arXiv:2304.04228v1 [cs.CV] 9 Apr 2023

深度图像检索中的无监督多准则对抗检测

Yanru Xiao 老道明大学 美国弗吉尼亚州诺福克yxiao002@odu.edu

Cong Wang*浙江大学 中国杭州
c
wang85@zju.edu.cn

Xing Gao 特拉华大学 美国特拉华州纽瓦克xgao@udel.edu

摘要

深度学习算法供应链中的脆弱性给下游图像检索系统带来了新的挑战。在各种技术中，深度哈希正日益普及。由于继承了深度学习的算法后端，近期已有不少攻击被提出，旨在扰乱正常的图像检索。不幸的是，*Softmax* 分类中的防御策略无法直接应用于图像检索领域。本文提出了一种高效、无监督的方案，用于识别汉明空间中的独特对抗行为。具体而言，我们从汉明距离、量化损失和去噪三个角度设计了三项准则，以防御非定向攻击和定向攻击，并共同限制对抗空间。在四个数据集上的大量实验表明，$2-23\%$ 检测率得到提升，同时计算开销极低，可满足实时图像查询的需求。

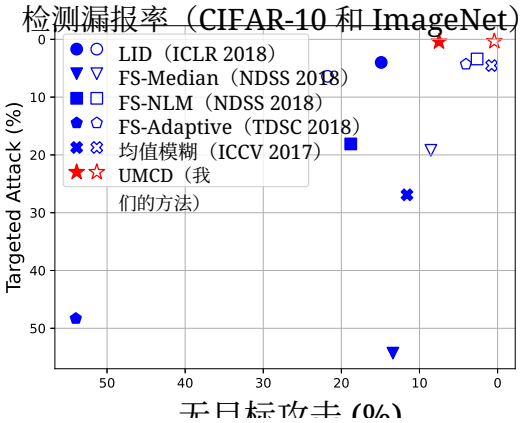


图1.深度哈希中针对非定向/定向对抗样本的不同检测方法的漏检率。实心和空心标记分别表示 CIFAR-10 和ImageNet。所提出的 UMCD 在两个数据集上都具有最低的漏检率。

1. 引言

在神经网络的支持下，深度哈希可实现大规模图像检索 [7, 8, 26, 27, 47, 49]。通过使用紧凑的二进制码表示高维图像，检索便可高效地计算汉明距离相似度。谷歌 [3], 必应 [2], Pinterest [4], 淘宝 [1] 均已将图像查询纳入其产品。尽管深度哈希取得了巨大成功，但它也继承了神经网络的漏洞 [36] ，并带来了新的攻击向量和影响。通过对查询或数据库图像施加对抗扰动，正常请求可能会被引导至无关结果（非定向攻击） [46]或特定类别（定向攻击） [6,40,42], 例如，将“哈士奇犬”查询引导至检索某品牌的“狗粮”，从而让攻击者免费为其产品打广告。

在攻击方面仅有少数几项努力 [6,40,42,46], 深度哈希在汉明空间中抵御对抗样本的能力仍然不足。对抗训练和检测是 Softmax 分类中两种常见的防御方法。然而，对抗训练必须应对鲁棒性与准确率之间不容忽视的权衡 [48]。根据我们的实现（见附录），在哈希函数下寻找极大极小鞍点变得更加困难，并且会导致准确率大幅下降。另一方面，检测旨在从原始像素 [15, 17], 特征分布 [17, 23, 28], Softmax 概率 [19] 和频率分量 [39] 等不同层面揭示对抗行为，并以有监督 [11] 或无监督的方式进行 [45]。基于对攻击方法的先验知识，有监督检测会训练分类器来区分对抗图像，但难以泛化到未知攻击。为此，本文探索无监督异常检测方向。与针对封闭类别概率集合进行的 S oftmax 分类不同，深度哈希会将相似/不相似图像映射为开放汉明空间中的二进制码。因此，我们的工作重点是挖掘其中独特的对抗行为

*通讯邮箱：cwang85@zju.edu.cn。

利用深度哈希检测非定向攻击和定向攻击。

从非定向攻击开始 [46], 我们首先从理论上推导出对抗性图像与其他类别之间的汉明距离分布，该分布渐近地接近高斯分布。对于定向攻击，我们发现了量化损失上的一种有趣的对抗性行为：当对抗性目标是生成与目标类别相同的哈希码时 [6, 40], 它会无意中使量化损失接近于零。因此，我们首先设计了两种阈值方法，分别以汉明距离和量化损失作为代理指标。然后，我们将这两个标准与基于去噪的检测相结合，以衡量输入与其去噪变换之间的不一致性。我们证明，这种组合能够成功防御灰盒攻击者（他们事先不了解检测方法）以及最强的白盒攻击者（他们知道检测的存在，并且能够实施对策）。整体框架如图2所示。

本文的主要贡献总结如下。据我们所知，这是首次以无监督方式防御深度哈希中的对抗攻击。基于新颖的发现与分析，我们提出了三项标准，用于揭示哈明空间中定向攻击和非定向攻击的对抗行为，并证明了这些标准在应对最强白盒攻击者时具有互补关系。在CIFAR-10、ImageNet、MS-COCO和 NUSWIDE 数据集上进行的大量实验表明，与当前最先进的防御方法相比，所提出的方法在检测率方面最高提升了 23%，同时针对实时图像查询仅产生可忽略的计算开销。

2. 预备知识

本节介绍深度哈希和对抗攻击的基础知识。

2.1. 深度哈希

给定一个包含 N个样本 $X = \{x_1, x_2, \ldots, x_N\}$、$x_i \in \mathbb{R}^D$ 及其对应标签 $Y = \{y_1, y_2, \ldots, y_N\}$、$y_i \in \mathbb{R}^C$的数据集，其中 x_i是第 i个样本，且$y_{c,i} = 1$ 表示第 i个图像与类别 c相关联。深度哈希通过符号操作学习函数 $f_\theta(x)$，将输入图像x 映射为 K位二进制编码 $h(x)$,

$$h(x) = sign(f_\theta(x)) \in \{-1, +1\}^K, \tag{1}$$

其中 θ 是通过最小化相似性损失的加权组合而学习得到的参数 $\mathcal{L}_S$ 和量化损失 $\mathcal{L}_Q$ [7, 8, 26, 27, 47, 49],

$$\theta = \arg\min_\theta \mathcal{L}_S + \lambda \mathcal{L}_Q. \tag{2}$$

$\mathcal{L}_S$ 表示两幅图像 x_i 和 x_j 之间的汉明距离 $D_h(h(x_i)\ h(x_j)$，其相似度为 $s(y_i, y_j)$,

$$s(y_i, y_j) = \begin{cases} +1, & \text{if } y_i y_j^T > 0 \\ -1, & \text{otherwise.} \end{cases} \tag{3}$$

$\mathcal{L}_Q$ 是量化损失，用于最小化 $f_\theta(x)$的连续输出与其二进制码$h(x)$之间的差异。当两个样本 x_i 和 x_j 相似时，目标是最小化它们之间的汉明距离$D_h(h(x_i)\ h(x_j)$；当它们不相似时，则最大化汉明距离，同时将连续 $f_\theta(x)$ 表示为二进制码。 $D_h(h_1, h_2)$和 $h(x)$对于其输入均不可微。一种常见的技术是在反向传播过程中使用 $D_h(h_1, h_2)$的可微形式，记为 $\frac{1}{2}(K - {h_1}^T h_2)$，其中 h_1、h_2 是 $[-1, +1]$中的连续浮点表示，而二进制哈希码 $h(x)$则由 $f_\theta(x)$的连续输出表示。这种连续表示与二进制表示之间的差距被视为量化损失 $\mathcal{L}_Q$，并在式 (2) 中予以最小化。

