Koo-Fu CLIP：基于Fukunaga–Koontz线性判别分析的视觉-语言模型闭式适配

Matej Suchanek、Klara Janouskova、Ondrej Vasatko、Jiri Matas

视觉识别组 网络控制论系 电气工程学院 布拉格捷克理工大学

arXiv:2602.01127v1 [cs.CV] 1 Feb 2026

摘要

诸如 *CLIP* 之类的视觉-语言模型能够提供强大的通用表示，但其原始嵌入并未针对监督式分类进行优化，通常表现出有限的类别分离能力和过高的维度。我们提出了 *Koo-Fu CLIP*，这是一种基于 ***Fu**kunaga–**Koo**ntz*线性判别分析的监督式*CLIP*适配方法，在白化嵌入空间中运行，以抑制类内变化并增强类间判别能力。由此得到的闭式线性投影重塑了*CLIP*嵌入的几何结构，在实现有效降维的同时提升了类别可分性，并为 *CLIP* 表示提供了一种轻量且高效的适配方案。

在大规模 *ImageNet* 基准测试中，在 *Koo-Fu CLIP*空间中进行最近视觉原型分类，可将 *ImageNet-1K* 上的 *Top-1* 准确率从 *75.1%* 提升至 *79.1%*；随着标签空间扩展至 *14K* 和 *21K*类，这一优势仍然稳定。该方法支持最高 *10–12*× 的大幅压缩，同时准确率几乎没有损失甚至毫无损失，从而实现高效的大规模分类和检索。

1. 引言

视觉-语言模型（VLM）通过在共享嵌入空间中学习联合图像—文本表征，使视觉与语言概念相互对齐，从根本上改变了计算机视觉，并解锁了零样本识别、开放词汇分类和跨模态检索等能力。

对比语言-图像预训练（CLIP） [19]是视觉语言模型（VLM）的典型代表，通过在大规模图像-文本对上进行对比学习训练而成。其强大的通用嵌入使其成为适用于多种视觉任务的基础模型。除了标准分类之外，CLIP嵌入还成功应用于各种问题，包括开放集识别和分布外检测 [11, 15], 领域适配 [1], 检索 [19], 以及深度伪造检测等安全相关任务 [17, 26]。

尽管用途广泛，原始 CLIP嵌入并未针对下游监督分类进行优化；用于 ImageNet-1K 或 ImageNet-21K 等固定标签空间时，往往表现出类别分离不足和维度过高的问题。人们提出了许多将 CLIP 适配到下游任务的方法，包括提示学习（例如，CoOp [30]和 CoCoOp [29]）、基于缓存的方法（例如，Tip-Adapter[28]）以及线性探测 [4, 19], ，这些方法在适配能力、计算成本和数据效率之间各有取舍。

本文提出 Koo-Fu CLIP，这是一种基于 **Fu**kunaga-**Koo**ntz 线性判别分析（LDA）的监督式 CLIP适配方法。Fukunaga-Koontz变换在白化特征空间中应用 LDA，以同时消除类内相关并最大化类间散度。由此得到一种闭式线性投影，可重塑 CLIP 嵌入的几何结构，提高类别可分性，并实现有原则的低维判别子空间。Koo-Fu CLIP 将对比表征几何、经典判别子空间学习与基础模型的轻量级适配结合起来。

在 Koo-Fu CLIP 视觉嵌入空间中进行最近视觉原型（NVP）分类，其表现显著优于使用原始嵌入。在 ImageNet-1K 上，利用 ImageNet-1K 训练样本学习变换，可将 Top-1 准确率从 75.1% 提升至 79.1%；在标准评估协议和重新标注的评估协议下，均观察到相近的提升。重要的是，随着候选类别数量增加，这些提升仍保持稳定；当标签空间扩展至 14K 和 21K类时，效果依然如此。此举引入了大量干扰项，使任务与图像检索场景十分相似。这种鲁棒性表明，Koo-Fu CLIP 能有效增强类别判别能力，而非过拟合于闭集设定。

此外，该方法还能大幅降低维度，同时几乎不损失准确率：将

嵌入维度缩小至原来的 3×（768 →256），NVP 性能的下降不到 0.5%，同时 10-最近邻分类的准确率提高 +1%。即使将维度更加激进地缩减至 10–12×（76–64 维），准确率仍与原始 CLIP空间持平或更高。在变换后的空间中，k-最近邻分类的准确率虽比 NVP 高出 1–2%，但需要存储并搜索所有训练嵌入，因此内存和计算成本约高出 1000× 。相比之下，NVP 仅使用类别原型便能取得具有竞争力的性能，这凸显了 Koo-Fu CLIP 作为高效大规模分类与检索方案的独特优势。

本文的贡献如下：

- 一种通过判别式白化对CLIP嵌入进行的新颖闭式适配，在纯视觉与零样本评估范式中均实现了稳定的准确率提升。
- 有效的降维（最高达10–12×），且准确率损失极小。
- 对正则化、降维、邻居数量、距离度量以及评估协议的影响进行了广泛的消融实验与分析，为部署提供实用指导。

2. 相关工作

所提出的Koo-Fu CLIP位于对比表示几何、监督判别式子空间学习以及基础模型高效适配的交叉点。

CLIP 及其扩展。 CLIP 等视觉-语言模型 [19] 通过大规模对比式预训练学习联合图文嵌入，从而在各种视觉识别任务中实现出色的零样本和迁移性能。后续模型（包括 SigLIP [27] 和 SigLIP2 [22], ）在保留共享嵌入空间这一核心思想的同时，改进了这一范式的训练目标和可扩展性。尽管这些模型能够生成高度通用的表征，下游适配通常仍通过线性分类器、提示优化或轻量级微调来完成，而底层嵌入几何保持不变。

由于采用对比式训练目标，CLIP 类模型会将视觉表征组织在高度结构化的几何空间中，其中语义相似度体现为角度上的接近，而类别信息则分布于多个方向。近期分析通过基于混合模型或 Fisher 风格的框架来解读此类表征 [2], ，并指出对于特定下游分类任务而言，嵌入空间中有相当一部分是冗余的 [5]。

CLIP适配与降维。 现有的大多数CLIP适配方法作用于分类器、提示词或相似度聚合 [28]层面。线性探测使用学习得到的线性分类器评估冻结表征，而基于适配器的方法，例如 Tip-Adapter [28]通过特征缓存优化决策边界。提示学习方法，包括CoOp、CoCoOp和反向提示 [18, 29, 30], ，通过优化调整CLIP的文本侧。尽管这些方法效果显著，但它们保留了原始嵌入空间，并且通常需要迭代训练。

