# FASTFILL：高效兼容的模型更新

**Florian Jaeckle***
牛津大学†

**Fartash Faghri**
苹果公司

**Ali Farhadi**
苹果公司

**Oncel Tuzel**
苹果公司

**Hadi Pouransari***
苹果公司

arXiv:2303.04766v1 [cs.CV] 8 Mar 2023

摘要

在许多检索系统中，原始高维数据（例如图像）通过学习得到的嵌入模型映射为低维特征。通过对特征进行相似度比较，系统可从图库集中检索出与给定查询数据最相似的数据。当嵌入模型更新时，它可能会生成无法与图库中使用旧模型计算得到的特征进行比较或兼容的特征。因此，需要使用新嵌入模型重新计算图库中的所有特征——这一计算成本高昂的过程称为回填。最近，有研究提出兼容表示学习方法，以避免回填。尽管这些方法取得了一定成效，但新模型的性能与其和旧模型的兼容性之间存在固有的权衡。在这项工作中，我们提出 FastFill：一种兼容模型更新流程，通过特征对齐和基于策略的部分回填，迅速提升检索性能。我们表明，以往的回填策略会导致性能下降，并展示训练目标和在线部分回填中的排序都至关重要。我们提出一种新的训练方法，利用不确定性估计对齐新旧嵌入模型之间的特征。与以往工作相比，我们在多种数据集上取得了显著提升的回填效果：ImageNet 上的平均精度均值为 (+4.4%)，Places-365 上为 (+2.7%)，VGG-Face2 上为 (+1.3%)。此外，我们证明，使用 FastFill 更新有偏模型时，只需少量部分回填，少数群体准确率差距便会迅速消失。[1]

## 1 引言

检索问题在许多实际应用中日益普及，例如人脸识别、语音识别、图像定位和物体识别。在图像检索场景中，我们有一个包含大量图像的数据集，称为图库集，其中带有预测标签；此外还有一组未知的查询图像。图像检索旨在将查询图像与图库集中相关的图像进行匹配，理想情况下匹配同一类别／身份的图像。在实践中，我们使用学习得到的嵌入模型生成的低维特征向量，而不是原始的高维图像来执行检索。

当我们获得更多或更优质的训练数据、模型架构或训练方案时，希望更新嵌入模型，以提升下游检索任务的性能。然而，即使使用相同的数据集和优化方法进行训练，并且具有相同的架构，不同的神经网络也很少能学会生成兼容的特征（Li 等人，2015年）。因此，使用新的嵌入模型计算查询特征，同时保留旧的图库特征，会因新旧嵌入模型不兼容而导致检索结果不佳（Shen 等人，2020年）。

最常见的解决方案，是将图库集中由旧模型生成的特征替换为新模型生成的特征。这一计算成本高昂的过程称为回填。在实践中，我们会离线执行回填：在后台使用新模型计算新的图库特征的同时，对于查询仍继续使用旧的图库特征和旧模型（见图1-右）。只有在完成整个图库集的更新后，我们才会使用新的图库特征。

*通讯作者：florian.jaeckle@gmail.com&mpouransari@apple.com
†在苹果公司实习期间完成。[1]代码可在此处获取：
https://github.com/apple/ml-fct。

