哈希基线：在预训练模型时代重新思考

Ilyass Moummad*[1] Kawtar Zaher*[1,2] Lukas Rauch[3] Alexis Joly[1]

[1]INRIA, LIRMM, Université de Montpellier, France
[2]Institut National de l' Audiovisuel, France
[3]University of Kassel, Germany

ABSTRACT

信息检索与紧凑二进制嵌入，也称为哈希，对于可扩展的快速搜索应用至关重要，然而最先进的哈希方法需要昂贵且特定于场景的训练。在这项工作中，我们介绍了哈希基线，这是一种强大的无需训练的哈希方法，利用了能够生成丰富预训练嵌入的强大预训练编码器。我们重新审视了经典的、无需训练的哈希技术——主成分分析、随机正交投影和阈值二值化——以产生一个强有力的哈希基线。我们的方法结合这些技术与来自最先进的视觉和音频编码器的冻结嵌入，从而在没有任何额外学习或微调的情况下获得具有竞争力的检索性能。为了展示这种方法的通用性和有效性，我们在标准图像检索基准测试以及新引入的音频哈希基准测试上对其进行评估。[1]

Index Terms— 哈希基线，图像检索，音频检索，二进制码，预训练编码器。

1. 介绍

快速准确地使用二进制嵌入进行检索对于大规模搜索至关重要。传统的哈希方法依赖于手工制作的描述符来生成紧凑的二进制代码，而更近一些的深度哈希技术——无论是监督学习还是无监督学习——通常需要从头开始训练模型。这个训练过程通常计算成本高昂且耗时 [1]。此外，这些方法一般缺乏灵活性，因为它们必须为每个特定场景重新训练，比如不同的代码长度或不同的数据集，这限制了它们的可扩展性和泛化能力。

与此同时，基础模型的出现通过从庞大而多样的数据集中生成抽象潜在空间中的强大嵌入来彻底改变了数据表示 [2]。这些表示捕捉了丰富的语义信息，为各种下游任务提供了一个强大的起点。这自然引发了一个问题：我们能否通过直接利用这些预训练的嵌入来重新思考哈希，而不是投入到昂贵、特定场景的哈希网络训练中？

为了解决这个问题，我们引入了哈希基线，这是一种无需训练的方法，重新审视了经典的哈希技术，即主成分分析（PCA）、随机正交投影 [3] 以及通过阈值进行二值化。当这些技术应用于预训练编码器的嵌入时，它们的结合始终能产生具有竞争力的，并且有时是前沿的表现，所有这些都是在无需进一步学习的情况下实现的。

受音频检索日益重要性的驱动，我们也建立了首个专门针对音频哈希的综合基准。与分类任务相比——现代预训练音频模型通常能实现非常高的准确率 [4]，使得进一步进步变得困难——检索提供了对音频理解更具挑战性和区分度的测试。即使相关性是在类别级别定义的，成功的检索也需要将所有相关的项目排在不相关的项目之前，这对嵌入空间结构施加了比分类更严格的约束，而分类只需要最顶上的标签正确即可。我们的基准涵盖了多种音频类型，包括音乐流派、语音情感、人类发声和环境声音，连同哈希基线一起，为音频哈希提供了一个全面且具有挑战性的测试平台。

*Equal contributions.

[1]代码发布于：https://github.com/ilyassmoummad/hashing-baseline

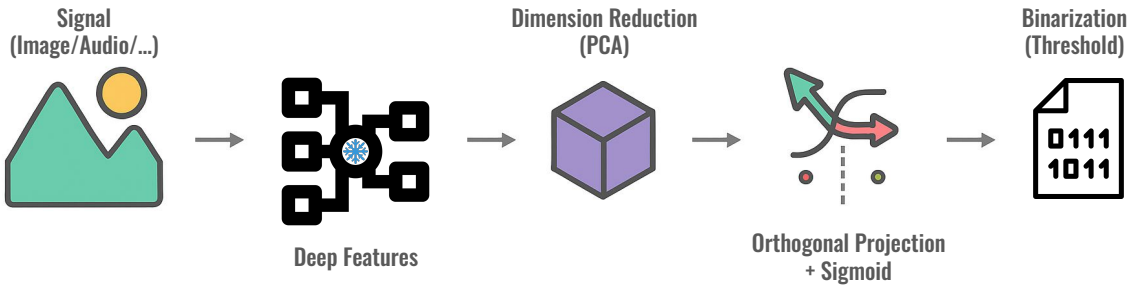

Fig. 1. 哈希基线概述：使用预训练的固定模型提取特征，然后通过 PCA 将其减少到目标比特长度。将减少后的特征正交投影，并使用 Sigmoid 函数后跟一个阈值进行二值化以生成紧凑的二进制码。