深度哈希由两个主要部分组成：一个数据库 和一个模型。数据库存储图像及其预先计算的哈希码。给定一张查询图像 x，其哈希码为 $h(x)$，系统会从数据库中返回排名前k 的图像，这些图像是 $h(x)$的 k个最近邻，邻近程度由汉明距离确定。检索性能通过平均精度均值（mAP）计算，即与 x相似的图像所占的比例。本文基于当前最先进的方法——中心相似度量化（CSQ） [47]构建哈希框架。CSQ 基于 Hadamard 矩阵预先确定最优哈希码，并从 Hadamard 矩阵中随机选取一组彼此距离足够远的哈希码作为哈希中心（如果 Hadamard 矩阵不可用，则从随机二进制矩阵中选取）。由于不同的哈希技术都具有公式（2）所示的共同目标，我们的防御方法也适用于其他技术 [7, 8, 26, 27, 49]。

2.2. 对抗攻击

无目标攻击 [46] 通过最大化对抗样本与原始图像的哈希码之间的汉明距离来寻找对抗图像 x'，同时满足$\mathcal{L}_\infty$ 界限 ϵ。

$$\begin{aligned} &\max_{x'} D_h\big(h(x'), h(x)\big) \\ &\text{s.t. } \|x - x'\|_\infty \le \epsilon \end{aligned} \tag{4}$$

通过将原始图像推向哈希空间中最远的汉明距离，它能有效降低平均精度均值。

定向攻击 [6, 40, 42] 尝试将与 x' 到目标哈希码 h_t 之间的汉明距离最小化

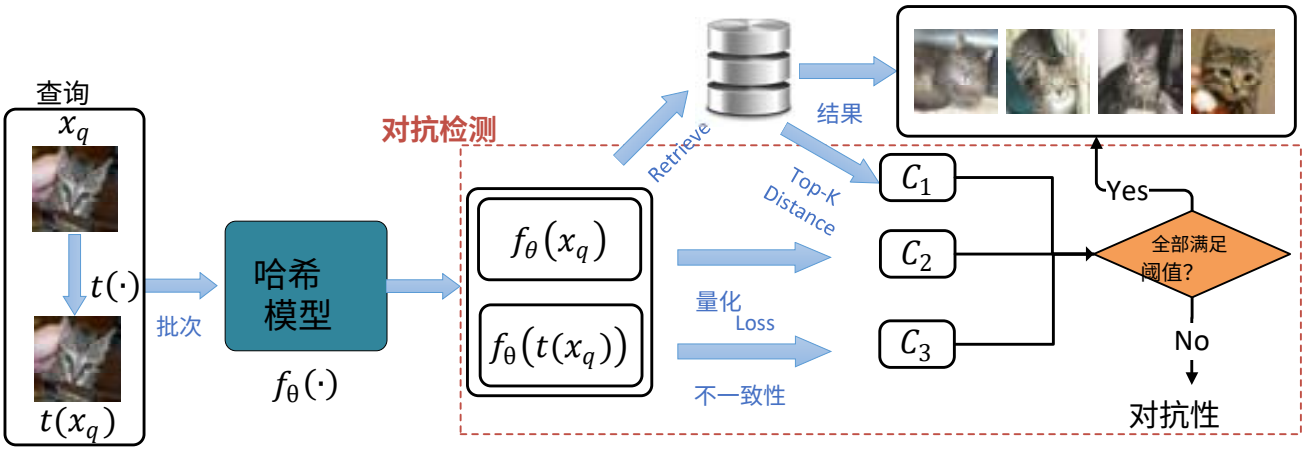


图2。Th　所提出的检测框架：由虚线　表示。

特定类别，

$$\min_{x'} D_h(h(x'), h_t) \quad \text{s.t. } \|x - x'\|_\infty \le \epsilon \tag{5}$$

攻击者将对抗图像嵌入数据库后，定向攻击便可使用户通过查询检索到特定类别的图像。例如，如图 [42]，查询面霜时，数据库可能会错误地返回数据库中的品牌啤酒广告。

攻击模型。攻击者可以实施非定向攻击和定向攻击。具体而言，我们考虑两种类型的灰盒 和 白盒 攻击者。灰盒 攻击者可以访问所有信息，包括网络架构、权重和数据，但不知道存在对抗检测。更强大的白盒 攻击者既了解模型函数/参数，也知道检测机制的存在。因此，他们会采用不同的规避策略，如第4节所述。

3. 汉明空间中的对抗行为

在分类网络中，对抗图像会留下多种痕迹，其中最明显的“对抗行为”之一来自 softmax函数 [19,23]。由于指数运算增长迅速，它会放大 logits 中的微小变化 [19]，并通过抑制其他类别，在存在对抗图像时变得过度自信 [23]。与 softmax 不同，softmax 根据封闭类别集做出决策，而哈希则将相似图像映射到开放汉明空间 $\{-1,1\}^K$中的紧凑汉明球。本节定义了三项准则，用于识别汉明空间中的对抗行为。

3.1. 检测非定向攻击（C_1）

我们首先讨论最大化 x' 与 x [46]之间汉明距离的非定向攻击。尽管这种行为很容易辨别，我们仍希望从理论上回答：当攻击能力达到最大时，$h(x')$的分布如何。

假设1。网络能够学习到完美的哈希码，使类内距离最小（即为零），且彼此之间的间隔最大。

如果假设1成立，那么对抗性 h（x'）与其余哈希码之间的汉明距离是多少？为回答这一问题，我们首先通过以下引理建立最大类间距离的分布。

引理1. 给定汉明空间中若干个具有（理想情况下）紧凑哈希码的 C 类别，K 位汉明空间中的类间汉明距离服从二项分布 $\mathcal{X} \sim B(K, p)$，其中 $p = \frac{C}{2(C-1)}$，并且当 C 很大时，$p \approx \frac{1}{2}$。

证明. 对于所有 C 个类别，每次只考虑一个比特位置。当 C 个类别中 $\frac{C}{2}$ 、 $\{+1\}$ 和 $\{-1\}$ 代码的数量相等时，汉明距离达到最大值。两个比特之间的汉明距离为 0 或 1。因此，在 $\binom{C}{2}$ 个配对选择中，汉明距离等于 1 的概率为 $\binom{C}{2} \cdot \frac{C}{2} / \binom{C}{2} = \frac{C}{2(C-1)}$。由于所有 K 个比特均可独立选取，h_i 与 h_j 之间的类间汉明距离等于 d 的概率为：

$$Pr\big(D_h(h_i, h_j) = d\big) = \binom{K}{d} p^d (1-p)^{K-d}, \quad p = \frac{C}{2(C-1)}, \tag{6}$$