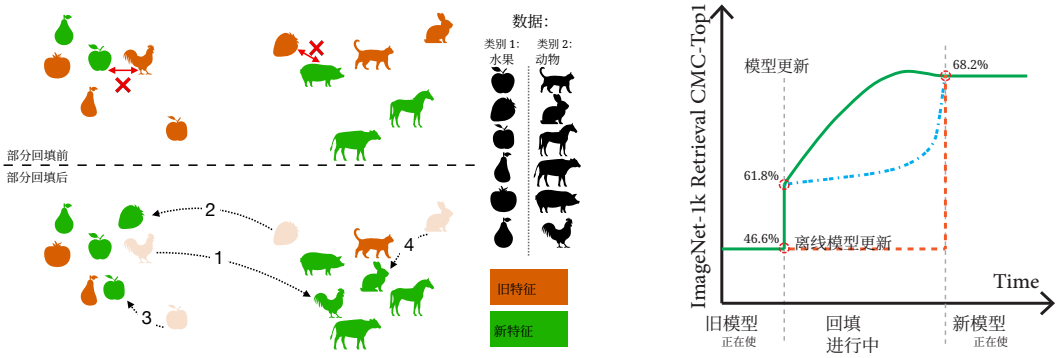



图 1：**左：** 水果与动物二分类设置下的旧特征和新特征。旧特征和新特征具有一定的兼容性，但仍存在一些不匹配之处，如顶部的红色叉号所示。我们可以按照特定顺序，将部分旧特征回填为质量更高的新特征，从而提升检索性能（底部）。在部分回填场景中，准确率的提升取决于回填顺序。**右：** 采用不同回填策略更新嵌入模型时，ImageNet 的检索性能随时间的变化。更快达到新模型性能的回填曲线更优。

然而，对于大规模系统而言，这一过程的计算成本很高，可能需要数月时间。在许多真实世界的系统中，尽管有更准确的模型可用，回填成本仍是模型更新的阻碍因素。

这促成了对兼容表示学习的研究。Shen 等人（2020）；Budnik 和 Avrithis（2021）；Zhang 等人（2021）；Ramanujan 等人（2022）；Hu 等人（2022）；Zhao 等人（2022）；Duggal 等人（2021）均提出了更新嵌入模型的方法，以在实现更高性能的同时，仍与旧模型生成的特征保持兼容（见图 1-左-上）。尽管取得了一定成功，兼容性学习仍不够完善：使用新旧特征混合进行检索时，其准确率低于将所有旧特征替换为新特征时的准确率。在本文中，我们致力于缩小这一性能差距。此外，一些以提升模型与旧模型兼容性为目标的既有方法会导致新模型性能下降（Shen 等人（2020）；Hu 等人（2022）），或者要求提供侧信息，即来自独立自监督模型的额外特征（Ramanujan 等人（2022））（而现有系统可能无法获得这些信息）。在本文中，我们放宽了这两个约束。

为了更快、更经济地受益于更准确的嵌入模型，我们可以以在线持续的方式回填图库中的部分或全部图像：我们在部分回填图库集上运行下游检索任务，其中部分特征由旧模型计算，另一部分则由新模型计算（见图1-左下）。

在实践中，我们可以考虑两种情形：1）在较长时间内以随机顺序回填整个图库集，并希望最大化回填过程中的平均性能（见图1-右）；2）我们有固定的部分回填预算，并希望在回填允许数量的图像后达到最佳检索性能（例如，我们只能回填图库中10% 的图像）。在这两种情形下，我们都希望以尽可能少的图像回填实现尽可能高的性能。我们证明，训练目标以及图像回填顺序都至关重要。事实上，如果采用文献中提出的训练损失并选择随机排序，性能甚至可能在出现提升之前先有所下降（见图2a中的 FCT-Random），这是由于出现了“负向翻转”（Zhang 等人，2021年）——即使用旧模型时分类正确、使用新模型时却被错误分类的图像。

在这项工作中，我们提出了 FastFill：一种高效的兼容表示学习与在线回填方法，可显著提升检索准确率，如图1右侧所示。在 FastFill 中，我们学习一种计算效率高的整体变换，将旧模型嵌入空间中的特征向量与新模型的嵌入空间对齐。新模型独立训练，不会为了兼容性而牺牲性能。我们提出了一种新的训练损失，其中包含点对点和点对簇目标。我们使用所提出的损失训练变换模型，并采用

