arXiv:2603.00951v1 [cs.LG] 1 Mar 2026

边际钳位何时会影响训练方差？对比式前向-前向学习中的数据集依赖效应

预印本

Joshua Steier*独立
研究者

摘要

对比式前向-前向学习（CFF）逐层训练视觉 Transformer，使其适应监督式对比目标。CFF 训练可能对随机种子敏感，但这种不稳定性的来源尚未得到充分理解。我们聚焦于一个实现细节：对比损失中的正对边际通过饱和相似度钳制 min($s + m$，1) 来实现。我们证明，另一种做法——在对数概率之后减去边际——在正样本均值归约下是梯度中性的。在 CIFAR-10 上（2×2 因子设计，每个单元使用 n=7 个随机种子），钳位使汇总测试准确率方差提高 $5.90\times$（p=0.003），但平均准确率没有差异。对钳制激活率、逐层梯度范数以及缩减边距探针的分析表明，早期层存在由饱和驱动的梯度截断。这种效应并不能直接推广到其他数据集：在 CIFAR-100、SVHN 和 Fashion-MNIST 上，钳位产生的方差相同或更低。两个因素可以解释这种差异。首先，每个批次中的正样本对密度决定了饱和发生的频率。其次，当准确率较高时，任务难度会压缩种子间差异。SVHN 难度扫描在单个数据集上验证了这种交互作用：方差比从高准确率时的 $0.25\times$ 变为激进数据增强下的 $16.73\times$。在准确率适中且每批次包含许多同类样本对的情况下，切换到梯度中性减法参考可消除这种方差膨胀，同时不损失平均准确率。测量第0层钳制激活率，可作为判断问题是否适用的简便方法。

1 引言

要信任一项结果，需要多少个随机种子？对于标准的反向传播训练模型，这个问题一直备受关注 [Dodge 等人，2020年，Bhojanapalli 等人，2021年，Picard， 2021], 但对于层局部训练方法，答案仍未揭晓。本文探讨损失函数中的常见实现选择，是否会系统性地改变某种此类方法的种子间方差。

前向-前向（FF）学习 [Hinton， 2022] 以层局部目标取代端到端反向传播。对比式前向-前向（CFF）将这一思路扩展到视觉 Transformer，通过使用监督对比损失训练各层 [Aghagolzadeh和Ezoji，2025年，Chen 等人， 2025]。由于各层独立优化，CFF 的优化动力学与标准训练有着本质区别。以往研究主要关注平均准确率，并将损失函数的形式视为既定条件。损失函数中的实现细节是否会影响训练方差，尚未经过检验。

聚焦。 我们考察其中一个细节：正对边际的应用方式。现有 CFF 代码库的默认做法是采用饱和相似度钳制，min($s + m$，1)。我们将其与对数概率后减法这一替代方案进行比较，并证明，在正样本均值归约下，减法形式是梯度中性的（命题

*联系邮箱：joshsteier@gmail.com

4.1），因此可作为真正的无边际基线。由此，这项比较能够区分饱和的影响与边际本身的影响。

主要结果。 在 CIFAR-10 上（2×2 因子设计，每个单元有 n=7 个随机种子），钳位使测试准确率的方差增加了 $5.90\times$（p=0.003），但未检测到均值差异。梯度范数和钳制激活率表明，截断发生在早期层，这些层的饱和最为频繁。该效应取决于数据集：CIFAR-100、SVHN 和 Fashion-MNIST 的方差比均呈反转。跨数据集分析和 SVHN 难度扫描确定了两个调节因素：正样本对密度（决定饱和发生的频率）和任务难度（限制不同随机种子之间的分歧幅度）。

贡献。 我们给出了 CFF 监督式对比损失的闭式定义，并明确列出边界变体（第 4 节），其中包括对减法形式的梯度中性性质的证明。随后，我们在 CIFAR-10 上开展了方差审计并提供相关诊断（第 5 和第 6 节），并探究了该效应在四个数据集上何时出现、何时不出现（第 7 节）。

2 背景与相关工作

2.1 前向-前向与层局部学习 ng

前向-前向（FF）学习 [Hinton, 2022] 使用局部目标函数而非单一全局损失来训练网络。它属于更广泛的贪心逐层方法系列，这类方法针对局部目标优化每一层（或模块） [Belilovsky et al., 2019, Nøkland and Eidnes, 2019, L¨owe et al., 2019]。CFF 通过在每一层应用监督对比目标，将 FF 扩展到视觉Transformer [Aghagolzadeh和Ezoji, 2025, Chen et al., 2025]。贪心逐层方法因其表征特性和可扩展性而受到研究，但其不同随机种子之间的稳定性却相对较少受到审视。

2.2 对比学习与监督对比学习

自监督对比目标函数通过在温度缩放softmax中拉近正样本对、推远负样本来学习表征 [van den Oord et al., 2018, Chen et al., 2020, He et al., 2020]。Khosla et al.[2020] 将其扩展到监督式设置，其中所有同类样本对都被视为正样本。CFF 使用的损失遵循这一监督式表述，并采用“正样本均值”归约。监督对比损失的变体在正样本对的归约方式，以及边际或相似度修改如何纳入目标函数方面存在差异 [Graf et al., 2021]；我们的工作考察了其中一种修改方式。

2.3 度量学习与对比学习中的边际

边际广泛用于度量学习，以确保不同类别之间相互分离。人脸识别损失中的加性角度边际和余弦边际，包括 SphereFace [Liu 等人、 2017], CosFace [Wang 等人 2018], 以及 ArcFace [Deng 等人 2019], ，通过非线性运算重塑逻辑值贡献，这可能产生梯度减弱或被截断的饱和区域。我们的设定有两点不同：（i）目标函数是独立应用于每一层的监督对比损失；（ii）前向-前向训练通过停止梯度来确保层局部更新。我们记录了诱发饱和的边际在这种层局部机制中的表现，并以种子间方差而非平均准确率作为主要评估指标。

2.4 训练方差与可复现性

深度学习中的种子间差异已在微调 [Dodge 等人、 2020], 预训练 [Bhojanapalli 等人 2021], 以及更广泛的模型选择 [D’Amour 等人 2022]中得到记录。Picard [2021] 指出，仅随机种子的选择就能显著改变报告指标。反向传播训练模型中的方差来源（权重初始化、数据排序、丢弃掩码）已得到充分研究，但针对层局部训练方法的类似分析几乎不存在。我们的工作弥补了 CFF 领域的这一空白。

2.5 优化中的梯度截断

梯度裁剪与截断既被作为有意设置的正则化项加以研究 [Pascanu 等人，2013年，Zhang 等人，2020] 也被视为激活饱和的意外后果。在我们的设定中，相似度钳制引入了一种依赖于数据且随各层变化的隐式梯度截断，从而有别于显式梯度裁剪策略。

符号	定义
$\mathcal{D}$	带标签数据集 $\{(x_n,\ y_n)\}_{n=1}^{N}$，包含 $y_n \in \{1,\dots,K\}$。
B	小批量大小（每次迭代的图像数量）。
$x_i^{(1)}, x_i^{(2)}$	图像 x_i 的两种随机增强。
L, ℓ	Transformer 层的数量；层索引 $\ell \in \{0,\dots,L-1\}$。
$z_{i,\ell}^{(v)} \in \mathbb{R}^d$	视图 v 在第 ℓ 层的 ℓ_2 归一化均值池化表示（公式 (3)）。
u, v	拼接视图集合 $\{1,\dots,2B\}$ 上的索引（定义 4.1）。
$s_{uv,\ell}$	余弦相似度：$z_u^{\top}{}_{,\ell} z_{v,\ \ell} \in [-1,1]$。
τ	温度参数（$\tau > 0$）。
m_ℓ	层 ℓ 处的非负间隔。
M_{uv}	正样本掩码：$M_{uv} = 1$ 当且仅当 $y\ (u)\ = y\ (v)$ 且 $u \neq v$。
$\mathcal{P}_u$	正样本集合：$\{v:\ M_{uv} = 1\}$。

Ta表 1：核心符号。诊断指标（CAR、梯度范数、VR）定义见第 4.4 节 .