总结而言，我们的贡献是两方面的：

- 我们提出了哈希基线，这是一种无需训练的强大方法，利用预训练模型的嵌入进行高效的检索；
- 我们建立了一个新的音频哈希基准，涵盖了多种音频领域（音乐类型、语音情感、人类发声和环境声音）。

2. 相关工作

预训练模型已经通过提供强大且预训练的表示显著推进了深度学习，这些表示适用于广泛的任务 [5, 6]。经过大量多样化的数据集训练，这些模型捕获丰富的语义信息，并在众多下游应用中表现出卓越性能 [2]。它们的出现减少了对特定任务训练的依赖，使得可以在冻结嵌入的基础上设计更高效且可扩展的解决方案。

哈希 通过将数据编码为紧凑的二进制表示形式，使大规模检索系统中的相似性搜索变得快速。传统的哈希方法应用预定义或基于数据驱动的转换——例如随机投影或频谱分析——来在高维空间中保持相似性。虽然这些方法是高效的，但它们难以捕捉从原始数据 [7] 中复杂的语义关系。基于深度学习的哈希方法则使用神经网络来学习依赖于数据的哈希函数。监督变体使用标记数据来保持语义相似性，而无监督变体则依赖于内在的数据特性 [1]。尽管深度哈希通常比传统方法产生更具辨别力的代码，但它通常需要针对特定数据集和代码长度进行计算成本高昂的训练，这限制了它的泛化能力和可扩展性。

我们提出的方法，哈希基线，通过利用预训练模型的强大表示能力和传统的无训练哈希技术来连接这两条研究路线。与传统哈希不同，哈希基线操作的是丰富的预训练嵌入而不是原始特征，并且与深度哈希相比，它不需要额外的训练。这种结合产生了一种简单、可扩展且令人惊讶地具有竞争力的基线方法，适用于图像和音频检索任务。

3. 哈希基线方法

我们用 $\boldsymbol{s} \in \mathbb{R}^T$ 表示来自任何模态的信号（例如，一张图像或一段音频）。一个预训练的编码器 $f_\phi(\cdot)$ 将 $\boldsymbol{s}$ 映射到一个 d 维特征向量：

$$\boldsymbol{x} = f_\phi(\boldsymbol{s}) \in \mathbb{R}^d. \tag{1}$$

哈希基线然后由三个简单的步骤组成：(i) 使用 PCA 进行降维，(ii) 随机正交投影，以及 (iii) 使用非对称汉明检索进行二值化。

3.1. 通过 PCA 进行降维

令 $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n \in \mathbb{R}^{n \times d}$ 表示一组标准化嵌入的训练集。我们计算一个截断奇异值分解：

$$\boldsymbol{X} \approx \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{S})\,\boldsymbol{V}^\top, \tag{2}$$

其中 $\boldsymbol{V}$ 的列是前 k 主方向。特征 $\boldsymbol{x}$ 然后被投影到降维的 k 维空间中：

$$\boldsymbol{z} = \boldsymbol{V}^\top \boldsymbol{x} \in \mathbb{R}^k. \tag{3}$$

Fig. 2. 在 Flickr25K 上的检索示例，展示了使用 SimDINO 特征及其 16 位哈希码的最近邻。

3.2. 随机正交投影和二值化

通过采样一个高斯矩阵然后进行 QR 分解生成随机正交矩阵 $\boldsymbol{R} \in \mathbb{R}^{k\times k}$。简化向量被变换为

$$\boldsymbol{u} = \boldsymbol{R}\boldsymbol{z}. \tag{4}$$

然后应用逐元素的 sigmoid 函数以获得比特概率：

$$\boldsymbol{p} = \sigma(\boldsymbol{u}) \in [0,1]^k, \tag{5}$$

并在阈值 0.5 处对 $\boldsymbol{p}$ 进行阈值处理，生成数据库项的二进制码 $\boldsymbol{b} \in \{0,1\}^k$。

3.3. 非对称汉明检索

对于查询嵌入 $\boldsymbol{x}_q$，我们计算

$$\boldsymbol{p}_q = \sigma(\boldsymbol{R}\boldsymbol{V}^\top \boldsymbol{x}_q). \tag{6}$$

我们不二值化查询，而是使用非对称汉明检索 [8]，它通过以下方式比较 $\boldsymbol{p}_q$ 与二进制数据库代码 $\boldsymbol{b}_i$：

$$\mathrm{sim}(\boldsymbol{x}_q, \boldsymbol{b}_i) = -\sum_{j=1}^{k} |b_{i,j} - p_{q,j}|\,. \tag{7}$$