不确定性估计，与以往工作相比，显著提升了部分回填性能。

我们的主要贡献总结如下：

- 我们提出了 FastFill，这是一种新型兼容训练方法，与以往工作相比，在所有部分回填程度下都能提升检索性能。
- FastFill 使用了一种新颖的训练目标，并采用一种新的策略来安排待回填样本的顺序；我们证明，这对提升性能至关重要。
- 我们展示了如何应用 FastFill，以极低的计算开销更新有偏嵌入模型，这对于高效修正大规模有偏检索系统至关重要。
- 在多种数据集上，我们证明 FastFill 达到了最先进的结果：在 ImageNet 上的平均精度均值为 +4.4%，在 Places-365 上为 +2.7%，在 VGG-Face2 上为 +1.3%。

## 2 问题设定

**图像检索** 在图像检索系统中，我们有一组称为图库集的图像 $\mathcal{G}$，它们带有预测标签，并可划分为 $k$ 不同的类别或簇。在推理时，我们还会有一个查询集 $\mathcal{Q}$。对于查询集中的每张图像，我们的目标是从图库集中检索一张属于同一类别的图像。在基于嵌入的检索系统中，我们用经训练的嵌入模型 $\phi$ 生成的特征，替换图库集和查询集中的原始$D$维数据（$3 \times h \times w$ 对于图像而言）：$\mathbb{R}^D \mapsto \mathbb{R}^d$，其中 $d \ll D$。模型 $\phi$ 是在图像 $\boldsymbol{x} \in \mathcal{D}$上训练的，而该数据集与 $\mathcal{G}$ 和 $\mathcal{Q}$均不相交。对于查询图像 $\boldsymbol{x}_q \in \mathcal{Q}$，检索算法会从图库集中返回满足以下条件的图像$\boldsymbol{x}_i$：

$$\text{(single-model retrieval)} \qquad \boldsymbol{x}_i = \arg\min_{\boldsymbol{x}_j \in \mathcal{G}} \mathfrak{D}(\phi(\boldsymbol{x}_j), \phi(\boldsymbol{x}_q)), \tag{1}$$

对于某个距离函数 $\mathfrak{D}$ （例如 $l_2$ 或余弦距离）。在这种设定下，我们只需计算一次图库集的特征 $\phi$（$\boldsymbol{x}$），并将其存储起来以供后续查询使用。此后，我们无需再访问图库中的原始图像。如 $d \ll D$所示，用特征替换图像可大幅降低维护图库和执行检索所需的内存与计算需求，这对于私有的设备端检索至关重要。根据上下文，图库可能指原始图像，也可能指对应的特征。

**学习兼容表示**。在实际应用中，我们希望每隔几个月更新一次嵌入模型，以提高其准确率：1）适应现实世界的变化，例如支持佩戴口罩的人使用人脸识别系统；2）覆盖代表性不足的子群体，例如新增族裔、不同的光照条件或不同的数据采集传感器，可通过向训练集添加更多数据实现；或 3）采用增强的架构和训练优化。我们分别用 $\phi_{old}$ 和 $\phi_{new}$ 表示旧的和新的嵌入模型，用 $\mathcal{D}_{old}$ 和$\mathcal{D}_{new}$表示这两个模型的训练数据集。模型更新后，对于每个新的查询图像，我们使用 $\phi_{new}$ 计算特征，因此检索任务转变为：

$$\text{(cross-model retrieval)} \qquad \boldsymbol{x}_i = \arg\min_{\boldsymbol{x}_j \in \mathcal{G}} \mathfrak{D}(\phi_{old}(\boldsymbol{x}_j), \phi_{new}(\boldsymbol{x}_q)). \tag{2}$$