相比之下，Koo-Fu CLIP 对视觉嵌入本身执行闭式有监督变换。通过应用线性判别分析，我们获得了紧凑的判别子空间，在提升类别可分性的同时实现有理论依据的降维，且无需反向传播或额外的推理时组件。

深度表示中的线性判别分析与判别子空间。线性判别分析作为一种有监督降维技术已有悠久历史，但其在现代深度表示中的应用相对有限。一个值得注意的计算机视觉例外是 WDiscOOD [5]，该方法利用 CLIP嵌入的白化线性判别分析子空间进行置信度估计和分布外检测。相比之下，我们证明了同样的判别结构可以直接用于表示级适配，在压缩 CLIP表征的同时显著提升分类准确率。因此，我们重新定位了线性判别分析：它不再只是辅助分析工具，而是基础模型嵌入的一种有效且可扩展的适配机制。

ImageNet ImageNet-21K [6] 是一个根据 WordNet 层级结构组织的大规模视觉识别数据集 [14], ，包含超过 21,000 个物体类别，具有较高的语义粒度和类别不平衡问题。随后，ImageNet-1K [20]从该层级结构中衍生而来，作为一个经过筛选的子集，聚焦于 1,000 个常见类别，并配备固定的训练集、验证集和测试集划分，以支持标准化基准评测。

尽管存在众所周知的局限性，ImageNet-1K 仍然是大规模图像分类和表示学习事实上的基准，并作为文献中普遍采用的参照点 [24]。为解决原始 ImageNet-1K 标注中的已知问题 [3, 12, 16, 21, 23], ，包括标签噪声、语义重叠以及图像包含多个有效标签等问题，研究者提出了大量重新标注工作。在本文中，我们遵循以往做法，采用通过汇集这些修正工作所获得的更干净的验证协议 [12]进行评测。

3. 方法

本节首先概述 Fukunaga-Koontz变换的理论背景，然后说明如何将其应用于 CLIP 模型。最后，我们概述实验中使用的基于CLIP的分类器。

3.1. 预备知识

线性判别分析 线性判别分析（线性判别分析）是一种监督统计方法，用于

在保留特征空间判别能力的同时降低其维度 [8]。它的优点包括简单高效，因为推理时无需访问训练数据，因此甚至适用于包含数百万个数据点的数据集学习。

LDA 在线性映射中，将原始特征空间映射到一个低维子空间，使各类别的均值向量尽可能分离，同时最小化每个类别内样本的方差（散度）。形式化地，给定带标签数据集 $\{(\mathbf{x}_i,\ y_i)\}_{i=1}^{N}$，其中包含样本 $\mathbf{x}_i \in \mathbb{R}^D$ 和类别标签 $y_i \in \{1,\dots,K\}$，令

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i} \mathbf{x}_i \tag{1}$$

表示全局数据均值向量，

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{i:y_i=k} \mathbf{x}_i \tag{2}$$

表示各类别均值向量，其中 N_k 是类别 k中的样本数量，

$$\mathbf{S}_k = \sum_{i:y_i=k} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \tag{3}$$

表示类别 k的条件散度矩阵，且

$$\mathbf{S}_w = \sum_{k=1}^{K} \mathbf{S}_k, \tag{4}$$

$$\mathbf{S}_b = \sum_{k=1}^{K} N_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^\top, \tag{5}$$

分别表示类内散度矩阵和类间散度矩阵。LDA 旨在找到一个投影矩阵 $\mathbf{W} \in \mathbb{R}^{D\times L}$，使目标函数

$$\frac{\det(\mathbf{W}^\top \mathbf{S}_b \mathbf{W})}{\det(\mathbf{W}^\top \mathbf{S}_w \mathbf{W})} \tag{6}$$

达到最大值。通过求解广义特征值问题，可以用闭式解得到该问题的解

$$\mathbf{S}_b\mathbf{v} = \lambda \mathbf{S}_w\mathbf{v}. \tag{7}$$

最后，为了执行到低维空间的映射，将特征值 λ_j 按降序（$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$）排列，并通过取对应于最高特征值的前 L 个特征向量 $\mathbf{v}^j$ 构成最终的投影矩阵。该矩阵定义了到 L维子空间的映射：

$$\mathbf{W}_L = [\mathbf{v}_1, \dots, \mathbf{v}_L]. \tag{8}$$

请注意，由于矩阵$\mathbf{S}_b$本身具有奇异性，产生的特征向量数量始终少于类别数量 K。

Fukunaga-Koontz 变换 线性判别分析可以通过所谓的 Fukunaga-Koontz 变换 [10] 得到改进。该方法对标准线性判别分析提出了若干改进，从而实现其推广。具体而言，它引入了所谓的白化变换，随后进行判别旋转。通过对类内和类间散布矩阵进行同时对角化，实现了这一过程，从而将类内方差归一化、增强类别分离，并克服了产生方向数量上的固有限制。

白化从对类内散度矩阵进行特征分解开始：

$$\mathbf{S}_w = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top. \tag{9}$$

由此，我们计算逆平方根矩阵

$$\mathbf{S}_w^{-1/2} = \mathbf{V}\boldsymbol{\Lambda}^{-1/2}\mathbf{V}^\top, \tag{10}$$

其中，$\boldsymbol{\Lambda}^{-1/2}$ 是由特征值（$\sqrt{\frac{1}{\lambda_i}}$）的逆平方根构成的对角矩阵。满足

$$\mathbf{S}_w^{-1/2}\mathbf{S}_w\mathbf{S}_w^{-1/2} = \mathbf{I}, \tag{11}$$

矩阵 $\mathbf{S}_w^{-1/2}$有效地执行了对类内散度矩阵的对角化，将特征空间变换到另一个空间，使类内方差变得球形。

我们注意到，为了计算逆平方根，所有特征值都必须为正。然而，在实验中，我们遇到了具有负特征值的 $\mathbf{S}_w$ 矩阵，尽管根据定义它们是半正定的。我们认为，这归因于在高维特征空间或每类样本数量有限的情况下，经验 $\mathbf{S}_w$ 通常会变得病态或奇异 [9]。作为缓解措施，我们在白化 [13]之前应用由 λ 参数化的收缩正则化：

$$\mathbf{S}'_w = \mathbf{S}_w + \lambda\mathbf{I}, \tag{12}$$

$$\mathbf{Z} := \mathbf{S}_w^{-1/2}. \tag{13}$$

参数 λ 是所提出方法的唯一超参数。

我们将白化函数 $f(\cdot)$ 正式定义为中心化向量向球化空间的投影：

$$f(\mathbf{x}) = \mathbf{Z}\mathbf{x}. \tag{14}$$

然后，在白化空间中计算类间散度矩阵（方程（5））， 白化空间定义为矩阵 $\mathbf{S}_w^{-1/2}$ 的值域。这需要将全局均值和各类别均值映射到白化空间：

$$\mathbf{S}'_b = \sum_{k=1}^{K} N_k f(\boldsymbol{\mu}_k - \boldsymbol{\mu}) f(\boldsymbol{\mu}_k - \boldsymbol{\mu})^\top. \tag{15}$$