3 问题设定

令 σ 表示一个随机种子，它共同决定一次 CFF 训练运行的权重初始化、数据加载器顺序以及随机增强。固定实验条件 c，并令 $a_c(\sigma)$表示完整两阶段流程（第5.1节）之后在留出集上得到的测试准确率。

主要终点。 给定 n 个独立随机种子 $\{\sigma_1,\dots,\sigma_n\}$，我们测量样本方差

$$\widehat{\mathrm{Var}}_c = \frac{1}{n-1}\sum_{k=1}^{n}\big(a_c(\sigma_k) - \bar{a}_c\big)^2, \tag{1}$$

其中 $\bar{a}_c$ 是样本均值。我们检验边距类型之间总体方差相等的零假设：

$$H_0\colon \mathrm{Var}(\texttt{clamp}) = \mathrm{Var}(\texttt{subtract}). \tag{2}$$

实验因素。 我们采用 2×2 析因设计，交叉考察两个因素：（A）边距类型（截断与减法；第 4.3 节），以及（B）数值稳定模式（分离与直接；公式 (5)）。所有其他流程组件（架构、优化器、学习率、数据增强分布和训练时长）在各实验单元间保持不变。

诊断终点。 为探查观察到的任何方差差异背后的机制，我们还报告：（i）钳制激活率（CAR），用于量化每层发生边距诱导饱和的频率；（ii）逐层梯度 ℓ_2 范数；以及（iii）减小边距的剂量-反应探测，用于检验降低边距调度是否会减弱方差效应。

4 方法

4.1 符号与表示

表 1 总结了主要符号；除非另有说明，所有定义均限于单个层 ℓ。

表示。 令 $h_i^{(v)}{}_{,\ell} \in \mathbb{R}^{T\times d}$ 表示层 ℓ 对示例 i、视图 v的令牌级输出，其中 T 为词元数量。对比表示为

$$z_{i,\ell}^{(v)} = \frac{\bar{h}_{i,\ell}^{(v)}}{\|\bar{h}_{i,\ell}^{(v)}\|_2}, \qquad \bar{h}_{i,\ell}^{(v)} = \frac{1}{T}\sum_{t=1}^{T} h_{i,\ell,t}^{(v)}, \tag{3}$$

因此 $z_i^{(v)}{}_{,\ell} \in \mathbb{R}^d$ 位于单位球面上。

定义 4.1（拼接视图索引）。 给定一个包含 B 张图像、每张图像各有两个视图的小批量，我们构造矩阵 $Z_\ell = [z_{1,\ell}^{(1)},\dots,z_{B,\ell}^{(1)}, z_{1,\ell}^{(2)},\dots,z_{B,\ell}^{(2)}]^\top \in \mathbb{R}^{2B\times d}$，并用 $u \in \{1,\dots,2B\}$为其各行编索引，类别标签 $y\ (u)$ 继承自底层图像。

定义 4.2（停止梯度）。 对于标量或张量 x，sg(x)表示前向传播中的恒等映射，以及反向传播中的零（即在自动微分框架中进行分离）。

4.2 层级局部监督式对比目标

我们将监督对比损失写成闭式形式，以明确边界变体及其梯度行为。

基础逻辑值。给定修正后的相似度 $\tilde{s}_{uv,\ell}$ （第 4.3 节），定义

$$b_{uv,\ell} = \frac{\tilde{s}_{uv,\ell}}{\tau}. \tag{4}$$

Nume数值稳定性偏移。令 $\alpha_{u,\ell} = \max_{k \neq u} b_{uk,\ell}$。我们按行平移逻辑值以防止溢出 :

$$g_{uv,\ell} = \begin{cases} b_{uv,\ell} - \mathrm{sg}(\alpha_{u,\ell}), & (\texttt{detach mode}) \\ b_{uv,\ell} - \alpha_{u,\ell}, & (\texttt{direct mode}). \end{cases} \tag{5}$$

两种模式产生相同的前向值；它们的区别在于梯度是否经过最大值算子传播（定义 4.2）。

逐锚点对数概率。对于锚点 u与索引 $v \neq u$:

$$\log p_{uv,\ell} = g_{uv,\ell} - \log\Big(\sum_{k \neq u} \exp(g_{uk,\ell})\Big). \tag{6}$$

监督对比损失。 使用正样本均值归约:

$$\mathcal{L}_\ell = -\frac{1}{2B} \sum_{u=1}^{2B} \frac{1}{|\mathcal{P}_u|} \sum_{v \in \mathcal{P}_u} \log p_{uv,\ell}. \tag{7}$$

前向-前向局部性。 每个 $\mathcal{L}_\ell$ 仅使用第 ℓ层的参数独立优化；层间梯度被停止梯度操作阻断，因此第 ℓ 层的更新不会传播到更早的层。

4.3 正样本对边界变体

我们比较两种策略，二者的区别在于是否引入相似度饱和。

4.3.1 饱和相似度钳位

截断变体在相似度空间中加入间隔，并将结果上限设为1:

$$\tilde{s}_{uv,\ell} = \begin{cases} \min(s_{uv,\ell} + m_\ell,\, 1), & \text{if } M_{uv} = 1, \\ s_{uv,\ell}, & \text{if } M_{uv} = 0. \end{cases} \tag{8}$$

温度-间隔耦合。 由于逻辑值为 $\tilde{s}/\tau$ （式(4)），相似度空间中的间隔 m_ℓ 在饱和之前相当于 m_ℓ/τ 的有效逻辑值偏移。因此，钳位的实际强度同时取决于 m_ℓ 和τ。

4.3.2 梯度中性减法参考

减法基线根据未修改的相似度（$\tilde{s} = s$）计算对数概率，并在事后减去间隔:

$$\log \tilde{p}_{uv,\ell} = \log p_{uv,\ell} - m_\ell\, M_{uv}. \tag{9}$$

命题4.1（对数概率后减法是梯度中性的）。 设 $m_\ell \geq 0$为一个固定标量（不是模型参数的函数）。在正样本均值损失（7）下，将$\log p_{uv,\ \ell}$ 替换为$\log \tilde{p}_{uv,\ \ell}$ （见（9））， 会使每个锚点项偏移常数 m_ℓ， 并且关于模型参数的所有梯度保持不变。

证明。 对于固定锚点 u:

$$-\frac{1}{|\mathcal{P}_u|} \sum_{v \in \mathcal{P}_u} \log \tilde{p}_{uv,\ell} = -\frac{1}{|\mathcal{P}_u|} \sum_{v \in \mathcal{P}_u} (\log p_{uv,\ell} - m_\ell) = -\frac{1}{|\mathcal{P}_u|} \sum_{v \in \mathcal{P}_u} \log p_{uv,\ell} \;+\; m_\ell.$$

该项 m_ℓ 为常数 相对于所有模型参数（根据假设）为常数，因此求导后为零。 □

表2：**超参数配置**。

参数	值
架构	ViT（d=128、 H=4、 L=8）
图像块大小	4×4
批大小	512
阶段 1（表征学习）	600 个训练轮次，AdamW（η=4×10^{-3}、 β=(0.9,0.999)，wd=10^{-4}）
阶段 2（线性探测）	50 个训练轮次，AdamW（η=5×10^{-4}）
温度 τ	0.15
标准间隔调度	m_0=0.4 $\rightarrow$ m_{L-1}=0.1（跨层线性变化）
低间隔调度	m_0=0.2 $\rightarrow$ m_{L-1}=0.1
数据增强	随机裁剪(32, pad=12)、水平翻转、通道归一化