## 3 相关工作

正如 Li 等人 (2015年) 所指出的，不同模型并不总是会生成兼容的特征，即便它们是在同一数据集上训练的。这一观察推动了相对新兴的模型兼容性研究：给定一个性能较弱的旧模型 $\phi_{old}$，我们的目标是训练一个检索性能更好的新模型 $\phi_{new}$，如式（1）所示，同时仍与旧模型兼容，如式（2）所示。这样，我们就能提升性能，而不必等待数月并承担高昂的计算成本来补全图库。Shen 等人 (2020 年 ) 提出了反向兼容训练（ BCT ）。他们在新模型的目标函数中加入影响损失项，将旧模型的线性分类器应用于新模型生成的特征，以促进兼容性。 Budnik 和 Avrithis(2021年) 将反向兼容正则化表述为对比损失，以此训练新模型。 Zhang 等人 (2021年) 提出了一种回归缓解兼容训练（RACT）方法，在对比损失中加入一个

正则化项，以减少负向翻转：即旧模型分类正确、但新模型分类错误的图像。Ramanujan 等人 (2022年) 提出了前向兼容训练（FCT）方法，该方法首先训练一个独立的新模型，随后训练一个变换函数，将特征从旧模型的嵌入空间映射到新模型的嵌入空间。此外，该变换需要输入每个数据点的一些侧信息。 Iscen 等人 (2020年)， Wang 等人 (2020年) 和 Su 等人 (2022年) 也通过学习旧模型与新模型嵌入空间之间的映射来解决模型兼容性问题。值得注意的是，正如 Ramanujan 等人 (2022年) 所讨论的，与在图像上运行嵌入模型相比，应用变换的成本可以忽略不计（因此，图 1右图中的性能会瞬间跃升）。

**兼容预测学习**。一条相关的研究路线聚焦于学习兼容预测（而非兼容特征）。对于分类任务，Srivastava 等人（2020年）表明，即使学习设置的其余部分保持不变，更改随机种子也可能导致较高的预测不兼容性。此外，他们发现，不兼容点并不总是具有较低的置信度分数。Oesterling 与 Sapiro（2021年）旨在通过确保新模型在所有训练样本上的训练损失均小于旧模型的损失，来获得预测兼容性更好的模型。Träuble 等人（2021年）考虑了为大型未标注数据集分配标签的问题。具体而言，他们关注的是：在给定新模型时，如何确定哪些点需要重新分类，以及当旧预测与新预测彼此不同时应如何处理。与兼容表示学习类似，文献中的多项工作聚焦于正一致训练，其目标是减少预测场景中的负向翻转数量。Yan 等人（2021年）通过焦点蒸馏实现这一目标，该方法更加重视旧模型分类正确的样本。Zhao 等人（2022年）认为，大多数负向翻转是由具有较大间隔的逻辑值引起的，因此尝试减小较大的逻辑值偏移。

**持续学习（CL）和知识蒸馏（KD）**。其他相关性较低、关注跨领域和跨任务学习的研究方向包括持续学习（CL）和知识蒸馏（KD）。持续学习（CL）（Parisi 等人，2019年；Li & Hoiem，2017；Zenke 等人，2017）与兼容学习有关，因为它同样旨在训练一个能够保留旧模型所获取知识的新模型。然而，在持续学习场景中，新模型要处理的是一项新任务，而此处的任务（图像检索）是固定的。此外，在持续学习中，两个模型不会交互，也不会同时使用：新模型一经训练完成，我们便会舍弃旧模型。

在知识蒸馏（KD）（Hinton 等人，2015年；Gou 等人，2021年）中，我们同样会训练一个新模型（学生模型），使其继承旧模型（教师模型）的部分能力。然而，与模型兼容性场景相比，两者有两个主要区别。首先，在知识蒸馏中，教师模型通常比学生模型更大、更强，这与模型兼容性场景恰好相反。此外，教师模型和学生模型在实践中从不同时使用。因此，在知识蒸馏场景中，两个模型的特征无需兼容。 Zhang 等人 (2021年）表明，无论是持续学习方法还是知识蒸馏方法，在模型更新场景下都无法取得良好的检索性能，因为旧模型和新模型的特征并不兼容。