接下来，我们对 $\mathbf{S}'_b = \mathbf{U}\boldsymbol{\Gamma}\mathbf{U}^\top$ 进行特征分解，其中 $\mathbf{U}$ 的列表示类别可分性最大的方向。与线性判别分析类似，维度

降维通过对白化类间散度矩阵$\mathbf{S}'_b$（$\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_D$）的特征值进行排序来实现并截断旋转矩阵$\mathbf{U}$，仅保留对应于前 L 个特征值的特征向量：

$$\mathbf{U}_L = [\mathbf{u}_1, \ldots, \mathbf{u}_L] \tag{16}$$

最终的投影矩阵$\mathbf{T} \in \mathbb{R}^{L \times D}$ 通过将截断后的旋转矩阵与白化变换组合得到：

$$\mathbf{T} = \mathbf{U}_L^\top \mathbf{Z}. \tag{17}$$

该投影将原始特征映射到一个经过优化的 L 维判别子空间。

3.2. Koo-Fu CLIP

我们所提出的方法**Koo-Fu CLIP**将Fukunaga-Koontz 变换应用于基于 CLIP 的模型的嵌入空间。这些模型充当图像编码器——将图像转换为紧凑的嵌入（向量），这些嵌入来自一个能够以数学方式表达视觉相似性和语义相似性的特征空间。我们方法的核心思想是在编码过程中加入一种变换，使隐式特征空间适应下游任务，并提高常见分类技术的准确率。这种变换通过监督方式，基于人工标注数据集进行学习。

整体方法可概括为以下流程：

1. 使用基于 CLIP 的模型提取训练图像的嵌入。2. 按照第 3.1 节所述计算必要的统计量。3. 执行（交叉）验证，为给定的分类方法选择最佳变换模型。例如，可以对 λ正则化参数和保留维度数进行消融实验。4. 在推理时，将提取的嵌入映射到优化特征空间（公式（17）），然后执行分类。

3.3. 分类器

我们还介绍了通常应用于基于 CLIP 的模型的高维特征空间中的分类方法。

基于k-最近邻（k-NN）的分类该算法保留包含全部训练嵌入的数据库。在推理过程中，通过检索k 个最近的训练样本，并选择其中最常见的标签来对样本进行分类。 k 的最优值由（交叉）验证集确定。这一成熟方法的局限在于必须将整个数据库存储在内存中；除非可以接受不精确的结果，否则便要以增加内存开销来换取速度。

基于最近视觉原型（NVP）的分类在推理之前，训练图像首先按类别累积为原型。这通常是每个类别中所有样本的平均嵌入向量（或加权平均）。在推理时，样本依据最近原型进行分类。由于原型数量与训练样本数量相比通常微不足道，而k-NN可能需要访问全部训练样本，因此该方法的内存效率更高。

此外，由于可以使用来自训练数据中未出现的概念的少量代表性样本构成的新原型来扩展原型空间，因此还能够更平滑地从闭集分类任务过渡到开放集分类任务。这也称为“少样本分类”。

零样本分类 该技术利用基于 CLIP 的模型的能力，通过自然语言查询检索图像。这些模型经过训练，以对比方式将图像和文本查询嵌入到共同的特征空间中——文本嵌入应与其视觉表征的嵌入相似 [19]。这消除了为给定概念收集足够数量图像实例的需要（因此称为“零样本”）；相反，只需对该概念的标签（名称）进行编码即可直接找到原型，从而实现从闭集分类任务向开放集分类任务的过渡。

然而，由于基于 CLIP 的模型通常使用从互联网上抓取的带有图像说明文字的图像进行训练，因此仅由名称组成的 [19],搜索查询（称为提示词）可能与训练分布不匹配（例如，图像更有可能被描述为“一张苹果的照片”，而不仅仅是“苹果”）。因此，我们的实验还讨论了不同提示词策略和模板的影响。

4. 实验

4.1. 设置

模型 我们选择 OpenAI 的原始预训练 CLIP ViT-L/14 模型 [19] 作为实验中的视觉语言模型，其生成 768 维嵌入。

数据集 我们使用*ImageNet-21K* [6]（2021年冬季版，缩写为 IN-21k），这是一个按照所描绘的概念对图像进行分组的大规模数据集。该版本包含超过 13,150,000 张图像，分为 19,167 个类别。

此外，我们使用 *ImageNet-1K* [20]（缩写为 IN-1k），它是 ImageNet-21K 的一个子集，包含超过 1,280,000 张图像，分为（在大多数情况下）数量均匀的 1,000 个类别。与 ImageNet-21K 不同，该子集还附带一个独立的验证集，包含 50,000 张图像。除一个类别外，IN-1k 的类别都包含在 IN-21k 中。我们还在 IN-21k 的一个专用类别子集上进行评估，该子集记为 14K，对应于 ImageNet 类别树中的 14,271 个叶节点。有关 ImageNet 类别集合及其重叠情况的概述，请参见表 1 。

缩写			描述	类别数量
21K	14K	1K	属于 1k，不属于 21k	1
			属于 1k，也属于 21k	999
			1k 之外的叶节点	13,271
			1k 之外的 1k 的祖先	757
			1k 之外的 1k 的后代	1,234
			其他	2,906
			总计：19,167 + 1	

表 1. ImageNet 类别（WordNet 同义词集 [14]），根据其与 ImageNet-1K 的关系及其在 WordNet 层级结构中的位置进行分类。前两行构成 ImageNet-1K 类别。它们与非 1K 叶节点的并集（在实验中记为 14K）包含 14,271 个类别，彼此之间不存在祖先—后代关系；对于描绘单个 ImageNet 对象的图像，这些类别集合互斥。修订版移除有问题的同义词集后，ImageNet-21K 包含约 19,000 个类别 [25]。

#	空间 $f(\cdot)$	类别。	Top-1 干净	Top-1 Val.	Top-5 干净	Top-5 Val.
1	(–)		85.82	75.08	96.84	93.42
2	(1K)	1K	**89.71**	**79.05**	**98.19**	**95.54**
3	(21K)		88.11	78.40	98.03	95.31
4	(–)		65.76	57.79	73.27	70.03
5	(1K)	14K	**68.28**	**60.52**	**74.15**	**71.39**
6	(21K)		67.96	60.09	74.08	71.23
7	(–)		50.96	42.73	80.15	71.28
8	(1K)	21K	**55.37**	**46.84**	**84.25**	**75.70**
9	(21K)		53.71	45.32	83.41	74.76