解读。 减法基线不会重塑梯度，而是作为梯度中性参考。相比之下，钳位（式（8））会截断任何发生饱和的正样本对的梯度。

4.4 诊断指标

方差比（主要终点）。

$$\mathrm{VR} = \frac{\widehat{\mathrm{Var}}(\texttt{clamp})}{\widehat{\mathrm{Var}}(\texttt{subtract})}, \tag{10}$$

$\widehat{\ }$其中，Var 表示不同随机种子下测试准确率的样本方差（式（1））。

钳制激活率（CAR）。令 $\mathcal{P}_\ell = \{(u,v) : M_{uv} = 1\}$ 表示层 ℓ上的正样本对集合。CAR 为

$$\mathrm{CAR}_\ell = \frac{1}{|\mathcal{P}_\ell|} \sum_{(u,v)\in\mathcal{P}_\ell} \mathbb{1}[s_{uv,\ell} + m_\ell > 1]. \tag{11}$$

梯度范数。对于每一层 ℓ 及其参数集 Θ_ℓ:

$$\|\nabla_\ell\| = \sqrt{\sum_{p\in\Theta_\ell} \left\|\nabla_p \mathcal{L}_\ell\right\|_2^2}, \tag{12}$$

即， ℓ_2 范数： $\mathcal{L}_\ell$ 相对于 ℓ第个编码器块中所有参数的完整梯度向量的范数。

5 实验

5.1 实验设置

数据集。 主要分析使用 CIFAR-10，并采用标准的 50k/10k 训练集/测试集划分；我们留出 5k 张训练图像作为验证集，因此得到 45k 张训练图像、5k 张验证图像和 10k 张测试图像。泛化实验（第 7 节）在 CIFAR-100（100 个类别）、SVHN（10 个类别，街景门牌号）和 Fashion-MNIST（10 个类别，灰度图像）上重复进行方差审计。数据集特定的详细信息见第 7 节。

架构和超参数。 表 2 总结了完整配置。我们使用 Vision Transformer [Dosovitskiyet al., 2021]，其嵌入维度为 $d = 128$，包含 4 个注意力头和 $L = 8$ 个编码器层。输入图像（32×32）被划分为 4×4 个图像块，每张图像由此生成 $T = 64$ 个词元。

数据增强（CIFAR-10 和 CIFAR-100）。每张图像都会经过 RandomCrop(32, padding=12) 和 RandomHorizontalFlip 处理，随后进行逐通道归一化（μ =(0.491，0.482，0.447)， σ =(0.202，0.199，0.201)）。不使用 RandAugment、AutoAugment 或颜色抖动。所有条件共享相同的数据增强流程；数据增强的随机抽样由随机种子控制。

表 3：**CIFA R-10 因子结果（标准边距** $0.4 \to 0.1$）。各 n=7 单元格的随机种子测试准确率 .

边距类型	稳定性	均值 (%)	标准差 (%)	Var
截断	分离	78.52	0.927	0.8590
截断	直接	78.44	1.158	1.3408
减法	分离	78.73	0.467	0.2178
减法	直接	78.30	0.224	0.0500

数据增强（SVHN 和 Fashion-MNIST）。SVHN 使用与 CIFAR-10 相同的 12 像素裁剪填充，并采用数据集特定归一化（$\mu =$ (0.438，0.444，0.473)，$\sigma =$(0.198，0.201，0.197)）。Fashion-MNIST 图像（28×28，灰度）会转换为三通道并调整至 32×32，使用 $\mu = \sigma =$(0.5，0.5，0.5) 进行归一化，并通过 RandomCrop(32, padding=4) 和 RandomHorizontalFlip 进行数据增强。较小的裁剪填充（4 像素，而非 12 像素）体现了 Fashion-MNIST 相较于自然图像较低的空间复杂度。

SVHN 难度扫描数据增强。 难度扫描（第 7.4 节）在固定数据集和架构的情况下改变增强强度：(i)简单 （标准）：RandomCrop(32, padding=12) 和 HFlip；(ii) 中等：RandomCrop(32, padding=4)、RandomRotation(10) 和 ColorJitter(0.2, 0.2, 0.2)；(iii) 困难：RandomCrop(32, padding=6)、 HFlip(0.5)、 RandomRotation(15)、RandomErasing(p=0.5, scale=(0.1, 0.3)) 和 ColorJitter(0.4, 0.4, 0.4, 0.1)。

训练协议。 CFF 训练分两个阶段进行：

1. 表征学习。 每个 L=8 层均使用各自的监督对比损失（式（7））训练600轮。2. 线性探测。 在线性分类器上使用冻结表征训练50轮。报告的测试准确率取此阶段验证准确率最高的检查点。

随机种子控制。 每个随机种子 σ 共同决定权重初始化、数据加载器打乱顺序以及随机数据增强的采样。所有超参数在不同随机种子和条件下均保持固定。

边距调度。 边际随层数线性递减：

$$m_\ell = m_0 + (m_{L-1} - m_0)\,\frac{\ell}{L-1}, \quad \ell \in \{0, \ldots, L{-}1\}. \tag{13}$$

析因设计。 在标准间隔调度下，我们交叉考察两个因素：（A）边距类型（截断与减法）和（B）稳定性模式（分离与直接），每个单元使用 n=7 个独立随机种子（共28次运行）。

剂量-反应探测。 为检验降低边距是否会减弱方差效应，我们采用低边距调度（m_0=0.2 $\to$ m_{L-1}=0.1），对截断分离和直接钳位分别使用7个随机种子（共14次截断运行）。减法对照组复用标准间隔析因实验中的7个减法分离随机种子；这是有效的，因为减法基线在梯度上是中性的（命题4.1），因此边距调度值不会影响减法条件下的梯度或训练所得模型。由此得到的合并比较不平衡（n=14 截断对比 n=7 减法）；根据析因实验中稳定性模式不影响方差的发现（$p > 0.82$；表4），我们节省计算资源。

计算。所有运行均在 NVIDIA V100-SXM2-32GB GPU 上进行。每次 CIFAR-10 运行约需 7 个 GPU 小时。完整的实验活动（CIFAR-10 析因实验、剂量-反应实验和泛化数据集）总计约 88 次运行和 ∼550GPU 小时。

5.2 结果：CIFAR-10

5.2.1 标准间隔析因结果

表 3 报告了标准边距调度下 2×2 析因中每个单元格的测试准确率统计。每个随机种子的准确率列于附录 A。

表4： **析因方差敏感性检验（标准边距）。**

Test	边距类型	稳定性模式	交互作用
Levene	0.061	0.824	0.058
Brown-Forsythe	0.058	0.856	0.137

表 5： **按边距类型（标准边距，** n=14 **每组）汇总的测试准确率。**

条件	n	均值（%）	标准差（%）	Var
截断（合并）	14	78.48	1.008	1.0170
减法（合并）	14	78.51	0.415	0.1724

析因方差检验。 为评估变异性是否取决于边距类型、稳定性模式或二者的交互作用，我们采用 Levene 和 Brown-Forsythe 检验程序（分别对偏离单元格均值和中位数的绝对偏差进行双因素方差分析）。表4 报告了检验结果。两项检验均未发现稳定性模式影响方差的证据（两者均为$p > 0.82$）。边距类型的影响处于临界水平（$p \approx 0.06$），但方向上与下文的合并比较一致。在 Brown-Forsythe 检验下，交互作用未获支持（$p = 0.137$）；Levene 检验中处于临界水平的交互作用（p =0。058）似乎是由直接钳位单元格较高的方差所驱动的。