其中均值为 Kp，方差为 $Kp(1-p)$。 □

由引理*1*，我们可以进一步推导出下一个定理。

定理1。对于非定向攻击，对抗性图像到任何其他类别的汉明距离服从高斯分布 $\mathcal{N} \sim (K(1-p)\ Kp(1-p))$。

证明。在理想情况下，无目标攻击会最大化 $D_h(h(x')\ h(x_i))$ 与 K 之间的汉明距离。

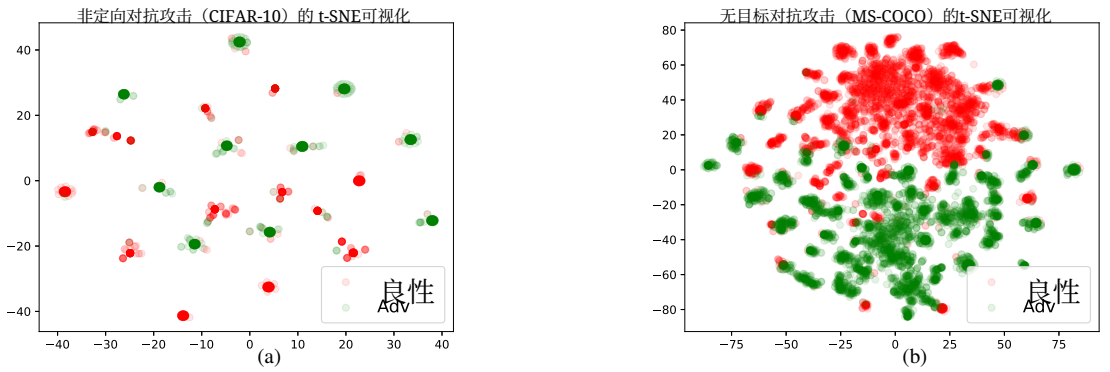


图3. 不同数据集的非定向对抗图像与原始图像的t-SNE可视化 (a) CIFAR-10. (b) MS-COCO.

因此，对于任意其他哈希码 $h(x_j)$，汉明距离为 $D_h(h(x') h(x_j)) = K - D_h(h_i, h_j)$，这同样服从均值为 $K(1-p)$、方差相同的二项分布。当哈希位数 K 较大时，可将其近似为高斯分布$\mathcal{N} \sim (K(1-p)\ Kp(1-p))$ [34]。 □

示例。当 $K = 64$ 位，且 C 较大时（$p \to_1 {}_2$），使用三西格玛法则，置信区间为（$K(1-p) - 3\sqrt{Kp(1-p)}$ $K(1-p) + 3\sqrt{Kp(1-p)}$）。换句话说，有99.73%的置信度认为，从无目标对抗攻击图像到任何其他类别的汉明距离都将位于 $[20, 44]$ 汉明距离区间内，其均值为 $K/2 = 32$，这在汉明空间中具有足够的可区分性。请注意，上述分析构成了理论上的上界，因为实现假设1仍是一项持续进行的工作 [20,47]。为了展示一些示例，我们在图 3中将CIFAR-10 和 MS-COCO 上的无目标对抗攻击图像与良性图像的 t-SNE可视化。可以观察到，尽管存在少量样本，但大多数对抗图像都能基于汉明距离充分区分。因此，我们正式给出第一个检测标准。

标准1（汉明距离）。对于查询 x，收集前-k 个哈希码 $\mathcal{H}_k$ 并计算到 $h(x)$ 的平均汉明距离。

$$C_1 = \frac{1}{|\mathcal{H}_k|} \sum_{h(x_k) \in \mathcal{H}_k} D_h\big(h(x_k), h(x)\big) \tag{7}$$

C_1 是前k 个检索结果的平均汉明距离，即一个标量值，我们可以将其与根据良性样本计算得到的阈值 $\mathcal{T}_1$ 进行比较。C_1 的计算过程遵循使用前k 个哈希码的常规检索流程。为了检测定向攻击，我们提出下一个判据。

3.2. 检测定向攻击 (C_2)

虽然非定向攻击试图诱导发生位翻转，使 $h(x') = -h(x)$，但定向攻击会最小化 $h(x')$ 与任意目标代码之间的汉明距离 h_t（例如，通过共识投票计算得出 [6] 的类别）。为了找到一种适合识别它们的指标，我们有如下观察。

观察 1. 对于良性图像 $\mathcal{L}_Q^b$ 和目标图像 $\mathcal{L}_Q^t$的量化损失，关系式 $\mathcal{L}_Q^b > \mathcal{L}_Q^t \approx 0$ 成立。

为说明这一观察，回顾一下公式（5）中的原始定向目标函数对于 x'的定向二进制代码不可微。该实现通过连续松弛进行近似，其目标是最小化$\tanh(\cdot)$ 函数的连续输出与目标代码 [8]之间的距离。当它们之间的距离被最小化时，采取更多的梯度下降步骤会使量化损失 $\mathcal{L}_Q^t \to 0$。这与Softmax 分类中的对抗图像非常相似：定向概率会变得过度自信 [23]，而且令人惊讶的是，深度哈希中的量化损失也呈现出类似现象。相比之下，对于所有良性样本，很难找到最优的模型参数，在训练过程中将 $\mathcal{L}_Q^b$ 推向零。图 44(a) 展示了一个ImageNet 示例，因为定向攻击与良性样本之间存在可区分的间隔。将其与图 44(a) 中无目标攻击的量化损失进行比较也很有意思，后者大于零。这是因为对于无目标攻击，在翻转所有比特之前找到一个将平均精度均值降低至零的对抗子空间即可实现攻击，这比定向攻击容易得多。基于这些观察，我们提出了第二个检测标准。

标准 2（量化损失）。计算网络 $f_\theta(x)$的输出（符号函数之前的 logits）与其哈希码 $h(x)$之间的 $\boldsymbol{L}_p$ 距离，

$$C_2 = \|h(x) - f_\theta(x)\|_p \tag{8}$$

C_2 是查询的量化损失 x。在这里，我们使用$\boldsymbol{L}_1$ 距离（$p = 1$）并在线下良性样本上获得一个阈值 $\mathcal{T}_2$。图44（b）（c）展示了对抗性图像与良性图像之间量化损失的分布。如 $C_2 \to 0$ 对于定向攻击，我们可以看到，使用 $\mathcal{T}_2$

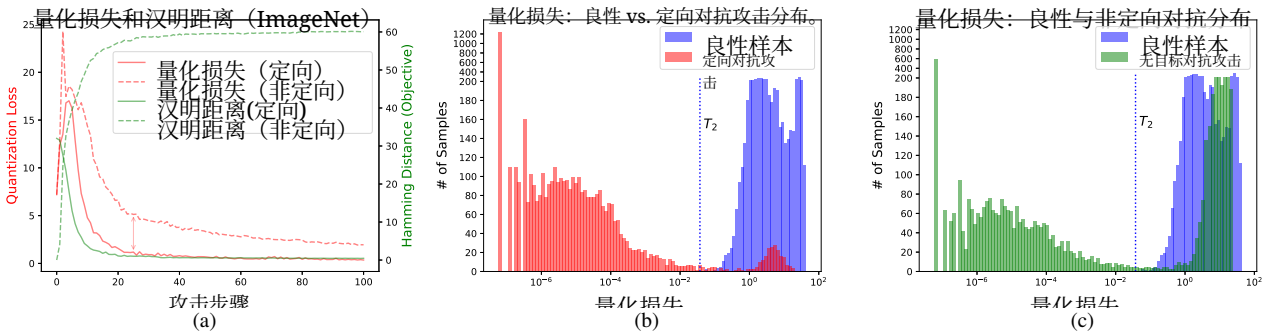


图 4. 基于 ImageNet 上的量化损失识别定向攻击的示例。(a) 与良性样本相比，定向攻击的量化损失集中在零附近。(b) 与非定向攻击相比，定向攻击会将量化损失推向零。(c) 60% 的非定向攻击也集中在零附近。