## 4 方法

### 4.1 特征对齐

一些已有工作旨在通过在训练新模型时，直接强制新特征与现有旧模型特征对齐，来实现模型兼容性（Shen 等人，2020年；Budnik 和 Avrithis，2021年；Zhang 等人，2021年）。然而，由于这些方法面临着最大化新模型性能与同时使其兼容性能较弱的旧模型之间固有的权衡，因此通常会大幅降低式（1）中定义的新模型单模型检索性能。因而，其他工作提出学习一个从旧模型嵌入空间到新模型嵌入空间的独立映射，以对齐二者的特征（Wang 等人，2020年；Meng 等人，2021年；Ramanujan 等人，2022年）。Wang 等人 （2020年）和 Meng 等人 （2021年）仍会修改新模型的训练过程，这是不理想的，因为这可能限制新模型的性能增益，并且对实践者而言具有侵入性且较为复杂。Ramanujan 等人（2022年）独立训练新模型和对齐映射，但为了实现更好的对齐，依赖于侧信息的存在

（为每个数据点计算的、由无监督模型生成的额外特征）。这同样是不理想的，因为在没有可用侧信息的情况下，它不适用于已有系统。附录 B.2 中给出了存在此类侧信息时的结果。在此，我们以 Ramanujan 等人 （2022年）中的相同特征对齐作为基线，但不使用侧信息：我们学习一个映射 $h_\theta : \mathbb{R}^d \to \mathbb{R}^d$ 从旧模型特征空间映射到新模型特征空间，并使用带有参数 $\theta$ 的 MLP 架构进行建模。因此，对于跨模型检索，我们使用 $h_\theta$（$\phi_{old}$（$\boldsymbol{x}_j$），而不是 $\phi_{old}$（$\boldsymbol{x}_j$），见式（2）。

### 4.2 FASTFILL

尽管取得了一定成功，所有先前方法仍存在局限性，未能实现完全兼容：在每种情况下，式（2）中定义的跨模型检索，其性能都显著低于使用新模型进行的式（1）中定义的单模型检索。为了弥合这一差距，需要进行（部分）回填：通过使用新模型更新旧图库特征，持续地从跨模型检索过渡到单模型检索。也就是说，将变换后的旧特征$h_\theta$（$\phi_{old}$（$\boldsymbol{x}_i$））替换为新特征 $\phi_{new}$（$\boldsymbol{x}_i$），用于图库集（$\boldsymbol{x}_i \in \mathcal{G}$）中的部分或全部图像（图1）。为此，我们提出了新方法 FastFill，主要包含两项贡献：首先，通过调整对齐模型训练，改变随机策略下的回填曲线行为；其次，我们提出了一种更优策略，用于确定要更新的样本顺序，从而迅速弥合跨模型检索与单模型检索之间的差距。

我们提出了一种用于对齐映射的训练损失 $h$，使其在无回填时具有较高的检索性能（模型更新后图 1右侧所示的瞬时跃升），同时具有良好的回填曲线，从而确保我们能够迅速缩小与新模型性能之间的差距。为此，我们加入了旨在促进点对点对齐和点簇对齐的目标。