合并后的主要比较。由于稳定性模式不会影响方差，因此主要分析合并了不同稳定性模式的数据，每种边距类型有 n=14 个随机种子（表5）。

均值比较。Welch 双样本 t 检验未发现边距类型之间存在均值差异的证据（t(19.4) $= 0.10, p = 0.92$; $\mu_{\texttt{clamp}} - \mu_{\texttt{subtract}}$ 的 95%置信区间： $[-0.64, 0.59]$）。因此，方差差异不能归因于集中趋势的变化。

方差比较。方差齐性的 F检验得出 $F(13, 13) = 5.90$, $p = 0.003$，拒绝 H_0（式 (2)）。此外，我们还报告：

- Levene检验： $p = 0.058$（边际显著，方向一致）。
- 方差比的自助法95%置信区间： $[1.62, 15.80]$ （对样本方差比进行10,000次百分位自助重抽样）。

Shapiro-Wilk检验未能拒绝任一合并组的正态性假设（截断组：$W = 0.951$、 $p = 0.58$ ； 减法组： $W = 0.952$、$p = 0.59$ ； 见附录D中的表23），支持F检验假设。
图1展示了每个随机种子的准确率分布。

5.2.2 剂量-反应探测

如果方差效应由饱和介导，降低起始间隔应会减弱该效应。表6比较了两种边距调度。

将起始间隔减半，从 $m_0 = 0.4$ 降至 $m_0 = 0.2$ ，使方差比从 $5.90\times$ 降至 $2.98\times$，这与剂量-反应关系一致。低边距组的方差比本身未达到显著性（$F(13, 6) = 2.98, p = 0.19$；方差比的自助法95%置信区间： $[0.79, 30.77]$），这可能反映了效应量较小且设计不平衡导致检验效能较低（n=14 对比 n=7）。因此，我们将此探测结果视为支持饱和介导机制的方向性证据，而非独立的显著性结论。
对低边距合并均值进行 Welch t 检验，结果为 $t(18.4) = -1.59, p = 0.13$ （$\mu_{\texttt{clamp}} - \mu_{\texttt{subtract}}$ 的 95%置信区间: $[-1.03, +0.14]$ 个百分点），表明均值差异不显著，不过与标准边距下的结果（$p = 0.92$）相比，这一趋势的零效应特征不那么明显。
该探测存在两个局限。首先，由于 τ 固定不变，降低 m_0 会同时减小有效 logit 偏移m_0/τ，从而使饱和频率的变化（公式（11））与整体 logit 尺度的变化相混淆。其次，低边距方差比的自助法置信区间包含 1.0，因此，在这一缩减后的安排下，数据同样与不存在方差差异的情况相一致。

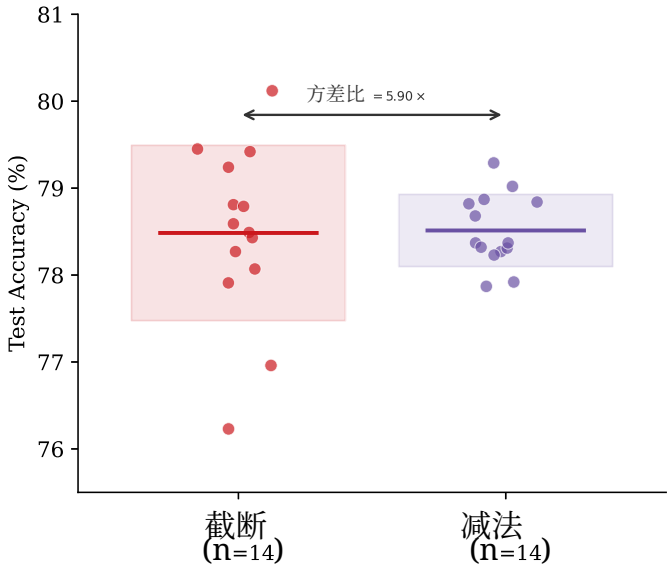


图1：**按边距类型（标准边距，汇总）划分的各随机种子测试准确率**。每个点代表一个随机种子（每组 n=14 个）。横线表示组均值；阴影带表示 ±1 标准差。

表6：**剂量-反应探测**。 将边距调度从 $0.4 \to 0.1$ 降低至 $0.2 \to 0.1$ 会降低观测到的方差比。低边距比较并不平衡（n=14 截断与 n=7 减法；见第5.1节）。

调度	方差截断	减法方差	VR	F检验 p
$0.4 \to 0.1$(标准)	1.0170	0.1724	5.90×	0.003
$0.2 \to 0.1$(低)	0.6498	0.2178	2.98×	0.19

6 机制分析

本节提供诊断证据，以说明为何截断变体在 CIFAR-10 上表现出更高的种子间方差。我们强调，这些诊断结果证实了所提出机制的必要条件，但并不构成因果证明。

6.1 梯度截断欠饱和

$f(x) = \min(x,\ 1)$ 的导数为 $f'(x) = 1$，当 $x < 1$ 时；当 $x > 1$ 时，$f'(x) = 0$。根据式 (8)，对于层 ℓ 上任意一对正值 $(u,\ v)$，若 $s_{uv,\ \ell} + m_\ell > 1$，则修改后的相似度 $\tilde{s}_{uv,\ \ell} = 1$ 是关于 $s_{uv,\ \ell}$ 的常数，因此该配对对锚点 u 的损失项所产生的直接（分子）梯度贡献为零。

分母路径。沿直接路径进行梯度截断，并不意味着完整导数 $\partial \mathcal{L}_\ell / \partial s_{uv,\ \ell}$ 为零。钳制后的相似度 $\tilde{s}_{uv,\ \ell} = 1$ 仍会通过公式 (6) 进入其他 锚点的 Softmax 分母 $u' \neq u$。因此，梯度可以沿这条间接（分母）路径传播。饱和的净效应是部分而非完全的梯度截断；这种效应取决于数据，并随每个小批量中恰好超过饱和阈值的配对而变化。

6.2 按层统计截断激活率

表 7 报告了两种边距调度下全部 8 层的 CAR（公式 (11)），该值是在第一阶段的最后一轮（第600轮）通过对小批量和随机种子取均值计算得出的。

更高的起始间隔会在早期层产生显著更高的CAR，差距单调缩小，并在第7层消失，两种调度在该层共享 $m = 0.1$。在低间隔调度下，尽管第1层的间隔更低，CAR仍从第0层（40.3%）略微上升至第1层（43.7%）；这可能反映了第1层更紧密的类内余弦相似度（即更聚集的表征），不过我们并未进一步探究这一点。

表 7：**各层钳制激活率（CAR）（CIFAR-10）**。

层	边际（标准差 / 低）	CAR（标准差）	CAR（低）	归约
0	0.40 / 0.20	60.7%	40.3%	33.7%
1	0.36 / 0.19	58.3%	43.7%	24.9%
2	0.31 / 0.17	54.0%	42.5%	21.3%
3	0.27 / 0.16	49.6%	40.5%	18.3%
4	0.23 / 0.14	45.8%	38.6%	15.6%
5	0.19 / 0.13	41.6%	36.6%	12.0%
6	0.14 / 0.11	37.5%	34.5%	7.8%
7	0.10 / 0.10	32.3%	32.3%	0.2%

表 8：**各层的均值梯度范数（最终轮次，CIFAR-10）**。数值为均值 ± 随机种子间的标准差。

层	截断	减法	比率（减法 / 钳位）
L0	0.400 ± 0.028	1.614 ± 0.045	$4.0\times$
L1	0.115 ± 0.008	0.243 ± 0.008	$2.1\times$
L2	0.048 ± 0.002	0.049 ± 0.008	$1.0\times$
L3	0.023 ± 0.003	0.018 ± 0.007	$0.8\times$

6.3 各层的梯度范数

表 8报告了第0–3层在最终训练轮次的梯度 ℓ_2 范数（式(12)）。我们将分析限制在这些层，因为在标准调度下，到第3层时CAR已降至50%以下，且第4–7层在各条件之间的差异可忽略不计（梯度范数的相对差异为$<5\%$；未列入表中）。