能够有效识别大多数攻击；使用 C_2 还能识别约60%的非定向攻击，

3.3. 检测预测不一致性 (C_3)

C_1 和 C_2 单独使用并不足够。原则上，检测的工作原理是将攻击者的行动空间限制在受限区域内。扰动通常可以视为一种带有高频分量的人工噪声 [39]。因此，一种常见方法是应用局部、非局部平滑滤波器 [45], 、自编码器去噪器 [30], 、颜色位深缩减 [45], 量化 [25], , 并测量对去噪图像的响应敏感度。对抗图像更容易产生不同的结果，而良性样本则不那么敏感。这些去噪操作降低了熵（随机性），并减少了扰动可作用的对抗空间输入维度。

我们将这一原理扩展到深度哈希中，以构建准则 3。将变换 [25, 30, 45] 记为 $t(\cdot)$。对于查询 x， C_3 根据符号函数作用前的输出，度量变换后的 $t(x)$ 与 x 之间的汉明距离。

准则 3（预测不一致性）。

$$C_3 = D_h\big(f_\theta(t(x)), f_\theta(x)\big) \tag{9}$$

换言之， C_3 量化原始输入与变换后输入之间的差异，可将其与在良性样本上离线计算的阈值 $\mathcal{T}_3$ 进行比较。

3.4. 整合所有内容

整体检测结合了三项标准：给定查询图像 x，我们计算 $\{C_1, C_2, C_3\}$ 并与阈值 $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\}$进行比较。如果 (a) $C_1 < \mathcal{T}_1$; (b) $C_2 > \mathcal{T}_2$; (c) $C_3 < \mathcal{T}_3$，则将输入视为良性；否则，只要有任一项不满足，就会将输入拒绝为对抗性示例。由于需要进行两次检索，计算时间上限为 C_3 。为尽量缩短计算时间，系统可以将原始查询及其去噪副本合并为一个批量查询。如果 GPU 有足够的资源，则开销应当很小，正如第4.5节所述。

4. 实验

4.1. 实现

我们在深度哈希中常用的 CIFAR-10、ImageNet、MS-COCO 和 NUSWIDE 数据集上评估了我们的机制 [7, 8, 26, 27, 47, 49]，并采用 CSQ [47]和 ResNet50 [18] 作为基础模型。使用 RM-SProp 优化器 [37] ，学习率为 10^{-5} ，训练 150 个周期。量化损失的权重设为 10^{-4}。在这四个数据集上，我们训练的模型分别达到 0.854、0.883、0.884 和 0.843 的平均精度均值。

我们将其与多个最初为 Softmax 分类设计的基准方法进行比较：局部内在维度（LID） [28]，中值平滑（FS-Median）、非局部均值（FS-NLM） [45], FS-Adaptive [25], 和均值模糊 [23]。FS-Adaptive 使用输入的熵作为度量，通过标量量化和空间平滑滤波器自适应地缩减输入空间。我们选择均值模糊作为我们的方法的去噪技术。真阳性率（TPR）、假阴性率（FNR，漏检率）和曲线下面积（AUC）被用作评估指标，其中对抗样本被视为阳性，良性样本被视为阴性。因此，被检测出的对抗样本计为真阳性，而被误识别的良性样本计为假阳性。

所有检测方法均针对非定向 [46] 和定向 [6] 深度哈希对抗攻击（基于 PGD 攻击）以及一种非定向深度哈希 CW [10] 攻击[1]进行评估。前两种攻击的步长设为 1.0，100 步，并受扰动范数 $\boldsymbol{L}_\infty$ 限制 $\epsilon = 8$。对于 CW 攻击，学习率设为 0.01 ，且为

[1]CW攻击最初是为 Softmax 分类设计的，因此只需优化少数几个 logits。但对于深度哈希，需要优化所有哈希位，而这对于定向 CW攻击来说很少能成功。因此，我们专注于深度哈希中的非定向 CW攻击。

500步。其他实验细节和结果可在附录中找到。

4.2. 灰盒攻击的检测

		非定向 [46]	定向 [6]	非定向 CW [10]
CIFAR-10	LID [28]	14.90	**4.00**	12.16
	FS-Median [45]	13.40	54.30	**2.40**
	FS-NLM [45]	18.80	18.10	14.41
	FS-Adaptive [25]	54.00	48.30	52.55
	均值模糊 [23]	**11.60**	26.90	**2.40**
	UMCD（我们的方法）	**7.50**	**0.40**	**0.15**
ImageNet	LID [28]	21.78	6.40	75.95
	FS-Median [45]	8.54	19.20	1.41
	FS-NLM [45]	2.62	**3.40**	8.02
	FS-Adaptive [25]	4.05	4.30	13.68
	均值模糊 [23]	**0.76**	4.54	**0.94**
	UMCD (我们的方法)	**0.42**	**0.34**	**0.47**
MS-COCO	LID [28]	3.24	**1.20**	54.59
	FS-Median [45]	0.26	7.10	**0.00**
	FS-NLM [45]	**0.00**	1.72	0.08
	FS-Adaptive [25]	**0.00**	2.84	0.08
	均值模糊 [23]	**0.00**	2.12	**0.00**
	UMCD（我们的方法）	**0.00**	**1.12**	**0.00**
NUSWIDE	LID [28]	25.34	**0.08**	47.96
	FS-Median [45]	3.52	16.33	**0.00**
	FS-NLM [45]	**0.00**	2.90	**0.00**
	FS-Adaptive [25]	**0.00**	4.57	**0.00**
	均值模糊 [23]	0.19	4.29	**0.00**
	UMCD (我们的方法)	**0.00**	**1.29**	**0.00**

表1. 与当前最先进方法的比较针对深度哈希的对抗攻击的检测漏检率（假阴性率），在良性样本上允许 5% 的假阳性率时。漏检率越低越好。每列中排名前两位的数值以**粗体**显示，最佳结果以**红色**显示，第二佳结果以**蓝色**显示。

我们的方法。 表1 展示了在将 FPR 固定为 5% 时，不同方法的检测漏报率。所提出的方法能够对定向攻击进行稳健检测，在大多数情况下的漏报率低于 1%。与最先进水平的基准方法相比，我们的方法平均2.13%到 23.44% 优于其他方法。非定向攻击相对更容易检测，因此总体漏报率低于定向攻击。我们的方法仍可使基线方法提升1.15%-14.33%。

与 LID 比较。 尽管 LID 在大多数定向攻击上表现良好，但无法推广到 CW攻击。此外，与 soft max 网络不同，后者通常由最后几层提供更好的检 [28], 测效果，LID 对于在深度哈希中应从哪些层提取特征非常敏感，这增加了额外的配置开销。总而言之，LID 并非始终能有效应对所有类型的攻击。