不对称汉明距离允许减少查询级别的量化损失，从而提高检索精度。

3.4. 理论动机

哈希基线是基于经典结果的：(i) 主成分分析（PCA）减少了预训练嵌入中的信息冗余和噪声，并仅保留最具有信息量的维度，同时在降维空间中保持几何结构，这由 Johnson–Lindenstrauss 引理 [9] 保证；(ii) 随机正交投影将集中方差重新分配到各个位上；(iii) 随机场次散列 [10] 将汉明距离与角相似性联系起来。

这些属性共同表明，将预训练模型嵌入与经典哈希结合可以提供强大的检索性能而无需任何额外的学习。

4. 实验

我们评估了提出的哈希基线在图像和音频检索基准上的表现，测量平均精度均值（mAP）。对于图像，我们遵循标准做法，并在 CIFAR-10 上报告 mAP@1000，在 Flickr25K、COCO 和 NUS-WIDE [11] 上报告 mAP@5000。对于音频，mAP 是通过对所有数据库项进行计算得出的。

4.1. 图像检索

我们评估了哈希基线在三个最先进的 ViT-Base 视觉编码器上的表现：DFN [13]，通过对比学习在 2B 图像文本对上训练；DINOv2 [5]，在一个包含 1.42 亿图像的精选数据集上训练；以及 SimDINOv2 [14]，这是一个 DINOv2 变体，通过在 ImageNet-1K 上使用余弦相似性和编码率正则化进行训练。对于每个模型，我们报告了使用在每个下游数据集的训练集上通过 PCA 降维得到的嵌入结果，以及原始嵌入和通过随机正交投影和符号阈值化（$d \in 16, 32, 64$）获得的 d 位二进制码的结果。

为了进一步研究哈希基线的泛化，我们在 SimDINOv2 上进行了消融实验，使用在 ImageNet-1K 上拟合的单个 PCA（全局 PCA）应用于所有数据集，并且还评估了以下效果：(i) 没有随机正交投影的全局 PCA，以及 (ii) 不使用 PCA 的随机正交投影。此设

Table 1. 图像检索使用不同的 ViT-Base 模型。*原始*：完整嵌入；*浮点数*：PCA 降维特征；*二进制*：随机投影码（平均 ± 标准差，经 10 次运行）。SOTA 表示无监督哈希的最新技术。

模型	特征	CIFAR10				FLICKR25K				COCO				NUS-WIDE			
		Orig	16	32	64	Orig	16	32	64	Orig	16	32	64	Orig	16	32	64
SOTA	Binary	–	87.6[11]	91.2[11]	92.6[11]	–	81.8[12]	83.8[11]	84.9[11]	–	76.0[11]	78.9[12]	81.6[12]	–	81.2[11]	83.2[11]	84.4[11]
DFN [13] (DFN-2B)	Float	93.3	94.6	94.4	94.22	80.7	83.7	83.9	83.6	85.3	77.1	82.3	85.3	83.2	81.9	83.1	83.2
	Binary	–	91.0±0.7	91.8±0.4	92.6±0.2	–	80.2±0.3	80.8±0.3	81.1±0.1	–	70.4±0.6	77.4±0.2	82.2±0.1	–	76.7±0.4	79.9±0.2	81.1±0.1
DINOv2 [5] (LVD-142M)	Float	95.4	95.9	96.0	95.9	76.3	77.8	78.2	77.7	88.3	81.2	86.5	88.8	79.8	76.4	78.0	78.7
	Binary	–	93.4±0.5	95.0±0.1	94.7±0.1	–	74.3±0.2	74.9±0.1	74.7±0.1	–	74.9±0.5	83.5±0.2	86.7±0.1	–	70.6±0.5	73.8±0.2	75.9±0.1
SimDINOv2 [14] (IN-1K)	Float	89.6	90.8	91.1	91.13	81.1	81.6	81.6	81.4	87.4	82.7	86.0	87.3	84.3	83.2	83.7	83.6
	Binary	–	84.4±0.4	86.4±0.4	88.0±0.2	–	77.2±0.5	78.0±0.1	78.6±0.1	–	75.5±0.4	81.8±0.3	84.9±0.1	–	78.0±0.2	80.3±0.2	81.5±0.1
	消融研究																
	全局 PCA（IN-1K）																
	Float	89.6	80.9	87.7	89.0	81.1	79.7	81.2	81.2	87.4	78.3	82.8	85.2	84.3	78.7	81.9	83.0
	Binary	–	66.0±1.7	79.4±0.8	84.0±0.5	–	75.1±0.9	77.3±0.5	78.3±0.2	–	69.2±0.9	77.2±0.3	81.8±0.2	–	72.7±0.6	77.7±0.2	80.5±0.2
	全局主成分分析无需随机正交投影																
	Binary	–	65.9	77.9	81.5	–	73.6	75.1	74.2	–	68.9	76.6	81.1	–	73.6	78.6	80.0
	随机正交投影无需主成分分析																
	Binary	–	40.7±2.8	59.3±2.7	71.9±1.4	–	61.1±1.0	65.2±1.4	69.1±0.8	–	54.9±1.2	68.0±1.0	76.3±0.5	–	55.4±2.2	65.4±1.5	73.7±0.7