在 CAR 最高（60.7%）的第0层，钳位使梯度范数相较于减法基线降低了 $4.0\times$ 倍。到第1层时，降幅缩小至 $2.1\times$，并在第2层消失。各层 CAR 与梯度范数降幅之间的这种对应关系，符合以下解释：饱和诱发的截断是早期层梯度差异的主要原因。

图 2 将两个剖面绘制在一起。

6.4 机制解释

诊断证据符合以下定性解释。在标准调度下，第0层超过 60% 的正样本对超出钳位阈值（表 7）；降低起始间隔会使这一比例减半（表 7，“归约”列）。这种频繁饱和会截断早期层的梯度：钳位时第0层的梯度范数比减法基线低 $4.0\times$（表 8），且降幅随各层的 CAR 剖面变化（图 2）。

截断并非均匀的：哪些正样本对发生饱和，取决于实际实现的余弦相似度，而这些相似度随初始化和数据排序而变化。因此，不同随机种子在各小批量上会经历不同的梯度置零模式，使优化轨迹的分离程度大于非截断基线。 $5.90\times$方差比（表5）是总体结果；将间隔减半会使方差比降至 $2.98\times$（表6），这与饱和减少从而分离减少相一致。

这条链条中的前两个环节（频繁饱和与梯度范数降低）已由我们的测量直接验证。后续环节——从随机截断到轨迹分离，以及从分离到准确率方差——是根据证据的联合模式推断得出的，而非经因果证明。具体而言，我们并未在训练过程中跟踪CAR或梯度范数（仅在第600轮），也未测量CAR的逐种子变异，而后者能更直接地检验截断率更高的随机种子是否会产生更极端的最终准确率。

7 跨数据集泛化

CIFAR-10 的结果表明，当 CAR 较高时，钳位会使方差增大。为了探究这种效应出现的条件，我们在另外三个数据集上重复了方差审计；这些数据集在类别数量、正样本对密度、任务难度和视觉域方面各不相同。表 9 总结了主要结果；下面将分别讨论各数据集。

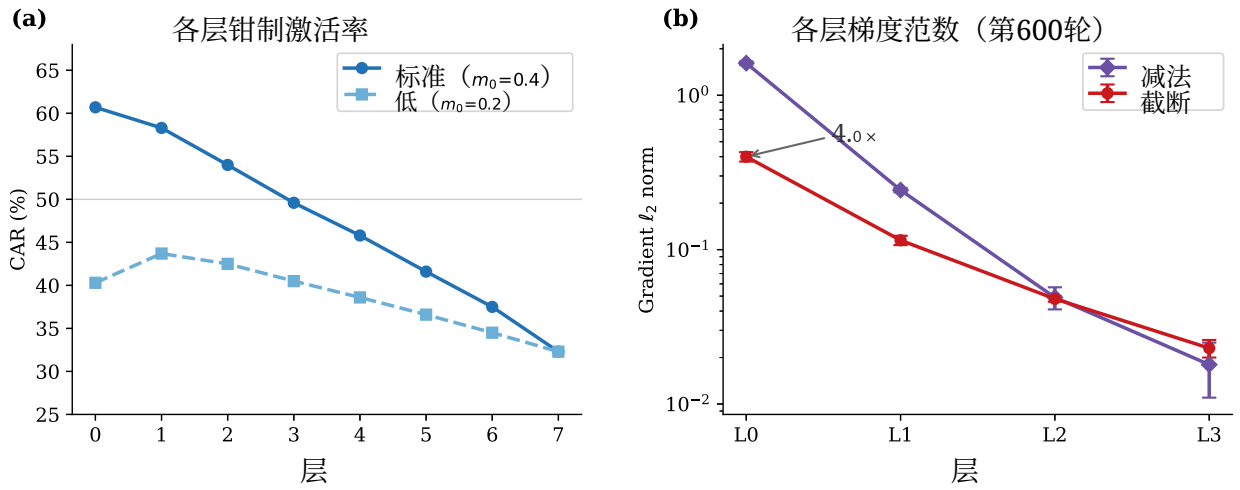


图 2：**按层划分的诊断剖面（CIFAR-10）**。（a）标准边际和低边际调度下的 CAR。（b）最终轮次的梯度ℓ_2 范数。条件之间的梯度范数差距与 CAR 剖面一致，并在第 2 层趋于消失。

表 9：**跨数据集摘要**。 方差比（VR）、第0层CAR、每个小批量中的近似正样本对数量（批大小为 512）以及平均钳位准确率。CIFAR-10 是唯一呈现 VR >1 的数据集。

数据集	K	正样本对	L0 CAR	准确率（%）	VR
CIFAR-10	10	∼25,700	60.7%	78.5	5.90×
CIFAR-100	100	∼2,580	29.0%	51.8	0.39×
SVHN	10	∼30,500	51.0%	96.7	0.25×
Fash-MNIST	10	∼25,700	—	92.6	0.08×

7.1 CIFAR-100：低正样本对密度

在100个类别、批大小为512的情况下，每个小批量大约包含2,580个正样本对，约比CIFAR-10的 ∼25,700少 10×。我们进行了 2×2 析因实验（每个单元格n=5 个随机种子；架构和训练与CIFAR-10相同，但 K=100除外）。

方差比为 0.39×，与 CIFAR-10 中的情况相反，且不显著（$F(9,\ 9) = 0.39, p = 0.17$）。CIFAR-100 上的 CAR 测量结果揭示了原因：L0 CAR 仅为 29.0%（表 11），而 CIFAR-10 上为 60.7%；各层的平均 CAR 为 19.7%，而 CIFAR-10 上为 47.5%。由于每个批次中的正样本对较少，较少的相似度值会接近饱和边界，因此，在 CIFAR-10 上导致方差膨胀的梯度截断路径基本处于非激活状态。
逐层梯度范数证实，低 CAR 对应着梯度截断的缺失（表 12）。在第0层，钳位与减法的范数比为 0.58×：钳位产生的梯度更小，而非更大。这与 CIFAR-10 上的情况相反，在那里，同一比值为 0.25× （钳位范数 4.0× 较低，意味着减法范数占主导）。在 CIFAR-100 上，两种条件产生的梯度景观相似，这与不存在方差效应的情况一致。

7.2 SVHN：高准确率，高CAR

SVHN 是一个类别频率不均匀的 10 类数字识别数据集，每个小批量产生 ∼30,500 个正样本对。我们对直接钳位和直接减法各运行了 n=5 个随机种子。 – –
方差比为 0.25×（反转； $F(4,4) = 0.25, p = 0.21$）。SVHN 的 CAR剖面在性质上不同于 CIFAR-10 的：CAR增加 随深度（L0：51.0%，L7：78.3%），而 CIFAR-10 的则单调下降（L0：60.7%，L7：32.3%）。平均 CAR 为 70.6%，远高于 CIFAR-10 的 47.5%。然而方差比是反转的。
CAR 水平与方差膨胀之间的这种分离表明，高 CAR并不足以产生方差效应。我们推测，SVHN 的高准确率（∼97%）压缩了种子间差异的空间：当任务几乎解决时，无论是否进行梯度截断，所有随机种子都会收敛到相似的最优点。

表 10：**CIFAR-100 汇总结果（每组 n=10 ）**。

条件	n	均值（%）	标准差（%）	Var
截断（池化）	10	51.82	0.505	0.255
减法（池化）	10	51.38	0.814	0.663

表 11：**CAR 对比：CIFAR-10 与 CIFAR-100（标准边距）**。

层	CAR（CIFAR-10）	CAR（CIFAR-100）	差异
0	60.7%	29.0%	−31.7 pp
1	58.3%	27.8%	−30.5 个百分点
2	54.0%	24.2%	−29.8 个百分点
3	49.6%	21.4%	−28.2 个百分点
4	45.8%	18.4%	−27.4 个百分点
5	41.6%	15.5%	−26.1 个百分点
6	37.5%	12.3%	−25.2 个百分点
7	32.3%	8.7%	−23.6 个百分点
Mean	**47.5%**	**19.7%**	−27.8 个百分点