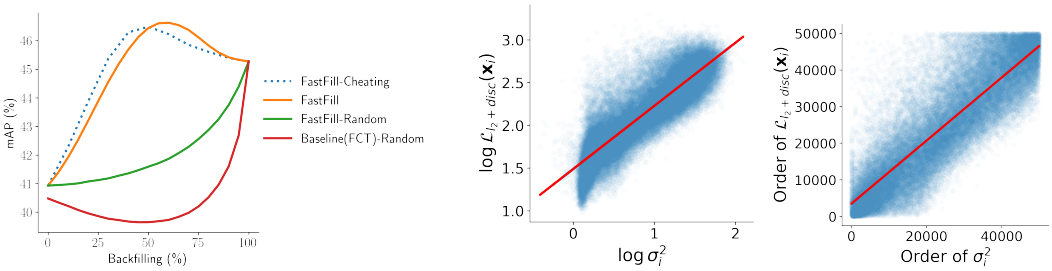
训练损失的第一部分是新特征与变换后的旧特征之间的成对平方距离：$\mathcal{L}_{l_2}(\boldsymbol{x}; h_\theta, \phi_{old}, \phi_{new}) = \|\phi_{new}(\boldsymbol{x}) - h_\theta(\phi_{old}(\boldsymbol{x}))\|_2^2$，其中 $\boldsymbol{x} \in \mathcal{D}_{new}$。这是 FCT（Ramanujan 等人，2022年）所使用的损失。此外，给定 $\phi_{old}$ 以及新模型的分类器头 $\kappa_{new}$（新模型就是基于此进行训练的），我们在变换后的特征上使用一种 $\kappa_{new}$ 运行的判别损失：$\mathcal{L}_{disc}(\boldsymbol{x}, y; h_\theta, \phi_{old}, \kappa_{new})$，其中（$\boldsymbol{x}$， $y$） $\in \mathcal{D}_{new}$ 是一个（图像、标签）对。$\mathcal{L}_{disc}$ 是新模型所使用的相同判别损失：在我们的实验中，对于 ImageNet-1k 和 Places-365数据集，它是带有标签平滑的交叉熵损失（Szegedy 等人，2016年）；对于 VGGFace2 数据集，则是 ArcFace 损失（Deng 等人，2019年）。该损失可以看作 BCT（Shen 等人，2020年）中影响损失的反向形式，但具有一项主要优势。与 BCT 不同，我们不要求旧模型和新模型在相同类别上进行训练，因为我们使用的是新模型的分类器头，而不是旧模型的分类器头。我们提出的对齐损失是这两种损失的组合：

$$\mathcal{L}_{l_2+disc}(h_\theta; \phi_{old}, \phi_{new}, \kappa_{new}, \boldsymbol{x}, y) = \mathcal{L}_{l_2} + \mathcal{L}_{disc}. \tag{3}$$

$\mathcal{L}_{l_2}$ 实现了成对兼容性，使每个变换后的旧特征与对应的新特征对齐。然而，在实践中，我们无法在查询集和图库集的未见数据上实现完美对齐。使用 $\mathcal{L}_{l_2}$，我们鼓励减小成对对齐误差$\phi_{new}(\boldsymbol{x}_i) - h_\theta(\phi_{old}(\boldsymbol{x}_i))$的幅度，但不限制误差的方向。$\mathcal{L}_{disc}$ 则鼓励误差向量朝向新模型的同类簇（由 $\kappa_{new}$给出），并远离错误簇。因此，我们避免变换后的旧特征簇靠近新特征的错误簇——这是回填改进缓慢的原因所在。

现在，我们着重寻找一种合理的排序，以便回填图库集中的图像（将$h_\theta(\phi_{old}(\boldsymbol{x}_i))$替换为 $\phi_{new}(\boldsymbol{x}_i)$）。寻找最优回填排序是一个棘手的组合问题，难以求解。我们先考察一种启发式的良好（但作弊的）排序：优先回填最差的图库特征。由于最小化对齐损失（3）能够得到良好的变换函数，我们便用它来衡量图库特征的质量：计算图库集中每张图像的训练损失（3），并按损失从高到低对图像排序。换言之，我们优先回填损失最高的图库图像，因为它们可能与新模型的兼容性较差。我们将此方法称为 FastFill-Cheating。图2a 显示，与随机排序回填相比，使用 FastFill-Cheating 可显著改善回填曲线。

遗憾的是，在实际应用中，我们无法对图库中的图像使用这种回填方法，因为在回填第 $i$张图像之前，我们无法访问 $\phi_{new}(\boldsymbol{x}_i)$。相反，我们的目标是估计这个