表2. **最近视觉原型分类（NVP）的准确率**，其中类别均值嵌入作为唯一原型，位于 CLIP嵌入空间中：未经修改（–）以及经 $f(\cdot)$变换；分别是在 IN-1k训练集（1K）上计算的类内散度矩阵$\mathbf{S}_{w:1K}$ 的逆平方根，以及在完整 IN-21k（21K）上计算的$\mathbf{S}_{w:21K}$的逆平方根。

报告 IN-1k 验证集和标签校正后的“Cleaner”验证集上的准确率。我们在标准的 ImageNet-1K 闭集识别任务（包含 1K 类）以及包含干扰项的设置中评估性能。有关 1K、14K 和 21K 类的详细信息，请参见表 1。

最近，人们投入了大量精力来评估并提高 IN-1k 标注的质量 [3, 12,16, 21, 23]。因此，我们也在一个经过标签校正的“更干净”的验证集子集上评估我们的方法；该子集由多个重新标注计划 [12]汇集而成，包含超过 36,000 张图像。请注意，该子集可能会因偏向较容易的样本（它排除了重新标注结果不一致的图像）以及采用更宽松的评估指标（见下一段）而夸大性能。

指标 我们的主要指标是准确率，即正确分类的样本所占比例。我们知道，这个指标可能过于简单，因为在发生以下情况时

#	空间 $f(\cdot)$	类别。	1-NN 干净	1-NN Val.	10近邻 干净	10近邻 Val.	15近邻 干净	15近邻 Val.
1	(–)		84.9	76.3	88.7	79.6	88.7	79.5
2	(1K)	1K	**86.1**	**77.6**	**89.9**	**81.0**	**90.1**	**81.0**
3	(21K)		85.7	77.3	89.6	80.7	89.8	80.7
4	(–)		59.2	51.2	68.0	58.7	68.5	59.2
5	(1K)	14K	**60.9**	**52.8**	**69.8**	**60.4**	**70.5**	**61.0**
6	(21K)		60.4	52.3	69.2	59.8	69.9	60.4
7	(–)		47.4	40.8	57.5	49.5	58.4	50.2
8	(1K)	21K	**48.8**	**42.2**	**59.0**	**50.9**	**60.2**	**51.9**
9	(21K)		48.3	41.7	58.3	50.2	59.5	51.3

表3。 **准确率：*k*-最近邻（*k*-NN）**，其中类别均值嵌入作为单个原型，位于 CLIP 嵌入空间中。符号和设置与表 2 类似。分类结果报告的是 1、10 和 15 个最近邻对应的结果，而非 Top-5准确率，且均对应于表 2 中的 Top-1 列。

空间 $f(\cdot)$	λ	同义词集 干净	同义词集 Val.	同义词集$_7$ 干净	同义词集$_7$ Val.	mod_7 干净	mod_7 Val.
(–)	—	79.60	69.79	83.42	73.50	85.53	75.38
(1K)	50	1.09	0.90	1.96	1.62	4.45	3.79
(1K)	150	11.93	10.06	26.61	22.34	45.34	38.55
(1K)	1k	71.89	62.08	80.58	70.73	83.29	73.36
(1K)	3k	78.42	68.52	83.50	73.56	85.51	75.47
(1K)	5k	79.37	69.54	83.77	73.81	85.79	75.76
(1K)	1万	**79.87**	**70.02**	**83.87**	**73.96**	**85.87**	**75.85**
(1K)	100k	79.72	69.92	83.52	73.63	82.92	72.65

表 4. 零样本分类，即文本最近原型（tNP），在经过逐步正则化白化的 CLIP空间中使用 $\mathbf{S}'_{w:1K}$。在 IN-1k Clean 和验证集上使用以下提示词评估 Top-1 准确率：同义词集 – “ {类别名称}的图像”，同义词集$_7$ – 来自 [12],的模板所生成的 7 个提示嵌入的均值 mod_7 – 7 个提示模板的均值，其中 IN-1k 类别名称替换为 OpenAI 优化后的名称 [12, 19]。结果显示，当 λ取值较低时，准确率会大幅下降；在这种情况下，NVP 在经过 Koo-Fu 变换的特征空间中表现最佳。

错误分类无法体现由此产生的错误的严重程度，也可能受到标注噪声的影响。因此，我们还报告广义 *top-k* 准确率。如果分类方法支持，我们会让分类器输出激活值最高的 k 个样本，而不只是一个。如果真实类别与检索出的任一样本相对应，则视为匹配。

与最初发布的 ImageNet-1K 不同，“更干净”的验证子集（见上文）为某些图像分配了多个标签。在这种情况下，我们采用 ReaL 准确率 [12]：只要检索结果中任一项与任一真实标签匹配，就视为判定正确

空间 $f(\cdot)$	Dim.	Top-1 干净	Top-1 Val.	Top-5 干净	Top-5 Val.
(–)	768	85.82	75.08	96.84	93.42
(1K)	768	89.71	79.05	98.19	95.54
(1K)	640	**89.72**	79.06	98.19	95.54
(1K)	512	89.69	**79.07**	**98.23**	**95.65**
(1K)	384	89.70	79.03	**98.23**	**95.65**
(1K)	256	89.53	78.80	98.14	95.49
(1K)	192	89.22	78.45	97.99	95.25
(1K)	128	88.39	77.50	97.61	94.57
(1K)	96	87.33	76.43	97.11	93.89
(1K)	76	86.18	75.30	96.64	93.17
(1K)	64	85.00	74.22	96.12	92.39
(1K)	32	76.98	66.21	92.43	87.55

表 5. **低维子空间中 NVP 的准确率。** 在 **ImageNet-1K** 验证集和“干净”子集上测得。子空间通过 Fukunaga-Koontz 变换获得，使用 $\mathbf{S}'_{w:1\mathrm{K}}$ ($\lambda = 50$) 和 $\mathbf{S}'_{b:1\mathrm{K}}$。舍弃与 $\mathbf{S}'_{b:1\mathrm{K}}$ （式（15））的最小特征值对应的特征向量对齐的维度。前两行给出了表 2第 1、2 行的全空间结果，以便于比较。

空间 $f(\cdot)$	Dim.	1-NN 干净	1-NN Val.	10近邻 干净	10近邻 Val.	15近邻 干净	15近邻 Val.
(–)	768	84.9	76.3	88.7	79.6	88.7	79.5
(1K)	768	86.1	77.6	89.9	81.0	90.1	81.0
(1K)	640	86.1	77.7	89.9	81.0	90.1	81.0
(1K)	512	86.7	78.4	90.4	81.6	90.6	81.7
(1K)	384	87.2	79.0	91.0	82.2	91.1	82.0
(1K)	256	**87.7**	**79.4**	**91.2**	**82.4**	**91.3**	**82.3**
(1K)	192	**87.7**	**79.4**	91.1	82.2	91.2	82.2
(1K)	128	87.3	78.8	90.8	81.8	91.0	81.7
(1K)	96	86.9	78.1	90.3	81.1	90.2	81.0
(1K)	76	86.2	77.2	89.7	80.5	89.6	80.2
(1K)	64	85.5	76.4	88.9	79.5	88.9	79.4
(1K)	32	79.0	69.1	83.3	73.3	83.3	73.1