与空间去噪方法比较。 LID 不像其余方法那样利用空间信息采用非局部均值、中值滤波等方法的四个基准测试。如表1所示，尽管它们在非定向攻击上的漏检率通常低于 10%，但在检测定向攻击方面存在 0.5-17.8%的差距。这些差距可以通过攻击机制来解释，因为定向攻击需要更多的梯度步数。这会使图像深入对抗空间，从而对去噪等像素级修改具有更强的鲁棒性 [21, 42]。不过，我们的方法还可从 C_2 提供额外一层防御，以专门监测量化损失的值，因为定向攻击会将其降至零，从而补充 C_3 在定向攻击将输入推入对抗空间深处时的作用。

4.3. 白盒攻击检测

接下来，我们展示如何检测白盒攻击者：他们知道检测机制的存在，并据此采取对策。攻击者采用反向传播可微近似 [5] 来估计梯度，并针对 C_1、 C_2和 C_3 制定不同策略：

针对 C_1。 C_1 依靠哈希码之间的汉明距离来检测异常值。因此，有效的规避方法是将对抗样本推入良性图像的邻域，例如，生成与某些目标图像相同的二进制哈希码 h_t:

$$\mathcal{L}_1 = \underbrace{D_h(f_\theta(x'), h_t)}_{\text{adv loss}} \tag{10}$$

针对 C_2。 C_2 通过读取符号函数之前的 logits 来检测接近零的量化损失。为绕过这一检测，攻击者旨在最大化量化损失，以削弱从 C_2中识别出的对抗行为。

$$\mathcal{L}_2 = -\underbrace{\left\|h(x') - f_\theta(x')\right\|_1}_{\text{quantization loss}} \tag{11}$$

针对 C_3。 C_3 检测f_θ（x'）与去噪副本 f_θ（t（x'））之间的不一致。攻击者通过约束 f_θ（x'）与 f_θ（t（x'））之间的距离较小，来最小化这种差异，

$$\mathcal{L}_3 = \underbrace{D_h(f_\theta(t(x')), f_\theta(x'))}_{\text{denoised adv loss}} \tag{12}$$

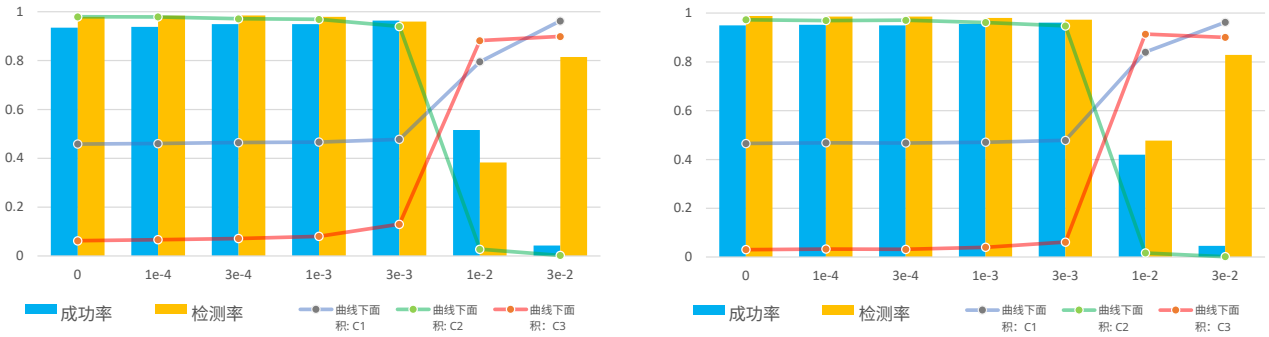
我们在此使用 MeanBlur [23] 作为变换 $t(\cdot)$。通过将它们结合起来，白盒攻击者构建了一个联合优化目标，

$$\min_{x'} \mathcal{L} = \mathcal{L}_1 + \lambda_1 \mathcal{L}_2 + \lambda_2 \mathcal{L}_3 \tag{13}$$

白盒攻击检测。 优化（13）通过寻找 λ_1， λ_2 同时违反三个标准。我们在图 5 中展示了将 λ_2 固定为 0。3 和 3 并将 λ_1 从 0 调整到 0。03。最佳

	CIFAR-10		ImageNet		MS-COCO		NUSWIDE		平均增益	
	非定向	Tgt	非定向	Tgt	非定向	Tgt	非定向	Tgt	非定向	Tgt
C_3 单独	0.8160	0.7460	0.9110	0.8338	0.9954	0.9076	0.9757	0.8904	–	–
$C_1 + C_3$	0.9870	0.7580	0.9522	0.8460	0.9966	0.9098	0.9757	0.8904	*0.0533*	*0.0066*
$C_2 + C_3$	0.8170	0.9830	0.9504	0.9828	0.9992	0.9784	0.9961	0.9847	*0.0161*	*0.1377*
$C_1 + C_2 + C_3$	0.9880	0.9950	0.9916	0.9956	0.9992	0.9784	0.9961	0.9847	*0.0692*	*0.1439*

表 2. 消融研究：不同组合的检测率（$\epsilon = 32$）



(a) $\lambda_2 = 0.3$ (b) $\lambda_2 = 3$

图5. 白盒攻击结果。x轴为 λ_1 值，y轴为百分比。当白盒攻击在 λ_1 =1e-2处达到较低的检测率时，生成此类对抗样本的成功率也会显著下降。通过观察曲线下面积值可以看出，三个检测标准相互补充：当 C_1, C_3 较低时， C_2 较高，反之亦然。

情况是，当 λ_1=0.01时，检测率已降至约0.4（黄色柱）。然而，能够成功优化的对抗样本数量（13）也降至约45%（成功率的蓝色柱）。因此，尽管最强的白盒攻击者仍有一些机会，但我们的检测通过增加攻击者的努力，成功限制了对抗空间。有趣的是，不同标准之间形成了相互补偿的关系，这一点可由AUC值体现。当 C_2（绿色曲线）下降时， C_1， C_3 会迅速上升，反之亦然。接下来的消融研究进一步验证了这一关系。

4.4. 消融研究

我们开展消融研究，以量化每个标准的贡献。我们单独使用 C_3 作为基线，并加入 C_1 和 C_2 ，其平均增益如表2所示。结果与防御目标一致，因为加入 C_1 和 C_2 分别有助于将非定向攻击和定向攻击的检测率提高**0。0533** 和**0。1377**。与此同时， C_1 和 C_2 对总体检测的贡献几乎相互独立，例如， $C_1 + C_3 = \mathbf{0.0533}$ 加上 $C_2 + C_3 = \mathbf{0.0161}$ 等于 $C_1 + C_2 + C_3 = \mathbf{0.0692}$对于非定向攻击，定向攻击也同样如此。这验证了这三个标准在检测中都是不可或缺的组成部分。