(a) (b) (c)

图 2：a) 在 ImageNet-1k 数据集上评估不同训练目标和回填排序的回填结果。与基线方法（FCT）相比，式（5）中定义的 FastFill 训练目标在使用随机排序时，随着回填过程的推进，平均精度均值显著提高。使用由预测不确定性所暗示的排序时，性能还会进一步提升 $\sigma_i^2$。我们还与 FastFill-Cheating 进行比较：该方法通过计算图库图像上的训练损失（3）来确定排序（因此属于作弊设置），结果表明其性能与之相当。b) 我们展示了图库图像上 $\log \sigma_i^2$（$\psi_\vartheta$ 的输出）与 $\log \mathcal{L}_{l_2+disc}(\boldsymbol{x}_i)$ 之间具有很强的相关性。c) 我们分别按 $\boldsymbol{x}_i$ 和 $\sigma_i^2$ 对 $\mathcal{L}_{l_2+disc}(\boldsymbol{x}_i)$ 排序一次，得到两种排序。这里，对于每个 $\boldsymbol{x}_i$，我们绘制其在第一种排序中的顺序与在第二种排序中的顺序，并观察到二者高度相关（Kendall-Tau 相关性=0.67）。

排序，借鉴贝叶斯深度学习的思路。当训练目标为 $\mathcal{L}_{l_2}$时，我们可以用一个多元高斯分布来建模对齐误差，其 $\sigma\mathbb{I}$ 协方差矩阵[2],并在训练过程中降低负对数似然。这与Kendall 和 Gal(2017)提出的方法类似。给定旧模型和新模型后，我们训练一个模型 $\psi_\vartheta$： $\mathbb{R}^d \mapsto \mathbb{R}$ 其参数由 $\vartheta$决定，该模型以旧模型特征 $h(\phi_{old})$ 作为输入，并估计$\log \sigma^2$。随后，我们可以通过最小化以下目标来联合训练 $h_\theta$和 $\psi_\vartheta$：

$$\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_{new}}\left[\frac{\|h_\theta(\phi_{old}(\boldsymbol{x})) - \phi_{new}(\boldsymbol{x})\|_2^2}{\sigma^2} + \frac{1}{\lambda}\log\sigma^2\right], \tag{4}$$

其中 $\log\sigma^2 = \psi_\vartheta(\phi_{old}(\boldsymbol{x}))$，并且 $\lambda$ 是 $d$的超参数函数。现在，我们将这一表述扩展到式（3）中的多目标训练。对于模型同时具有离散和连续输出的多任务设置，我们沿用 Kendall 等人（2018）在不确定性估计中的近似方法。此外，我们假设回归目标和判别目标（$\mathcal{L}_{l2}$ 和 $\mathcal{L}_{disc}$）共用相同的协方差参数 $\sigma^2$。因此，我们可以将式（ $\mathcal{L}_{l_2}$ （4）中的 $\mathcal{L}_{l_2+disc}$ 替换为此项，从而得到完整的 FastFill 训练损失：

$$\mathcal{L}(h_\theta, \psi_\vartheta; \phi_{old}, \phi_{new}, \kappa_{new}\mathcal{D}_{new}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{new}}\left[\frac{\mathcal{L}_{l_2+disc}}{\sigma^2} + \frac{1}{\lambda}\log\sigma^2\right]. \tag{5}$$