表6. **低维子空间中k-NN的准确率，**针对 $k = 1$、10 和15。设置与表5 – $\mathbf{S}'_{w:1\mathrm{K}}$相同，由CLIP 1K（$\lambda = 150$）中的样本计算得到，并在**ImageNet-1K**验证集上评估，仅保留来自$\mathbf{S}'_{b:21\mathrm{K}}$的最具判别性维度。前两行展示表3 中的先前结果，以便比较。

样本。

评估 在所有实验中，我们遵循第 3.2 节概述的流程。除非另有说明，我们分别从 ImageNet-1K（记为 1K）和 ImageNet-21K（记为 21K）的训练样本中，学习 Koo-Fu变换；使用 CLIP ViT-L/14 的 768 维特征空间，并在 ImageNet-1K 的验证集划分上评估其准确率。用于计算相应矩阵的数据集（见第 3.1 节中的公式）以下标表示（例如，$\mathbf{S}_{w:1\mathrm{K}}$）。不使用Koo-Fu变换的同类实验则视为基线。评估前，所有嵌入均归一化为单位长度，相当于使用 余弦相似度 作为距离计算指标。

4.2. 结果

纯视觉空间

初始实验研究在纯视觉特征空间中对样本进行分类，即使用最近视觉原型（NVP）和 k-NN 分类器。不进行降维（第 3.1 节），所有计算均在维度相同的变换后嵌入空间中进行。

结果见表 2。我们观察到，在未经修改的 CLIP 特征空间中，对 ImageNet-1K 验证集执行最近视觉原型分类时，其基线准确率在各项指标上均因 Koo-Fu变换而稳定提升（第 1–3 行）。具体而言，使用 ImageNet-1K 样本学习的变换，将 ReaL（Top-1）准确率从 85.2% 提高至 89.7 %，并将标准评估方案下的准确率从 75.1% 提高至 79.1%。同样，Top-5准确率分别提高了 1% 以上和 2% 以上，达到 98.2% 和 95.5%，从而减少失败次数约 40%。

从ImageNet-21K样本学习Koo-Fu变换时，观察到的提升略小，Top-1分别达到88.1%和78.4%，Top-5分别达到98.0%和95.3%。由于该变换是从十倍数量的样本和二十倍数量的类别中学习得到的，其通用性自然会为换取特定子集上的性能而有所牺牲。

我们还使用更大的类别集合进行了相同的实验，即14K和21K（见第4.1 节§数据集及表1）。这些ImageNet-1K的超集模拟了存在许多干扰图像的图像检索场景。即便在这些情况下，Koo-Fu CLIP仍复现了先前观察到的提升，将基线准确率提高至多4%（第4–9行）。

表 3展示了与表 2 中类似的实验结果：将基于最近视觉原型的分类替换为基于 k近邻（k近邻，简称 k）的分类，仅报告 Top-1 准确率，同时测试不同的 k 取值。总体而言，与 NVP 相比，Koo-Fu变换带来的提升较小。事实上，当 k为 10 或 15 时，$k = 10$ 近邻在未经修改的 CLIP空间中已比 NVP 高出 3–4.5%；而当 $k = 1$ 时，两者表现相当（第 1、4 和 7 行）。在变换后的空间中，无论是 ReaL 还是标准评估，无论是否包含干扰项，10近邻和 15近邻都比 NVP 高出 4–6%。不过，这意味着分类会变得 $1000\times$更耗内存且速度更慢。

如第3.1节所述，白化变换需要进行正则化，才能得到有意义的结果。表2 和3 报告了参数 $\lambda = 150$ 在 1K 空间和 $\lambda = 1500$ 在 21K 空间中的测量结果。有关参数 λ 对

性能的影响详见消融研究（第4.3节）。

零样本分类

接下来的实验通过文本嵌入进行分类，这称为零样本分类。这种设置是对 NVP 的一种改动：用相应的文本原型替换视觉原型，这些文本原型是通过 CLIP 文本编码器对类别名称进行编码得到的。

如前所述，CLIP 类模型的训练数据特性给最优评估带来了挑战。因此，按照 [12], ，我们采用更多提示策略：

1. 同义词集 – 使用提示模板“image of {类别名称}”，替换为原始 WordNet同义词集中的类别名称后生成的嵌入 [14],2. 同义词集$_7$ – 使用 [12], 中的模板和相同类别名称生成的 7 个提示嵌入的均值；3. 修改版$_7$ – 对相同的 7 个提示模板取均值，其中 IN-1k 类别名称替换为 OpenAI 优化后的名称 [19]。

结果如表 4所示。 我们观察到所选提示策略对准确率有相当大的影响，其中mod_7 表现最佳，并且维持与 NVP 方法相当（第4.2节）。然而，最显著的现象是，分类准确率在参数取值较低时会急剧下降，参数 λ(例如，具体取决于提示策略，降至 1–5%（含 $\lambda = 50$），揭示了我们的方法可能存在模态差距。为了保持未修改空间的零样本准确率，参数 λ 必须在 5k 左右。 然而，这会导致 NVP 准确率下降 1.5%（稍后讨论，见第4.3节），这暴露了模态选择上的权衡。

正则化空间中的零样本准确率也会受到降维的显著影响（更多内容请参见第4.2节）。

降维

最后一组主要实验除了考察白化变换外，还研究了 Koo-Fu CLIP 的降维，重点关注效率与准确率之间的潜在权衡。

表5 展示了一项实验的结果：我们反复应用从 CLIP 1K 学得的 Koo-Fu 变换（$\lambda = 150$），并在 ImageNet-1K 上进行 NVP 分类，同时逐步减少目标嵌入空间中保留的维度数量。我们观察到，维度数降至 256 之前，准确率仅略有下降（不到 0.5%）（即，3×降维）；如果仅去除最不显著的 20–30% 方向，准确率还会略有提升。保留 256 至 76 个维度的嵌入可在效率与准确率之间取得权衡。请注意，Fukunaga–Koontz 变换空间中 76 个维度下的 Top-1 准确率仍优于原始10× 较大的 CLIP 空间。