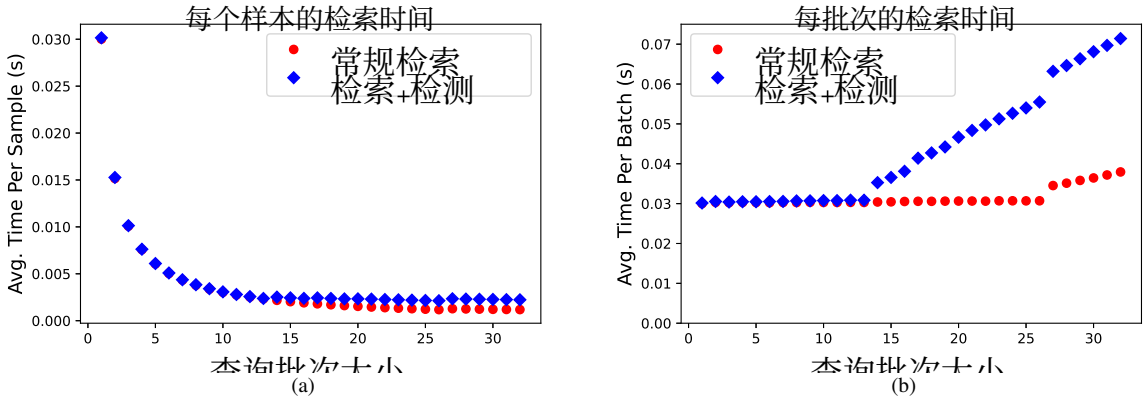
4.5. 计算时间

最后，我们评估了检测机制带来的计算开销。实际上，系统可以将查询累积成批次，以提高 GPU资源 的利用率并降低成本。次成批，以提高 GPU资源 的利用率并降低成本。图6展示了每个样本/批次的平均检索时间。首先可以看到，每个样本的平均耗时低于 50 毫秒，并且随着批次大小增加而进一步降低。当批次较小时，由于 GPU 未得到充分利用，检测带来的开销可以忽略不计；随着批次大小增加，从去噪副本进行额外检索会 C_3 扩大其与常规检索之间的差距，因为 GPU 资源已得到充分利用。因此，当系统累积的批次相对较小时，我们的检测机制带来的开销最小，并且能够实时响应查询。

5. 相关工作

5.1. 深度哈希

图像检索使用最近邻搜索，返回与查询输入在语义上相关的图像。传统上，它依赖人工设计的视觉描述符来降低相似度计算的计算成本 [14, 31]。在深度学习的推动下，端到端哈希学习将性能提升到了新的水平 [7, 8, 26, 27, 47, 49]。这类方法利用图像对之间的相似性，以监督方式训练深度哈希模型，将高维图像转换为紧凑的哈希码，并基于汉明距离高效地执行邻近搜索。为将连续输出转换为离散的二进制码，常见方法采用连续松弛，例如使用 sigmoid 或双曲正切



(a) (b)

图 6. 不同批次大小下的计算时间：a）每个样本；b）每个批次。

函数来近似离散二进制阈值化 [7, 8, 26, 27, 47, 49]。我们的工作利用了这一近似过程所产生的对抗行为，因此可应用于各种深度哈希模型。

5.2. 对抗攻击

众所周知，深度神经网络容易受到不可感知扰动 [36]的影响。快速梯度符号法（FGSM） [16] 沿符号梯度方向生成扰动，通过一次计算最大化损失函数。基本迭代法（BIM） [22] 和投影梯度下降（PGD） [29]通过迭代步骤（从随机初始化开始）实现更高的攻击成功率。还有其他几种变体 [5, 10, 32, 38],，例如CW攻击旨在尽量减小扰动，以逃避检测。

借助深度学习后端，深度哈希继承了神经网络的脆弱性。近期研究表明，只需稍作调整，对抗攻击也能误导图像检索系统 [6, 40, 42,43, 46]。攻击通常可分为非定向和定向攻击。非定向攻击会使查询偏离正确结果，导致系统检索到无关图像，甚至什么也检索不到。 [46] 提出了一种非定向攻击，通过最大化对抗样本与良性样本之间的汉明距离来实现攻击。 [12, 24] 基于从黑盒模型进行的迭代检索来构造对抗样本。 [43] 通过尽量减少私有图像周围的样本数量，将数据库中的私有图像隐藏在不可检索子空间中。定向攻击会使系统返回来自目标类别、且不同于输入的图像。 [6, 40] 通过最小化对抗样本与一组带有目标标签的图像之间的平均汉明距离来实现攻击。 [42] 通过在对抗样本生成过程中注入随机噪声，增强了对黑盒模型的有目标迁移性。我们的工作旨在防御深度哈希中的非定向攻击和定向攻击。

5.3. 对抗防御

大多数防御机制都基于 softmax 分类。作为主动防御措施，梯度掩蔽 [33] 和对抗训练 [13, 29, 35, 41] 能够学习鲁棒模型。 [33] 提出的早期防御始于一种错误的推测，即将对抗样本归因于高度非线性或过拟合，并据此开发出防御性蒸馏，以减少输入周围的变化。正如 [16] 所论证的， [5, 9, 10] 很快就攻破了该方法，指出问题的主要原因其实是神经网络的局部线性。

因此，大量研究聚焦于对抗训练 [13, 29, 35, 41]，通过求解一个极小极大鞍点问题来实现。然而，借助大量训练工作，在鲁棒性与准确率 [48], 之间取得平衡并非易事，这通常会导致干净图像准确率显著下降。正如我们的实验所示，将对抗训练应用于深度哈希领域会遭受更高的准确率损失（见附录）。对于图像检索系统，只要在输入端检测出对抗图像，我们就可以在不损失准确率且无需增加训练复杂度的情况下，有效阻止攻击。

对抗性检测从对抗样本在不同层面留下的伪影中提取特征：原始像素 [15, 17],特征分布 [17, 23], Softmax分布 [19] 和频率分量 [39]。通过分析对抗图像与自然图像的对比分布，检测器可以以有监督或无监督的方式进行高效训练。另一类研究则依靠预测结果的不一致性，通过利用去噪方法并衡量结果之间的分歧来实现检测 [25, 30, 45]。所有这些工作都基于 Softmax 分类。在本文中，我们从汉明空间中发现对抗行为，并提出了一组检测标准，包括防御最强的白盒攻击者。

6. 结论

本文提出了一种高效的无监督方法，用于检测基于深度哈希的图像检索中的对抗样本。

检索。我们设计了三个标准，用于在汉明空间中识别定向攻击和非定向攻击的对抗性行为，并考虑了知晓防御存在的白盒攻击者。大量评估表明，所提出的检测方法能够以较大间隔超越以往的防御技术，并且通过限制白盒攻击者的行动空间，对其也具有鲁棒性。