Table 2. 使用表 1 中相同的协议进行音频检索。

模型	特征	GTZAN				ESC50				**声音信号**				CREMA-D			
		Orig	16	32	64	Orig	16	32	64	Orig	16	32	64	Orig	16	32	64
掌声 [15]	Float	41.2	41.2	38.2	37.4	88.1	81.4	87.3	87.7	62.7	59.3	57.0	55.7	25.1	25.1	25.0	24.9
	Binary	–	37.9±1.0	39.3±0.7	40.9±0.3	–	70.0±1.5	81.2±0.5	84.7±0.3	–	58.3±0.9	60.3±0.5	61.6±0.2	–	24.0±0.1	24.5±0.2	24.9±0.2
达森 [6]	Float	38.4	40.6	39.1	36.7	29.8	27.4	35.1	39.4	27.8	31.7	31.8	31.6	25.0	24.9	25.3	25.2
	Binary	–	35.1±1.0	36.8±0.9	37.8±0.5	–	19.0±0.6	25.3±0.4	29.3±0.5	–	26.6±0.4	27.5±0.3	28.1±0.2	–	23.9±0.2	24.6±0.1	25.0±0.1
CED [16]	Float	51.5	53.7	50.0	48.3	82.7	50.0	72.8	83.2	60.2	58.7	58.5	58.5	19.3	20.6	20.6	20.7
	Binary	–	51.4±0.8	51.3±0.5	52.4±0.2	–	64.1±2.3	79.5±0.3	82.2±0.3	–	57.3±0.7	58.8±0.5	60.2±0.3	–	19.2±0.1	19.4±0.1	19.5±0.1

置使我们能够分离出 PCA 和正交投影对检索性能的贡献。结果总结在表 1 中。

分析。在所有数据集上，哈希基线保留了大部分原始嵌入质量（特别是在 64 位的情况下）。令人惊讶的是，即使只有 16 位，它也在几个基准测试中达到了非常高的 mAP，而无需任何额外的学习。所获得的结果表明，强大的预训练模型特征包含大量的冗余：虽然这种冗余在预训练期间可能是有益的，但它似乎对于下游检索任务并非严格必要。最后，我们的消融研究证实了 PCA 和随机正交投影——哈希基线的两个核心组件——起着互补作用，因为移除任何一个都会显著降低性能。

4.2. 音频检索

我们引入了一个新的音频哈希基准，涵盖了音乐、环境声音和语音领域。数据集的概述见表 3。

Table 3. 音频数据集检索基准。

Dataset	Train/Database	Validation	Query	Classes
GTZAN [17]	400	—	599	10
CREMA-D [18]	5,210	1,116	1,116	6
VocalSound [19]	15,531	1,855	3,591	6
ESC-50 [20]	400	—	1,600	50

基准测试包括：GTZAN [17]（音乐类型）、ESC-50 [20]（环境声音）、CREMA-D [18]（语音情感）和 **声音** [19]（人类发声）。由于这些数据集的大小与标准图像基准相比很小，我们遵循类似于 CIFAR-10 的设置，将训练集和数据库集视为相同。

我们评估了三种最先进的编码器：CED [16]，一个知识蒸馏框架，它集合教师模型以实现高效的音频标记；**大胜** [6]，一个大规模自监督掩码音频编码器，在多样化的音频上训练用于通用分类；以及 LAION-

CLAP [15]，一个多模态对比学习模型，对齐音频和文本配对。检索结果汇总于表 2 中，遵循与第 4.1 节相同的评估协议。

分析。性能趋势与在图像设置中观察到的一致：哈希基线相对于 PCA 降维特征略有损失，这种损失在更高的位长度时会减少。**掌声**由于其广泛的多模态训练通常能取得最佳性能，而 CED 在 ESC-50 和 GTZAN 上仍具有竞争力。**大盛**表现不佳，表明通过重构进行预训练可能产生不适合检索的嵌入。