7.3 Fashion-MNIST：高准确率，中等难度

Fashion- MNIST 是一个 10 类灰度数据集。我们分别对截断和减法条件运行了 n=5 个随机种子 ons. 方差比为 $0.08\times$，这是所有数据集中最强的反转，并且具有显著性（$F(4,\ 4) = 0.08, p = 0.029$）。Fashion-MNIST 的高准确率（∼92%）使其处于与 SVHN 相同的状态：任务难度足够低，因此尽管存在梯度截断，不同随机种子仍会收敛到相似的最优点。Fashion-MNIST 未测量 CAR；根据 SVHN 的模式（高准确率与较高 CAR 同时出现，这是由于表征聚类良好），我们预计其 CAR 为中等至较高，但不会造成方差膨胀。

7.4 SVHN 难度扫描

为了直接检验任务难度是否会调节钳位与方差之间的关系，我们在 SVHN 上重复了实验，逐步加大数据增强强度以降低准确率。在保持数据集、架构和类别数量不变的同时，只改变优化难度。

扫描结果揭示了一个明显的转变：在相同数据集和相同架构上，增强难度的提高会使方差比从 $0.25\times$变为 $16.73\times$。在强增强条件下，截断条件表现出双峰行为：部分随机种子收敛（准确率为∼35%），而其他随机种子则失败（准确率为∼18%），从而导致极端方差。这是最有力的证据，表明决定钳位是否会使训练不稳定的是任务难度，而非 CAR 本身。

7.5 综合：钳位何时会影响方差?

The 跨数据集证据表明，当以下两个条件同时出现时，钳位会增大方差 :

1. **较高的 L0 钳制激活率**。 CIFAR-100 的 L0 CAR 较低（29.0%），因此截断路径不会激活。较高的 CAR 是必要条件，但并非充分条件。

2. **中等任务难度**。 SVHN 和 Fashion-MNIST 的准确率达到 $>$ 92%，留给轨迹拉开差距的空间很小。另一方面，强增强下的准确率极低，会导致极端方差（表 15）。

CIFAR-10 是我们研究中唯一同时满足以下两个条件的数据集：较高的 L0 CAR（60.7%）和中等准确率（∼78%）。SVHN 难度扫描证实，这些条件会在单个数据集上相互作用：同一架构在准确率为 97% 时产生的方差比为 $=0.25\times$，而在准确率为 25% 时产生的方差比为 $=16.73\times$ 。

表 12：**各层的平均梯度范数（CIFAR-100，单个随机种子）**。与 CIFAR-10（表 8）对比，其中钳位将 L0 范数降低了 $4.0\times$。

层	边际	截断	减法	比率（C/S）
L0	0.400	1.265 ± 0.099	2.170 ± 0.279	$0.58\times$
L1	0.357	0.262 ± 0.019	0.358 ± 0.032	$0.73\times$
L2	0.314	0.080 ± 0.004	0.073 ± 0.004	$1.09\times$
L3	0.271	0.046 ± 0.002	0.029 ± 0.002	$1.60\times$

表13：**SVHN结果（每组n=5）**。

条件	n	均值（%）	标准差（%）	Var
截断	5	96.70	0.430	0.185
减法	5	95.29	0.853	0.727

8 讨论

没有平均代价的方差。在 CIFAR-10 上，截断变体产生了 $5.90\times$更高的测试准确率方差，但平均准确率没有可检测的差异（$p = 0.92$；表 5）。这意味着，在此设置下，切换到减法参考可以降低噪声，同时不会对中心估计产生可观测的代价。

随机种子预算的实际影响。 样本均值的标准误按 Std$/\sqrt{n}$ 缩放。具体来说：在 CIFAR-10 上使用钳位（Std ≈ 1.0 pp）时，若要达到 ± 0.3 pp 的标准误，需要 $n \approx 11$ 个随机种子；使用减法（Std ≈ 0.42 pp）时，$n \approx 2$ 个就足够。这种差异可能决定在有限计算资源下，超参数比较是否具有足够的统计效力。

层局部训练为何可能放大这种影响。 在端到端反向传播中，某一层的梯度截断会受到整条下游梯度链的调节，而来自其他层的补偿信号可以部分抵消局部伪影。在 CFF 中，每一层都独立优化，不存在层间梯度流，因此第0层由饱和诱发的截断无法通过更深层获得补偿。这种结构差异可能使 CFF 比端到端训练更容易受到边界诱发的梯度截断影响。

数据集依赖性与正样本对密度的作用。CIFAR-100 上反转的方差比（VR $=0.39\times$）可归因于其较低的正样本对密度（100 个类别，而非 10 个），因此每一层的 CAR 也相应较低（表 11）。这证实，梯度截断机制需要足够高的正样本对密度，才能引发广泛的饱和。逐层梯度范数也支持这一观点：在 CIFAR-100 上，第0层的截断与减法梯度范数比为 $0.58\times$（1.27 对 2.17），这表明当 CAR 较低时，钳位并不会放大梯度幅值。在 CIFAR-10 上，相应的比值在钳位条件下更高（表 8）。研究多类别问题的从业者不太可能遇到 CIFAR-10 上记录的方差膨胀现象。

任务难度作为调节因素。SVHN 和 Fashion-MNIST 表明，高 CAR 并不足以导致方差膨胀：两者的准确率均达到 $> 92\%$，无论截断程度如何，种子间差异都有限。SVHN 难度扫描（表 15 ）为这种交互作用提供了最直接的证据：在同一数据集上，准确率从 97% 降至 25%，方差比便从 $0.25\times$ 变为 $16.73\times$。因此，钳位的方差效应在中等难度区间最为显著：此时优化轨迹对梯度扰动较为敏感，但任务又没有简单到使所有随机种子无论如何都会收敛。

与度量学习中的间隔饱和的关联。人脸识别目标中的硬间隔（SphereFace、CosFace、ArcFace）旨在产生饱和logit区域。我们的发现表明，在评估这些场景中的间隔实现时，方差审计或许能提供有用信息，而不应只看平均准确率比较；不过，端到端梯度流和不同的损失归约方式限制了这一类比的适用范围。

实践指南。根据跨数据集证据：(i) 在类似 CIFAR-10 的设置中（准确率适中、正样本对密度高），切换到减法参考可大幅降低方差，且未观察到平均代价。(ii) 在高准确率设置（$> 90\%$）或多类别设置（正样本对密度低）中，钳位似乎不会增大方差，

表 14：**Fashion-MNIST 结果（每组 n=5）**。

条件	n	均值（%）	标准差（%）	Var
截断	5	92.57	0.113	0.0127
减法	5	91.87	0.406	0.1645

表 15：**SVHN 难度扫描**。在同一数据集上增加增强难度，会使方差比从 $0.25\times$（简单，准确率为 $\sim$97%）变为 $16.73\times$（困难，准确率为 $\sim$25%）。

难度	n	准确率（%）	方差截断	减法方差	VR	模式
简单（标准）	5	96.7	0.185	0.727	$0.25\times$	反转
中等	5	26.8	37.57	17.23	$2.18\times$	升高
Hard	5	25.0	85.45	5.11	$16.73\times$	极端