参考文献

[1] 阿里巴巴拍立淘。http://www.pailitao.com。访问日期：2022-02-03。1[2] 必应图像搜索。https://www.bing.com。访问日期：2022-02-03。1[3] 谷歌图像搜索。https://www.google.com/ imghp。访问日期：2022-02-03。1[4] Pinterest 视觉搜索工具。https : / / www . pinterest.com。访问日期：2022-02-03。1[5] Anish Athalye、Nicholas Carlini 和 David Wagner。《混淆梯度带来虚假的安全感：规避针对对抗样本的防御》。载于机器学习国际会议论文集，第274–283页。机器学习研究论文集，2018年。6, 8[6] Jiawang Bai、Bin Chen、Yiming Li、Dongxian Wu、Weiwei Guo、Shu-tao Xia 和 En-hui Yang。《针对深度哈希检索的定向攻击》。载于欧洲计算机视觉会议论文集，第618–634页。施普林格，2020年。1, 2, 4, 5, 6, 8[7] Yue Cao、Mingsheng Long、Bin Liu 和 Jianmin Wang。《用于汉明空间检索的深度柯西哈希》。载于IEEE计算机视觉与模式识别会议论文集，第1229–1237页，2018年。1, 2, 5, 7, 8[8] Zhangjie Cao、Mingsheng Long、Jianmin Wang 和 Philip S Yu。《HashNet：通过延续法进行深度学习哈希》。载于IEEE国际计算机视觉会议论文集，第5608–5617页，2017年。1, 2, 4, 5, 7, 8[9] Nicholas Carlini 和 David Wagner。《防御性蒸馏无法抵御对抗样本》。arXiv预印本 arXiv:1607.04311，2016年。8[10] Nicholas Carlini 和 David Wagner。《评估神经网络鲁棒性》。载于2017年IEEE安全与隐私研讨会（SP），第39–57页。IEEE，2017年。5, 6, 8[11] Fabio Carrara、Rudy Becarelli、Roberto Caldelli、Fabrizio Falchi 和 Giuseppe Amato。《特征距离空间中的对抗样本检测》。载于欧洲计算机视觉会议（ECCV）研讨会论文集，2018年9月。1[12] Mingyang Chen、Junda Lu、Yi Wang、Jianbin Qin 和 Wei Wang。《DAIR：针对图像检索系统的查询高效的基于决策攻击》。载于第44届ACM信息检索研究与开发国际会议论文集，第1064–1073页，2021年。8[13]Francesco Croce 和 Matthias Hein。《通过多样化无参数攻击集成实现对抗鲁棒性的可靠评估》。载于机器学习国际会议论文集，第2206–2216页。机器学习研究论文集，2020年。8

[14] NavneetDalal和 Bill Triggs。用于人体检测的定向梯度直方图。载于 2005年 IEEE计算机协会计算机视觉与模式识别会议（CVPR’ *05* 1, 886 893 2005 7），卷 页码 –。IEEE，。[15] ZhitaoGong、WenluWang和 Wei-Shinn Ku。对抗性数据和干净数据并非孪生数据。arXiv 预印本 arXiv:1704.04960，2017年。1，8[16] Ian J Goodfellow、Jonathon Shlens 和 Christian Szegedy。解释并利用对抗样本。arXiv 预印本 arXiv:1412.6572，2014年。8[17] Kathrin Grosse、Praveen Manoharan、Nicolas Papernot、MichaelBackes 和 PatrickMcDaniel。关于对抗样本的（统计）检测。arXiv预印本 arXiv:1702.06280，2017年。1，8[18] Kaiming He、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于IEEE计算机视觉与模式识别会议论文集，第 770–778 页，2016年。5[19] Dan Hendrycks和 Kevin Gimpel。神经网络中错误分类样本和分布外样本检测的基线。arXiv预印本arXiv:1610.02136，2016年。1，3，8[20] Jiun TianHoe、Kam Woh Ng、Tianyu Zhang、Chee SengChan、Yi-ZheSong和 TaoXiang。一种损失适用于所有任务：基于单一余弦相似度学习目标的深度哈希。神经信息处理系统进展，34，2021年。4[21] Shengyuan Hu、TaoYu、Chuan Guo、Wei-LunChao 和Kil-ianQWeinberger。针对对抗图像的新型防御：化弱点为优势。神经信息处理系统进展，32，2019年。6[22] Alexey Kurakin、Ian JGoodfellow和Samy Bengio。物理世界中的对抗样本。arXiv 预印本 arXiv:1607.02533，2016年。8[23] Xin Li和Fuxin Li。利用卷积滤波器统计量检测深度网络中的对抗样本。载于 IEEE国际计算机视觉会议论文集，第 5764–5772 页，2017年。1，3，4，5，6，8，11[24] Xiaodan Li、Jinfeng Li、Yuefeng Chen、Shaokai Ye、Yuan He、ShuhuiWang、HangSu 和Hui Xue。Qair：用于图像检索的实用查询高效黑盒攻击。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 3330–3339 页，2021年。8[25] BinLiang、Hongcheng Li、Miaoqiang Su、Xirong Li、Wen-changShi 和 Xiaofeng Wang。通过自适应噪声降低检测深度神经网络中的对抗图像样本。IEEE可靠与安全计算汇刊，18(1)：72–85，2018年。5，6，8[26] Kevin Lin、Huei-Fang Yang、Jen-HaoHsiao和 Chu-Song Chen。用于快速图像检索的二进制哈希码深度学习。载于 IEEE计算机视觉与模式识别会议研讨会论文集，第 27– 35 页，2015年。1，2，5，7，8[27] Haomiao Liu、RuipingWang、ShiguangShan 和 Xilin Chen。用于快速图像检索的深度监督哈希。载于 IEEE计算机视觉与模式识别会议论文集，第2064–2072 页，2016年。1，2，5，7，8[28] Xingjun Ma、Bo Li、Yisen Wang、Sarah M Erfani、SudanthiWijewickrema、Grant Schoenebeck、Dawn Song、Michael E

Houle 和 James Bailey。使用局部内在维度刻画对抗子空间。arXiv 预印本 arXiv:1801.02613，2018年。1，5，6[29] Aleksander Madry、Aleksandar Makelov、Ludwig Schmidt、Dimitris Tsipras 和 Adrian Vladu。迈向能够抵御对抗攻击的深度学习模型。载于《国际学习表征会议》，2018年。8[30] Dongyu Meng 和 Hao Chen。Magnet：一种针对对抗样本的双管齐下防御方法。载于《2017年 ACM SIGSAC 计算机与通信安全会议论文集》，第135 147, 2017 5, 8 –页。Aude[31] Oliva 和 Antonio Torralba。场景形状建模：空间包络的整体表征。《国际计算机视觉杂志》，42(3):145–175，2001年。7[32] Nicolas Papernot、Patrick McDaniel、Ian Goodfellow、Somesh Jha、Z Berkay Celik 和 Ananthram Swami。针对机器学习的实用黑盒攻击。载于《2017年 ACM 亚洲计算机与通信安全会议论文集》，第506–519页，2017年。8[33] Nicolas Papernot、Patrick McDaniel、Xi Wu、Somesh Jha 和 Ananthram Swami。蒸馏作为抵御深度神经网络对抗性扰动的防御方法。载于《2016年 IEEE 安全与隐私研讨会（SP）》，第582–597页。IEEE，2016年。8[34] Sheldon M Ross。《概率论入门》，第6卷。Prentice Hall，Upper Saddle River，新泽西州，2002年。4[35] Ali Shafahi、Mahyar Najibi、Amin Ghiasi、Zheng Xu、John Dickerson、Christoph Studer、Larry S Davis、Gavin Taylor 和 Tom Goldstein。免费进行对抗训练！arXiv 预印本 arXiv:1904.12843，2019年。8，10，11[36] Christian Szegedy、Wojciech Zaremba、Ilya Sutskever、Joan Bruna、Dumitru Erhan、Ian Goodfellow 和 Rob Fergus。神经网络的有趣特性。arXiv 预印本 arXiv:1312.6199，2013年。1，8[37] Tijmen Tieleman、Geoffrey Hinton 等。第6.5讲——RMSProp：将梯度除以其近期幅度的滑动平均值。COURSERA：机器学习神经网络，4(2):26–31，2012年。5[38] Florian Tram`er、Nicolas Papernot、Ian Goodfellow、Dan Boneh 和 Patrick McDaniel。可迁移对抗样本的空间。arXiv 预印本 arXiv:1704.03453，2017年。8[39] Haohan Wang、Xindi Wu、Zeyi Huang 和 Eric P Xing。高频分量有助于解释卷积神经网络的泛化能力。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第8684–8694页，2020年。1，5，8[40] Xunguang Wang、Zheng Zhang、Guangming Lu 和 Yong Xu。深度哈希的定向攻击与防御。载于《第44届国际 ACM SIGIR 信息检索研究与开发会议论文集》，第2298 2302, 2021 1, 2, 8 –页。Eric Wong、[41] Leslie Rice 和 J Zico Kolter。快速胜过免费：重新审视对抗训练。arXiv 预印本 arXiv:2001.03994，2020年。8[42] Yanru Xiao 和 Cong Wang。你看到的正是我想让你看到的：探索定向黑盒可迁移性攻击