表 6展示了类似实验的结果，其中我们采用k近邻分类。观察到的结果更加理想，降幅达三分之一

空间 $f(\cdot)$	维度	同义词集 干净	同义词集 Val.	同义词集$_7$ 干净	同义词集$_7$ Val.	mod_7 干净	mod_7 Val.
(–)	768	79.60	69.79	83.42	73.50	85.53	75.38
(1K)	768	11.93	10.06	26.61	22.34	45.34	38.55
(1K)	640	7.38	6.17	21.69	18.10	44.15	37.53
(1K)	512	72.37	77.11	79.03	68.99	81.27	71.11
(1K)	384	**77.11**	**67.05**	**80.46**	**70.25**	**82.30**	**72.01**
(1K)	256	76.07	66.11	79.60	69.39	81.52	71.19
(1K)	192	75.33	65.28	78.50	68.27	80.69	70.35
(1K)	128	70.51	61.02	74.28	64.49	76.98	66.82
(1K)	96	66.17	56.90	70.31	60.69	72.92	63.02
(1K)	76	60.99	52.28	65.79	56.80	68.00	58.82
(1K)	64	55.80	47.93	61.51	52.91	63.80	54.96
(1K)	32	29.81	25.63	33.94	29.31	36.17	31.35

表 7. **低维子空间中零样本分类的 Top-1 准确率**。 在 **ImageNet-1K** 验证集和“干净”上测量，其中$\mathbf{S}'_{w:1\mathrm{K}}$ 由 $\lambda = 150$正则化，丢弃与 $\mathbf{S}'_{b:1\mathrm{K}}$ 的最小特征值对应的特征向量对齐的维度。前两行重复表 4 的结果以供参考。

空间 $f(\cdot)$	Dim.	15近邻 干净	15近邻 验证集	搜索 time	加速比	内存 (MB)
(–)	768	88.7	79.5	2.47秒	—	3753.4
(1K)	768	90.1	81.0			
(1K)	640	90.1	81.0	2.37秒	1.04×	3127.8
(1K)	512	90.6	81.7	1.97秒	1.25×	2502.3
(1K)	384	91.1	82.0	1.57秒	1.57×	1876.7
(1K)	256	**91.3**	**82.3**	1.23秒	2.01×	1251.1
(1K)	192	91.2	82.2	1.00秒	2.47×	938.4
(1K)	128	91.0	81.7	0.76秒	3.24×	625.6
(1K)	96	90.2	81.0	0.63秒	3.92×	469.2
(1K)	76	89.6	80.2	0.59秒	4.20×	371.4
(1K)	64	88.9	79.4	0.55秒	4.52×	312.8
(1K)	32	83.3	73.1	0.45秒	5.49×	156.4

表8. **15近邻分类的效率与资源消耗**。 精度数据对应于表6。**搜索时间**与**内存**是在使用 FAISS [7] 的 H200 英伟达 GPU 上测量的。

采用256维投影时，15近邻在提升准确率的同时，实现了 2× 加速比。64维嵌入保持了基线准确率，并将效率提升了4.5×。

（256维）使准确率提高1%，10倍降维（76维）与未做降维相当，而12倍降维（64维）在完全不涉及白化时与基线准确率相当。这表明Koo-Fu CLIP降维具有多重益处，既能同时提升效率与准确率，也可在二者之间进行权衡。降维对k近邻其他方面的影响在消融研究（第4.3节）中讨论。

最后，表7 展示了维度为.....时的结果

空间 $f(\cdot)$	λ	Top-1 干净	Top-1 Val.	Top-5 干净	Top-5 Val.
(–)	—	85.82	75.08	96.85	93.42
(1K)	<28	—	—	—	—
(1K)	30	89.66	78.97	98.10	95.38
(1K)	50	89.65	78.98	98.14	95.43
(1K)	100	89.69	79.04	98.18	95.52
(1K)	150	**89.71**	**79.05**	98.19	95.54
(1K)	200	89.69	79.02	**98.20**	**95.55**
(1K)	250	89.67	79.03	98.19	95.54
(1K)	500	89.58	78.94	98.16	95.53
(1K)	1000	89.39	78.67	98.09	95.39
(1K)	5000	88.15	77.50	97.77	94.85
(1K)	10000	87.43	76.79	97.54	94.45

表9. $\mathbf{S}'_{w:1\mathrm{K}}$ 正则化对白化 CLIP 空间中 IN-1k NVP 准确率的影响。$\mathbf{S}_{w:1\mathrm{K}}$ 由超过 120 万张 IN-1k 训练图像累积得到，而$\mathbf{S}_{w:1\mathrm{K}}$ 对角线上的元素值域为 76 至 4318。

对于 $\lambda < 28$，scipyeigh()过程也会错误地输出负特征值（该矩阵是按对称半正定矩阵构造的）。通过正则化解决该问题后，λ 的值对准确率几乎没有影响。

空间 $f(\cdot)$	λ	Top-1 干净	Top-1 Val.	Top-5 干净	Top-5 Val.
(–)	—	85.82	75.08	96.85	93.42
(21K)	<215	—	—	—	—
(21K)	250	88.97	78.22	97.92	95.14
(21K)	500	89.00	78.29	97.96	95.19
(21K)	750	89.03	78.35	97.98	95.23
(21K)	1000	89.06	78.37	98.01	95.27
(21K)	1500	**89.11**	**78.40**	98.03	95.31
(21K)	2000	89.08	78.39	**98.06**	95.34
(21K)	2500	89.08	78.39	**98.06**	**95.35**
(21K)	3000	89.09	78.39	98.05	95.33
(21K)	5000	89.06	78.34	98.04	95.31
(21K)	10000	88.89	78.17	97.98	95.18

表10。$\mathbf{S}'_{w:21\mathrm{K}}$ 正则化对白化 CLIP 空间中 IN-1k NVP 准确率的影响。该矩阵由超过 1300 万张 IN-21k 图像累积得到。$\mathbf{S}_{w:21\mathrm{K}}$ 的对角元素值域为 924 至 39,291。

对于 $\lambda < 215$，scipy eigh() 程序也会输出负特征值。在实验中，不使用$\mathbf{S}_{w:21\mathrm{K}}$，因为$\mathbf{S}_{w:1\mathrm{K}}$ 的性能略好。

将 Koo-Fu CLIP 的维度降4.2应用于零样本分类。如第 4.2 节所示，如果从视觉类别原型切换为文本类别原型，却不进一步增大正则化系数，性能会严重受损。令人惊讶的是，降维使低 λ 时的准确率几乎恢复；降维 50%（384 维）时，准确率仅比未修改的 CLIP 空间中的基线准确率低 3–4%。我们无法对此现象作出确切解释。

4.3. 消融研究

正则化

我们讨论 λ正则化参数对 Koo-Fu CLIP 性能的影响。首先，找出最小的 λ，使得 $\mathbf{S}'_w$ （式（12））的所有特征值均为正。然后，我们逐步增大该值并测量准确率。