因此可能无需修改。(iii) 如有疑问，测量 L0 CAR 可提供一种简单的诊断方法：远低于 50% 的数值表明截断路径基本未被激活。

9 局限性与有效性威胁

单一架构和训练协议。 所有结果均采用固定的 ViT 配置（d=128, H=4, L=8）、训练计划（600 个第一阶段训练轮数）和数据增强流程。我们没有测试方差效应如何随模型规模、深度或其他训练计划而变化。

跨数据集设计不平衡。 CIFAR-10 的主要分析采用完整的 2×2 因子设计（每个单元 n=7，合并后 n=14）。CIFAR-100 每个单元使用 n=5（合并后 n=10）。SVHN 和 Fashion-MNIST 每个条件使用 n=5，不包含稳定性模式这一因素。这些设计差异反映了计算资源分配优先考虑主要数据集；相应地，泛化实验的统计功效较低。

SVHN 设计的局限性。 SVHN 实验仅使用直接稳定性模式，因为我们事后才发现分离条件中存在一个错误。CIFAR-10 因子实验表明，稳定性模式不会影响方差（$p > 0.82$），因此这一局限性不太可能改变结论，但我们仍指出这种不对称性。

缺少 Fashion-MNIST 的 CAR。 由于检查点可用性限制，我们未测量 Fashion-MNIST 的 CAR。在准确率为 92.6% 时观察到显著的方差反转（方差比 $= 0.08\times$, $p = 0.029$），这与 SVHN 上观察到的任务难度调节模式一致；但由于缺少 CAR 数据，我们无法确认截断路径在该数据集上是否处于激活状态。

温度与边际混杂。 截断（公式 (8)）在温度缩放（公式 (4)）之前应用，因此在达到饱和之前，其在 logit 空间中的有效强度为 m_ℓ/τ。在剂量-反应探测中，我们改变 m_0，同时固定 $\tau = 0.15$，这使饱和频率的变化与整体 logit 尺度的变化相混杂。

SVHN 难度扫描的混杂因素。 难度扫描改变了增强强度，这会同时影响任务难度、成对相似度的分布以及有效的正样本对信息量。我们无法区分究竟是哪一个因素驱动了方差比转变。

对数据增强的敏感性。我们的数据增强流程对 32×32 图像进行随机裁剪时使用 12 像素的填充，比标准的 4 像素填充更激进。由于增强强度会影响成对余弦相似度的分布，进而影响钳位饱和的频率，因此在较温和的数据增强下，方差效应的大小可能有所不同。实验内部的比较仍然有效（所有条件使用相同的数据增强），但不应假定绝对 CAR值和方差比能够适用于其他数据增强流程。

两阶段流程。 报告的测试准确率同时反映表征学习（阶段 1）和对冻结表征进行的线性探测（阶段 2）。我们的机制分析完全聚焦于阶段 1（CAR、梯度范数），但探测阶段方差也可能对观察到的不同随机种子间的差异有所贡献。我们没有区分这两个阶段对总方差的贡献。

诊断时点。 CAR 和梯度范数仅在第一阶段轮次的最后一轮报告。我们没有追踪这些量在训练过程中的变化。

逐种子的机制关联。 CAR 和梯度范数以跨随机种子均值的形式报告。我们没有测量这些诊断指标的逐种子变异；测量这些变异可以更直接地检验经历更频繁截断的随机种子是否会产生更极端的最终准确率。

统计功效。 标准间隔析因实验中每个单元使用 $n = 7$ 个随机种子（每种边距类型合并后共 $n = 14$ 个）。尽管 F 检验在 $p = 0.003$ 水平上拒绝方差齐性假设，但 VR 的自助法置信区间较宽（$[1.62, 15.80]$），因此该效应的大小比其方向更难以精确确定。泛化数据集的样本量更小（$n = 5$），进一步限制了对单个数据集进行显著性检验的统计功效。

归约范围。 命题4.1证明，在本文全程采用的正样本均值归约下，梯度具有中性。若采用其他归约方式（例如对正样本求和），则梯度中性这一结果不再成立。

10 结论

我们给出了对比式前向-前向训练中使用的监督对比损失的形式化规范，并证明了一种常见的间隔变体——对数概率后减法——在按正样本均值归约时是梯度中性的（命题 4.1），因此可将其确立为真正的无间隔参照。以此为参照，我们分离出另一种实现方式的影响：通过 min($s + m$, 1) 进行饱和相似度钳制。

在 CIFAR-10 上（2×2 因子设计，每个单元有 n=7 个随机种子），钳位与汇总测试准确率方差近乎六倍的升高相关（F-检验 $p = 0.003$），而平均准确率在统计上无法区分。三项诊断分析支持由饱和介导的解释：早期层的钳制激活率超过 60%，钳位时第 0 层的梯度范数低 $4.0\times$，且将起始间隔减半后，方差比从 $5.90\times$ 降至 $2.98\times$。

这一效应并不能一致地泛化。在 CIFAR-100、SVHN 和 Fashion-MNIST 上的复现表明，这三种情况下的方差比均发生反转。跨数据集分析识别出两个调节因素：正样本对密度（决定 CAR）和任务难度（在高准确率下限制轨迹分离）。SVHN 难度扫描在单一数据集上证实了这一交互作用，方差比从 $0.25\times$（97% 准确率）变为 $16.73\times$（25% 准确率）。

实际结论是有条件的：在以 CIFAR-10 为代表的情形中（中等准确率，每批次包含大量同类样本对），改用梯度中性减法参考可以在未观察到代价的情况下降低训练方差。在高准确率或多类别设置中，该效应不存在或发生反转，钳位可能无需修改。我们建议 CFF 实践者将 L0 CAR 作为一种简单诊断进行测量，并明确报告所选的间隔实现方式，以支持可复现性。

更广泛影响声明

本研究审查了对比损失中边距实现选择所导致的种子间方差。本研究未引入新的模型能力、数据集或面向部署的系统，我们也未发现任何可能造成负面社会影响的具体途径。

参考文献

Hossein Aghagolzadeh 和 Mehdi Ezoji。对比式前向-前向：一种视觉 Transformer 的训练算法。神经网络，192:107867，2025年。doi: 10.1016/j.neunet.2025.107867。

Eugene Belilovsky、Michael Eickenberg 和 Edouard Oyallon。贪心逐层学习可扩展至 ImageNet。载于第*36*届国际机器学习会议（*ICML*）论文集，第583–593页，2019年。

Srinadh Bhojanapalli、Ayan Chakrabarti、Daniel Glasner、Daliang Li、Thomas Unterthiner 和 Andreas Veit。理解 Transformer 在图像分类中的鲁棒性。载于*IEEE/CVF*计算机视觉国际会议（*ICCV*）论文集，第10231–10241页，2021年。

Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。用于视觉表征对比学习的简易框架。载于 第 *37* 届国际机器学习会议（*ICML*）论文集，第 1597–1607 页，2020年。

Xing Chen、Dongshu Liu、J´er´emie Laydevant 和 Julie Grollier。自对比前向-前向算法。《自然·通讯》，16:5978，2025年。doi：10.1038/s41467-025-61037-0。

Alexander D'Amour、Katherine Heller、Dan Moldovan、Ben Adlam、Babak Alipanahi、Alex Beutel、Christina Chen、Jonathan Deaton、Jacob Eisenstein、Matthew D Hoffman 等。规格不足给现代机器学习中的可信度带来了挑战。《机器学习研究杂志》，23(226):1–61，2022年。

Jiankang Deng、Jia Guo、Niannan Xue 和 Stefanos Zafeiriou。ArcFace：用于深度人脸识别的加性角度边界损失。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第4690–4699页，2019年。

Jesse Dodge、Gabriel Ilharco、Roy Schwartz、Ali Farhadi、Hannaneh Hajishirzi 和 Noah Smith。微调预训练语言模型：权重初始化、数据顺序与提前停止，2020年。

Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。《一张图像胜过 16x16 个词：大规模图像识别的 Transformer》。发表于 国际学习表征会议（*ICLR*），2021年。