用于基于哈希的图像检索系统。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第1934–1943页，2021年。1、2、3、6、8[43] Yanru Xiao、Cong Wang 和 Xing Gao。通过将私有图像藏匿于哈希空间来规避深度图像检索。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第9651–9660页，2020年。8[44] Cihang Xie、Mingxing Tan、Boqing Gong、Jiang Wang、Alan L Yuille 和 Quoc V Le。对抗样本提升图像识别能力。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第819–828页，2020年。11[45] Weilin Xu、David Evans 和 Yanjun Qi。特征压缩：检测深度神经网络中的对抗样本。arXiv预印本，arXiv:1704.01155，2017年。1、5、6、8[46] Erkun Yang、Tongliang Liu、Cheng Deng 和 Dacheng Tao。用于汉明空间搜索的对抗样本。《IEEE控制论汇刊》，2018年。1、2、3、5、6、8[47] Li Yuan、Tao Wang、Xiaopeng Zhang、Francis EH Tay、Zequn Jie、Wei Liu 和 Jiashi Feng。用于高效图像和视频检索的中心相似度量化。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第3083–3092页，2020年。1、2、4、5、7、8[48] Hongyang Zhang、Yaodong Yu、Jiantao Jiao、Eric Xing、Laurent El Ghaoui 和 Michael Jordan。鲁棒性与准确率之间的理论性权衡。载于《国际机器学习会议论文集》，第7472–7482页。机器学习研究论文集，2019年。1、8[49] Han Zhu、Mingsheng Long、Jianmin Wang 和 Yue Cao。用于高效相似性检索的深度哈希网络。载于《AAAI人工智能会议论文集》，第30卷，2016年。1、2、5、7、8

7. 附录

我们在附录中提供了关于对抗性训练和白盒攻击的额外实验。

7.1. 对抗训练

作为一种主动防御手段，对抗训练旨在求解一个极小极大优化问题：内层优化寻找对抗样本以最大化损失函数，而外层优化则最小化整体损失。因此，模型准确率与鲁棒性之间的权衡成为对抗训练的关键。接下来，我们将说明，在深度哈希中，这种差距会进一步扩大，传统对抗训练方法难以应对。

实现细节。我们的实现基于免费对抗训练（FreeAT）[35]。它将参数更新与对抗样本生成并行化，从而减少内部优化带来的开销。在训练过程中，它会重复使用相同的小批量数据 m 次，以更新模型参数，并为下一次迭代生成对抗样本。较小的 m 对其影响很小

鲁棒性，但较大的值会影响模型准确率。从 [35], Softmax 分类的最佳情况（当 $m = 8$）具有 0。46 来自对抗攻击（PGD 和 CW）的准确率提升，并且在 CIFAR-10上准确率约下降 0。1 。接下来，我们验证类似数值是否适用于深度哈希。

我们采用FreeAT 方法，在CIFAR-10 上使用其配置进行深度哈希。我们设置 $||\epsilon||_\infty = 8$，将四种不同的重复次数 m 从 2到 16，分别用于训练四个ResNet50模型。对于其中的每个模型，我们生成具有不同步数的PGD 对抗样本，分别记为 PGD-8、PGD-40 和 PGD-100，对应 8、40 和 100 步。

训练	评估对象			
	良性	PGD-8	PGD-40	PGD-100
无 FreeAT	0.854	0.028	0.030	0.033
FreeAT $m = 2$	**0.889**	0.070	0.035	0.035
FreeAT $m = 4$	0.838	0.222	0.070	0.055
FreeAT $m = 8$	0.769	**0.328**	**0.146**	**0.130**
FreeAT $m = 16$	0.683	0.278	0.122	0.114

表 3. 在 CIFAR-10 上对深度哈希实施 FreeAT [35] 时的干净样本和对抗样本平均精度均值

结果。 表3 第一行显示，未进行对抗训练时，PGD 攻击已成功将平均精度均值降至接近零。采用 FreeAT 后，PGD-8 下的平均精度均值有所提升，但面对 PGD-40 和 PGD-100 等攻击步骤较多的攻击时，表现仍然不佳。在这里的最佳情况下，当$m = 8$时，采用 100 个步骤的 PGD 攻击仍能将平均精度均值降至 0.130。对抗训练不仅防御保真度很低，还会降低模型在良性样本上的准确率。随着纳入更多 m 步骤，良性样本上的平均精度均值从 0.854 大幅降至 0.683[2]，而对抗鲁棒性并未显著提升。与 Softmax 分类相比，深度哈希中的差距越来越大，因此本文转而研究对抗检测。

7.2. 白盒实验

我们展示了白盒攻击的更多结果。由于攻击者需要针对检测方法设计特定目标，因此我们选择均值模糊 [23] 作为比较基准。检测率是根据良性样本上的假阳性率（FPR），在特定阈值下计算的真阳性率（ TPR）。我们使用 PGD 攻击，并采用 $||\epsilon||_\infty = 32$。为了获得最佳成功率并绕过检测，我们选择 $\lambda_1 = 0.0075$。在表4中，我们展示了在四个数据集上以 TPR 评估的检测结果。给定 FPR= 0.1 和 FPR= 0.2，我们的方法达到平均0.7309 和0.8403 的白盒检测率，而均值模糊的检测率接近于零。这进一步验证了我们的方法能够抵御最强的白盒攻击者。有趣的是，与灰盒攻击不同，两种方法在单标签数据集（CIFAR-10 和 ImageNet）上的表现都相对优于多标签数据集（MS-COCO 和 NUSWIDE）。其背后的原因还需要在未来进一步研究。

检测器	FPR	CIFAR10	ImageNet	MS-COCO	NUSWIDE
均值模糊 [23]	0.1	0.0340	0.0218	0.0044	0.0057
	0.2	0.0480	0.0250	0.0076	0.0076
我们的方法	0.1	0.9560	0.8824	0.4648	0.6204
	0.2	0.9790	0.9180	0.6742	0.7900

表4。针对白盒攻击（PGD、 $\epsilon = 32$、 $\lambda_1 = 0.0075$）的对抗样本检测率（TPR）

[2]唯一的例外是采用 $m = 2$的模型，它提升了良性样本上的性能。这与 [44], 的结论一致：对于某些参数设置，对抗训练能够提升干净图像上的准确率。