结果见表 9 和 10。我们分别在 1K 和 21K CLIP 空间的 ImageNet-1K 上进行验证。我们发现，下界已经能提供合理的估计，相较于表现最佳的取值，准确率仅降低约 0.1%（$\lambda = 150$ 适用于 1K，$\lambda = 1500$适用于 21K）。

一般而言，为λ选择特定取值的影响很小。只有当 λ 的取值相对较高时，准确率才会显著下降。

***k*-NN 降维**

我们还讨论了 Koo-Fu CLIP 降维对k-NN 实际应用的影响。我们的实测结果见表8。具体而言，应用针对最大值优化的降维最佳准确率可实现两倍加速比，而更大幅度的缩减在保持基线准确率的同时，可将运行时间降至 22%（内存需求也会按缩减比例相应降低）。

最近邻数量选择

在使用k-NN进行实验之前，我们搜索了最近邻的最佳数量。根据具体设置，该值位于 10–17 的值域内；详细结果见补充材料。

嵌入归一化

除了主要实验外，我们还测试了在放宽评估流程并省略嵌入归一化时 Koo-Fu CLIP 的表现，并根据欧氏距离在特征空间中选出最佳候选项。

我们的观察表明，原始嵌入的性能几乎与归一化后的嵌入持平，在大多数设置下仅相差 <1%。这进一步凸显了白化变换在我们的流程中作为特征空间正则化手段的作用。不过，我们也注意到，需要将参数 λ 设为非常大的值（数万或数十万）。

结果见补充材料。

5. 结论

我们提出了 Koo-Fu CLIP，一种简单而有效的视觉语言模型闭式适配方法，其做法是将正则化的 Fukunaga-Koontz（白化线性判别分析）变换直接应用于冻结的视觉语言模型嵌入。我们在 CLIP 这一典型视觉语言模型上展示了该方法。在不同的分类

设置中，这带来了持续的性能提升，并在准确率、维度与推理成本之间呈现出清晰的权衡。这些结果凸显了表示空间适配作为利用基础模型的一个前景广阔且尚未充分探索的方向。未来工作将聚焦于对该变换进行更深入的理论分析，并将其扩展到其他视觉语言及多模态基础模型。

参考文献

[1] Sravanti Addepalli、Ashish Ramayee Asokan、Lakshay Sharma 和 R Venkatesh Babu。利用视觉语言模型提升图像分类中的领域泛化能力。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 23922–23932 页，2024 年。1[2] Parikshit Bansal、Ali Kavis 等。通过高斯混合模型理解对比学习。载于第三十九届神经信息处理系统年会，2024 年。2[3] Lucas Beyer、Olivier J. H´enaff、Alexander Kolesnikov、Xi- aohua Zhai 和 A¨aron van den Oord。我们是否已经不再需要 ImageNet？2020 年。2、5[4] Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。用于视觉表征对比学习的简单框架。载于国际机器学习会议论文集，第 1597–1607 页。PmLR，2020 年。1[5] Yiye Chen、Yunzhi Lin、Ruinian Xu 和 Patricio A Vela。Wdiscood：通过白化线性判别分析进行分布外检测。载于《IEEE/CVF国际计算机视觉会议论文集》，第 5298–5307 页，2023 年。2[6] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。ImageNet：大型分层图像数据库。载于 2009 年 IEEE计算机视觉与模式识别会议论文集，第 248–255 页。Ieee，2009 年。2、4[7] Matthijs Douze、Alexandr Guzhva、Chengqi Deng、Jeff Johnson、Gergely Szilvasy、Pierre-Emmanuel Mazar´e、Maria Lomeli、Lucas Hosseini 和 Herv´e J´egou。FAISS 库。arXiv 预印本 arXiv:2401.08281，2024 年。7[8] Ronald A Fisher。多种测量在分类学问题中的应用。《优生学年鉴》，1936 年。3[9] Jerome H Friedman。正则化判别分析。《美国统计学会杂志》，1989 年。3[10] Keinosuke Fukunaga。《统计模式识别导论》。学术出版社，1990 年。3[11] Xue Jiang、Feng Liu、Zhen Fang、Hong Chen、Tongliang Liu、Feng Zheng 和 Bo Han。利用预训练视觉语言模型进行负标签引导的分布外检测。载于第十二届国际学习表征会议，2024 年。1[12] Nikita Kisel、Illia Volkov、Kateˇrina Hanzelkov´a、Klara Janouskova 和 Jiri Matas。ImageNet 的缺陷：计算机视觉领域最受青睐的数据集。载于 ICLR 2025 第四届博客文章赛道，2025 年。2、5、7[13] Olivier Ledoit 和 Michael Wolf。适用于高维协方差矩阵的良态估计量。《多元分析杂志》，2004 年。3[14] George A Miller。WordNet：英语词汇数据库。《ACM通讯》，38(11):39–41，1995 年。2、5、7[15] Atsuyuki Miyai、Qing Yu、Go Irie 和 Kiyoharu Aizawa。GL-MCM：用于零样本分布外检测的全局与局部最大概念匹配。IJCV，2025 年。1[16] Curtis G. Northcutt、Anish Athalye 和 Jonas Mueller。测试集中的普遍标签错误会破坏机器学习基准的稳定性，2021 年。2、5[17] Utkarsh Ojha、Yuheng Li 和 Yong Jae Lee。迈向能够跨生成模型泛化的通用伪造图像检测器。载于《IEEE/CVF计算机视觉与模式识别会议计算机视觉与模式识别会议，第24480–24489页，2023年。 1[18]Sarah Parisot、Yongxin Yang和Steven McDonagh。为视觉语言模型学习类别命名，2023年。 2[19] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark等。通过自然语言监督学习可迁移视觉模型。在国际机器学习会议上，第8748–8763页。PmLR，2021年。1 、 2 、 4 、 5 、 7[20] Olga Russakovsky 、 Jia Deng 、 Hao Su、Jonathan Krause、Sanjeev Satheesh、Sean Ma、Zhiheng Huang、Andrej Karpathy、Aditya Khosla、Michael Bernstein等。ImageNet大规模视觉识别挑战赛。《国际计算机视觉杂志》，115(3):211–252，2015年。 2、4[21] Vaishaal Shankar、Rebecca Roelofs、Horia Mania、Alex Fang、Benjamin Recht和Ludwig Schmidt。在ImageNet上评估机器准确率。在国际机器学习会议上，第8634–8644页。PMLR，2020年。2、5[22]Michael Tschannen、Alexey Gritsenko、Xiao Wang、Muhammad Ferjad Naeem、Ibrahim Alabdulmohsin、Nikhil Parthasarathy、Talfan Evans、Lucas Beyer、Ye Xia、Basil Mustafa等。SigLIP 2：具有改进语义理解、定位能力和密集特征的多语言视觉语言编码器。arXiv预印本arXiv:2502 14786，2025年 2 。 。[23] Dimitris Tsipras、Shibani Santurkar、Logan Engstrom、Andrew Ilyas和Aleksander Madry。从ImageNet到图像分类：对基准测试进展的语境化分析，2020年。 2、5[24] Ross Wightman。PyTorch图像模型。https : // github.com/rwightman/PyTorch-图像-模型，2019年。2[25] Kaiyu Yang、Klint Qinami、Li Fei-Fei、Jia Deng和Olga Russakovsky。迈向更公平的数据集：过滤并平衡ImageNet层级结构中人物子树的分布。https://www.image-net.org/ update-sep-17-2019.php，2019年。访问日期：2025-12-22。 5[26] Andrii Yermakov、Jan Cech和Jiri Matas。释放CLIP在通用深度伪造检测中的隐藏潜力。arXiv预印本arXiv:2503.19683，2025年。 1[27] Xiaohua Zhai、Basil Mustafa、Alexander Kolesnikov和Lucas Beyer。用于语言图像预训练的Sigmoid损失。在IEEE/CVF国际计算机视觉会议论文集上，第11975–11986页，2023年。 2[28] Renrui Zhang、Rongyao Fang、Wei Zhang、Peng Gao、Kunchang Li、Jifeng Dai、Yu Qiao和Hongsheng Li。Tip-Adapter：无需训练的CLIP-Adapter，实现更优的视觉语言建模。arXiv预印本arXiv:2111 03930，2021年 1、2 。 。[29] Kaiyang Zhou、Jingkang Yang、Chen Change Loy和Zi-wei Liu。面向视觉语言模型的条件提示学习，2022年。 1、2[30] Kaiyang Zhou、Jingkang Yang、Chen Change Loy和Zi-wei Liu。学习为视觉语言模型生成提示。《国际计算机视觉杂志》，130(9):2337–2348，2022年。 1、2

