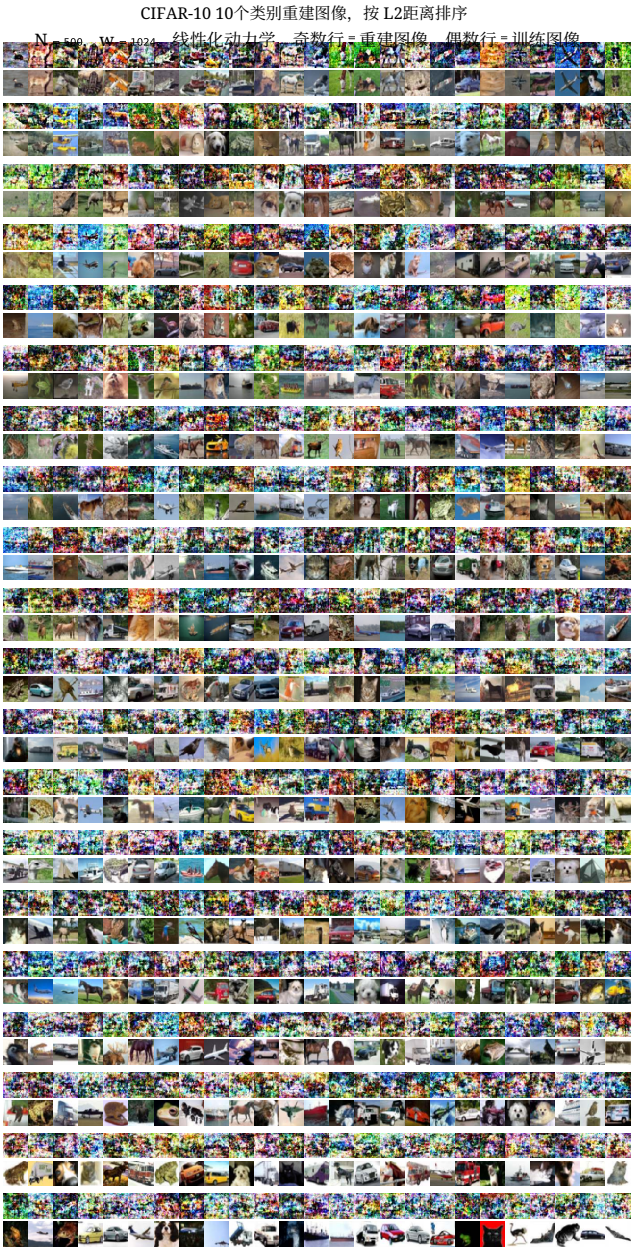


图38：CIFAR-10 10个类别的重建图像，标准动力学，4096宽度。



F图39：CIFAR-10 A10个类别的重建图像，线性化动力学，宽度为1024。

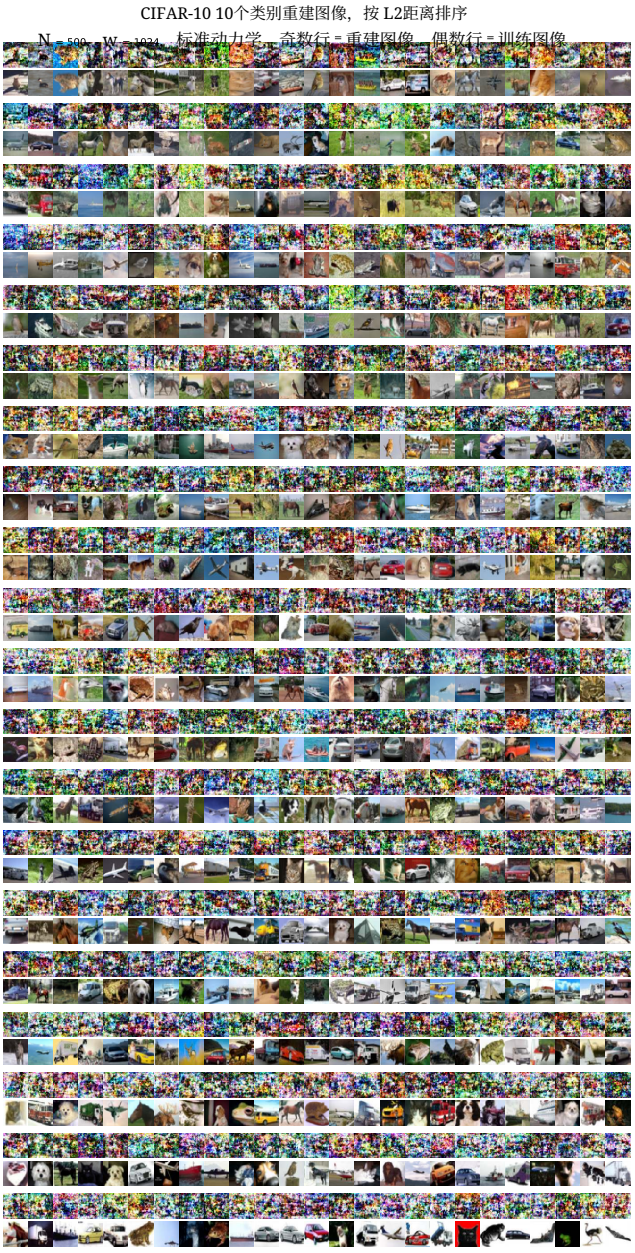


图40：CIFAR-10 10个类别的重建图像，标准动力学，宽度为 1024。