Florian Graf、Christoph Hofer、Marc Niethammer 和 Roland Kwitt。《剖析有监督对比学习》。发表于第 *38* 届国际机器学习会议（*ICML*）论文集，第 3821–3830 页，2021年。

Kaiming He、Haoqi Fan、Yuxin Wu、Saining Xie 和 Ross Girshick。《用于无监督视觉表征学习的动量对比》。发表于 *IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第 9729–9738 页，2020年。

Geoffrey Hinton。前向-前向算法：一些初步研究，2022年。

Prannay Khosla、Piotr Teterwak、Chen Wang、Aaron Sarna、Yonglong Tian、Phillip Isola、Aaron Maschinot、Ce Liu 和 Dilip Krishnan。有监督对比学习。载于神经信息处理系统进展，第 33 卷，第 18661–18673 页，2020年。

Weiyang Liu、Yandong Wen、Zhiding Yu、Ming Li、Bhiksha Raj 和 Le Song。SphereFace：用于人脸识别的深度超球面嵌入。载于*IEEE* 计算机视觉与模式识别会议论文集（*CVPR*），第 6738–6746 页，2017年。

Sindy L¨owe、Peter O'Connor 和 Bastiaan Veeling。终结端到端：表示的梯度隔离学习。载于神经信息处理系统进展，第 32 卷，2019年。

Arild Nøkland 和 Lars Hiller Eidnes。使用局部误差信号训练神经网络。载于第*36*届国际机器学习会议（*ICML*）论文集，第4839–4850页，2019年。

Razvan Pascanu、Tomas Mikolov 和 Yoshua Bengio。论循环神经网络训练的困难。载于第*30*届国际机器学习会议（*ICML*）论文集，第1310–1318页，2013年。

David Picard。Torch.manual seed(3407) 就是你所需要的一切：论随机种子对计算机视觉深度学习架构的影响，2021年。

A¨aron van den Oord、Yazhe Li 和 Oriol Vinyals。基于对比预测编码的表征学习。载于神经信息处理系统进展，第31卷，2018年。

Hao Wang、Yitong Wang、Zheng Zhou、Xing Ji、Dihong Gong、Jingchao Zhou、Zhifeng Li 和 Wei Liu。CosFace：用于深度人脸识别的大间隔余弦损失。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*）论文集，第5265–5274页，2018年。

Jingzhao Zhang、Tianxing He、Suvrit Sra 和 Ali Jadbabaie。《为何梯度裁剪能加速训练：自适应性的理论依据》。载于国际学习表征会议（*ICLR*），2020年。

A 逐随机种子结果

表 16 和表 17 报告了所有 CIFAR-10 运行的单次测试准确率。在每个条件下，随机种子均一致编号为 S1–S7；不同条件下相同的随机种子索引对应相同的随机种子值。减法分离行在两张表中完全相同，因为减法基线是梯度中性的（命题 4.1）：边距调度值不会影响训练后的模型。

表16： **CIFAR-10 各随机种子结果（标准边距** $0.4 \rightarrow 0.1$**）**。

条件	S1	S2	S3	S4	S5	S6	S7
截断分离	78.49	78.27	78.43	76.96	78.81	78.59	80.12
直接截断	78.07	77.91	79.42	76.23	79.24	78.79	79.45
减法分离 -	78.82	79.29	78.87	79.02	77.87	78.37	78.84
直接减法 -	78.27	78.31	78.68	78.23	78.37	78.32	77.92

表17：**CIFAR-10 各随机种子结果（低裕度** $0.2 \rightarrow 0.1$**）**。减法分离行与表16相同；详见正文。

条件	S1	S2	S3	S4	S5	S6	S7
钳位分离	78.97	78.49	79.04	78.15	78.46	78.49	77.99
直接钳位	80.16	77.36	78.33	77.74	76.88	77.55	78.35
减法分离 -	78.82	79.29	78.87	79.02	77.87	78.37	78.84

B SVHN CAR剖面

表18：**CAR比较：CIFAR-10与SVHN（标准边距）**。SVHN的CAR随深度增加，与CIFAR-10相反。

层	CAR（CIFAR-10）	CAR（SVHN）	差异
0	60.7%	51.0%	-9.7 pp
1	58.3%	65.3%	$+7.0$ 个百分点
2	54.0%	75.0%	$+21.0$ 个百分点
3	49.6%	78.3%	$+28.7$ pp
均值（L0–L7）	**47.5%**	**70.6%**	$+23.1$ pp

C 每个随机种子的结果：泛化数据集

表 19–22 报告了所有跨数据集实验的单次测试准确率。在每个条件下，随机种子均一致编号为 S1–S5。

表 19： **CIFAR-100 各随机种子结果（标准边距** $0.4 \rightarrow 0.1$**,** n=5 **每个单元格）**。

条件	S1	S2	S3	S4	S5
钳位分离	51.98	52.06	52.32	51.62	52.36
直接钳位	50.84	51.60	51.62	52.43	51.36
减法分离 -	50.85	52.25	52.31	50.17	50.90
直接减法 -	50.78	52.52	51.98	50.81	51.23

表 20：**SVHN 各随机种子的结果（每组** n=5 **个，仅限**直接**模式）**。

条件	S1	S2	S3	S4	S5
截断	97.16	96.42	97.11	96.17	96.66
减法	95.09	96.11	95.36	93.95	95.92

表 21： **Fashion-MNIST 的逐随机种子结果（每组 n=5 ）。**

条件	S1	S2	S3	S4	S5
截断	92.39	92.69	92.63	92.55	92.57
减法	92.00	91.32	92.06	92.36	91.62

表22：**各随机种子的SVHN 难度扫描结果（n=5 每个条件）**。数据增强细节见第5.1节。

难度	条件	S1	S2	S3	S4	S5
中等	截断	28.71	21.05	36.37	23.89	23.92
	减法	22.01	30.19	28.50	21.22	28.02
Hard	截断	18.11	35.75	19.02	34.41	17.62
	减法	23.67	25.93	21.98	21.65	19.97

D 正态性检验

表 23 报告了用于主要方差比较的合并组的 Shapiro–Wilk检验。

表23： **合并组的 Shapiro–Wilk正态性检验。**

组	n	W	p 值
截断（标准边距）	14	0.9513	0.5816
减法（标准边距）	14	0.9516	0.5853
截断（低裕度）	14	0.9608	0.7355
减法（低裕度）	7	0.9074	0.3780

E 可复现性详情

软件。 所有实验均在 NVIDIA V100-SXM2-32GB GPU 上使用 PyTorch 2.6.0 与 CUDA 12.4（系统驱动 CUDA 12.8）。CFF实现基于 Aghagolzadeh和Ezoji [2025] 的代码库，并做了修改以支持 subtractionmargin变体和诊断日志记录。

随机种子值。 CIFAR-10 实验使用七个随机种子：1、2、3、4、5、6、7。跨数据集实验（CIFAR-100、SVHN、Fashion-MNIST）使用随机种子 1–5。每个随机种子都会传入 torch.manual seed、numpy.random.seed、random.seed 以及 torch.cuda.manual seed all，以确保完全确定性。

自助法程序。 所有自助法置信区间均使用 10,000 次百分位重抽样。对于方差比，每次重抽样分别从各组独立有放回地抽取 n 个值，计算每个重抽样组的样本方差，并取其比值。所报告的95%置信区间为所得分布的第 2.5 与第 97.5 百分位数。

诊断计算。 CAR值（公式 (11)）在最后一个阶段 1 轮次（第600轮）计算：对训练集中的每个小批量求值公式 (11)，并对各小批量的结果取平均，然后再对各随机种子的结果取平均。梯度范数（公式 (12)）在最后一轮的单个小批量上计算，并对各随机种子的结果取平均。