5. 结论与展望

我们引入了哈希基线，这是一种简单而有效的方法，它结合使用强大的预训练模型嵌入和经典哈希技术。通过生成紧凑的二进制代码，它可以实现快速、内存高效且低复杂度的检索，同时完全不需要训练，并在图像和音频领域都达到了具有竞争力的表现。

在二进制嵌入捕获预训练表示的全部丰富性方面仍有改进的空间。未来的工作可以探索通过参数高效微调进行轻量级哈希，使预训练模型能够以最小的计算开销适应不同的位约束。

端到端表示学习与哈希联合优化也提供了一个令人兴奋的机会。共同训练特征提取器和哈希模块可以使潜在空间更好地与量化约束对齐，从而在保持效率的同时进一步提高检索准确性。

最后，将哈希基线扩展到无训练跨模态检索（例如音频—文本、图像—文本）可以解锁可扩展的高性能多模态检索系统，在无需完整端到端训练的情况下保持紧凑二进制码的优势。

6. REFERENCES

[1] Jingdong Wang, Ting Zhang, Jingkuan Song, Nicu Sebe, and Heng Tao Shen, "A survey on learning to hash," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 769–790, 2017.

[2] Rishi Bommasani and et al., "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021.

[3] Dimitris Achlioptas, "Database-friendly random projections: Johnson-lindenstrauss with binary coins," Journal of computer and system sciences, vol. 66, no. 4, pp. 671–687, 2003.

[4] Junbo Zhang, Heinrich Dinkel, Qiong Song, Helen Wang, Yadong Niu, Si Cheng, Xiaofeng Xin, Ke Li, Wenwu Wang, Yujun Wang, et al., "The icme 2025 audio encoder capability challenge," arXiv preprint arXiv:2501.15302, 2025.

[5] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al., "Dinov2: Learning robust visual features without supervision," arXiv preprint arXiv:2304.07193, 2023.

[6] Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang, and Bin Wang, "Scaling up masked audio encoder learning for general audio classification," arXiv preprint arXiv:2406.06992, 2024.

[7] Jingdong Wang, Heng Tao Shen, Jingkuan Song, and Jianqiu Ji, "Hashing for similarity search: A survey," arXiv preprint arXiv:1408.2927, 2014.

[8] Mihir Jain, Hervé Jégou, and Patrick Gros, "Asymmetric hamming embedding: taking the best of our bits for large scale image search," in proceedings of the 19th ACM international conference on multimedia, 2011, pp. 1441–1444.

[9] William B Johnson, Joram Lindenstrauss, et al., "Extensions of lipschitz mappings into a hilbert space," Contemporary mathematics, vol. 26, no. 189-206, pp. 1, 1984.

[10] Yaniv Plan and Roman Vershynin, "Dimension reduction by random hyperplane tessellations," Discrete & Computational Geometry, vol. 51, no. 2, pp. 438–461, 2014.

[11] Hu Cao, Lei Huang, Jie Nie, and Zhiqiang Wei, "Unsupervised deep hashing with fine-grained similarity-preserving contrastive learning for image retrieval," IEEE Transactions on Circuits and Systems for Video Technology, vol. 34, no. 5, pp. 4095–4108, 2023.

[12] Zeyu Ma, Siwei Wang, Xiao Luo, Zhonghui Gu, Chong Chen, Jinxing Li, Xian-Sheng Hua, and Guangming Lu, "Harr: Learning discriminative and high-quality hash codes for image retrieval," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 20, no. 5, pp. 1–23, 2024.

[13] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar, "Data filtering networks," arXiv preprint arXiv:2309.17425, 2023.

[14] Ziyang Wu, Jingyuan Zhang, Druv Pai, XuDong Wang, Chandan Singh, Jianwei Yang, Jianfeng Gao, and Yi Ma, "Simplifying dino via coding rate regularization," arXiv preprint arXiv:2502.10385, 2025.

[15] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.

[16] Heinrich Dinkel, Yongqing Wang, Zhiyong Yan, Junbo Zhang, and Yujun Wang, "Ced: Consistent ensemble distillation for audio tagging," in ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2024, pp. 291–295.

[17] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals," IEEE Transactions on speech and audio processing, vol. 10, no. 5, pp. 293–302, 2002.

[18] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," IEEE transactions on affective computing, vol. 5, no. 4, pp. 377–390, 2014.

[19] Yuan Gong, Jin Yu, and James Glass, "Vocalsound: A dataset for improving human vocal sounds recognition," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 151–155.

[20] Karol J Piczak, "Esc: Dataset for environmental sound classification," in Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015–1018.