A. 附加实验

	(–)		(1K)		(21K)	
k	干净	Val.	干净	Val.	干净	Val.
1	84.89	76.31	86.07	77.63	85.73	77.31
5	88.01	79.22	89.25	80.62	88.94	80.22
9	88.67	79.63	89.84	80.97	89.52	80.64
10	88.73	**79.64**	89.88	81.00	89.58	80.64
11	**88.74**	79.62	89.96	80.97	89.61	80.67
13	88.66	79.57	90.06	**81.01**	89.69	**80.72**
15	88.69	79.45	90.07	80.95	**89.77**	80.70
16	88.68	79.39	**90.09**	80.93	89.71	80.69
17	88.66	79.36	**90.09**	80.92	89.72	80.61
19	88.67	79.22	90.07	80.84	89.72	80.54
32	88.51	78.76	90.02	80.28	89.70	80.06
64	88.04	77.89	89.63	79.61	89.39	79.32

表 11. 在未变换空间和白化 CLIP 空间中，针对不同邻居数量（k）进行的 ImageNet-1K 上k-NN 准确率消融实验。使用单个邻居（$k = 1$）时精度最低；增加 k 则能持续提升性能，并在 $k = 10$ 至 $k = 17$之间趋于稳定。值得注意的是，在 k的所有取值下，Koo-Fu CLIP 的表现始终比基线高出 1% 以上。我们将后续实验中的 $k = 15$ 固定为该值，以平衡精度与计算成本。

#	空间		Top-1		Top-5	
	$f(\cdot)$	类别	干净	Val.	干净	Val.
1	(–)		85.37	74.58	96.67	93.12
2	(1K)	1K	**89.58**	**78.92**	**98.10**	**95.34**
3	(21K)		88.91	78.18	97.84	95.02
4	(–)		65.44	57.47	73.16	69.90
5	(1K)	14K	**68.22**	**60.44**	**74.08**	**71.22**
6	(21K)		67.78	59.88	73.95	71.01
7	(–)		50.57	42.34	79.49	70.64
8	(1K)	21K	**55.12**	**46.63**	**83.87**	**75.31**
9	(21K)		53.48	45.08	83.04	74.32

表 12. **不同任务上的 NVP 分类准确率**（欧几里得距离）。我们比较了基线 CLIP 嵌入（–）与使用 1K 和 21K 统计量变换后的空间。值得注意的是，1K 空间在所有基准测试中始终取得最高准确率，使验证集基线的 Top-1 和 Top-5 准确率提升超过 4%。(1K) 使用 $\lambda = 50$k，(21K) 使用 $\lambda = 100$k。

空间		Top-1		Top-5	
$f(\cdot)$	Dim.	干净	Val.	干净	Val.
(–)	768	85.37	74.58	96.67	93.12
(1K)	768	89.03	78.16	97.94	**95.10**
(1K)	640	89.05	78.20	**97.93**	**95.10**
(1K)	512	**89.08**	**78.30**	97.92	95.09
(1K)	384	89.01	78.23	97.90	95.07
(1K)	256	88.79	77.98	97.81	94.97
(1K)	192	88.60	77.71	97.73	94.78
(1K)	128	87.63	76.73	97.56	94.28
(1K)	96	86.52	75.54	97.03	93.59
(1K)	76	85.27	74.29	96.61	92.98
(1K)	64	84.05	73.00	96.18	92.30
(1K)	32	75.16	64.12	92.23	86.90

表 13. **降维后的 NVP 分类性能**。结果采用经过 $\mathbf{S}'_{w:1\mathrm{K}}$ ($\lambda = 50000$) 变换的空间中的欧几里得距离。值得注意的是，即使降至 76 维，Koo-Fu CLIP 的表现仍优于 768 维基线（–）。在大多数配置下，该变换使 Top-5准确率保持在 95% 以上。

空间		Top-1		Top-5	
$f(\cdot)$	λ	干净	Val.	干净	Val.
(–)	—	85.37	74.58	96.67	93.12
(1K)	<9400	—	—	—	—
(1K)	10k	89.51	78.80	98.04	95.26
(1K)	25k	89.57	78.88	98.06	95.33
(1K)	50k	**89.58**	**78.92**	**98.10**	95.34
(1K)	75k	89.53	78.90	**98.10**	95.34
(1K)	100k	89.49	78.88	**98.10**	**95.37**
(1K)	150k	89.44	78.83	98.08	95.34
(1K)	200k	89.39	78.76	98.05	95.32

表14. **1K 统计量变换的正则化消融实验**。我们报告了使用欧几里得距离时准确率随 λ 的变化情况。对于 $\lambda \geq 10$k，性能非常稳定，并在约 50k 处达到峰值。低于 9.4k 的值会导致协方差矩阵条件不佳。