在实践中，我们对 $h_\theta$ 和 $\psi_\vartheta$使用共享骨干网络，因此额外开销极小。有关架构、超参数和训练设置的详细信息，请参阅附录 A.2。在图 2b 中，我们展示了使用式（5）训练的 $\psi_\vartheta$ 所预测的 $\sigma_i^2$是 $\boldsymbol{x}_i \in \mathcal{G}$ 的 $\mathcal{L}_{l_2+disc}(\boldsymbol{x}_i)$ 的良好近似。具体而言， $\mathcal{L}_{l_2+disc}$ 和 $\sigma$ 所隐含的排序非常接近。在图 2c 中，我们展示了图库中各个 $\boldsymbol{x}_i$ 按 $\mathcal{L}_{l_2+disc}(\boldsymbol{x}_i)$ 排序后的顺序，与按 $\sigma_i^2$排序后的顺序具有很强的相关性：在 ImageNet 测试集上，两种排序之间的 Kendall-Tau 相关性为 0.67。在 FastFill 中，我们根据预测的 $\sigma_i^2$ 所隐含的排序（从高到低），对 $\boldsymbol{x}_i$ 进行回填。

最后，在图 2a 中，我们展示了使用由低成本计算得到的 $\sigma_i^2$'s 所隐含的排序时，FastFill 获得了与作弊设置相似的回填曲线。还要注意，即使采用随机排序，FastFill 相比我们的基线方法（采用随机排序的 FCT）也取得了显著提升，这得益于我们提出的对齐损失的优越性。在第6 节中，我们进一步展示了我们提出的训练目标和回填排序的重要性。

[2]我们还考虑了使用任意对角协方差矩阵，而非缩放单位矩阵，但未观察到显著改进。

## 5 实验

现在，我们通过在多种数据集上与不同基线方法进行比较，来分析我们方法的性能。我们首先介绍用于评估和比较不同兼容性方法与回填方法的指标。

### 5.1 评估指标

**兼容性指标**与以往研究模型兼容性的工作类似，我们主要采用以下两项指标来评估模型的性能和兼容性。累计匹配特性（*CMC*）用于评估前-$k$ 检索准确率。我们计算单个查询特征与每个图库特征之间的距离，然后返回距离最小的 $k$ 个图库特征。如果返回的图库图像中至少有一张与查询图像的标签相同，我们就认为检索成功。我们还计算平均精度均值（*mAP*）分数，即召回率在 0.0 到 1.0 之间时，精确率-召回率曲线下的面积。

给定一个指标 $M$ （前-$k$ CMC 或 mAP）、图库集 $\mathcal{G}$和查询集 $\mathcal{Q}$ ，我们用 $M(\mathcal{G},\ \mathcal{Q})$表示检索性能。模型更新后，我们使用 $\phi_{new}$ 计算查询集的特征，并使用 $\phi_{old}$ 计算图库集的特征。与以往工作类似，对于检索指标，我们将每个数据集的完整验证/测试集同时用作查询集和图库集。使用图像进行查询时，我们会将该图像从图库集中移除，以避免出现简单匹配。

**回填指标** 在部分回填中，我们将图库中的部分旧特征替换为新特征，即此处的 $h(\phi_{old}(\boldsymbol{x}_i))$ 和 $\phi_{new}(\boldsymbol{x}_i)$。为了评估部分回填图库的检索性能，我们引入以下记号：给定图库集中的图像排序 $\pi$：$\boldsymbol{x}_{\pi_1}, \ldots, \boldsymbol{x}_{\pi_n}$以及回填比例 $\alpha \in [0, 1]$，定义$\mathcal{G}_{\pi,\alpha}$ 为部分回填图库集。我们使用新模型为 $\boldsymbol{x}_i \in \pi[:\ n_\alpha]$ 计算特征，并使用旧模型为其余图像计算特征，即使用 numpy 数组记法，对所有 $\boldsymbol{x}_i \in \pi[n_\alpha :]$, 均如此，其中 $n_\alpha = \lfloor \alpha n \rfloor$且 $n$为图库大小。

给定指标 $M$，为了在模型更新的整个过程中比较不同的回填和模型更新策略，我们提出以下平均量：

$$\widetilde{M}(\mathcal{G}, \mathcal{Q}, \pi) = \mathbb{E}_{\alpha\sim[0,1]} M(\mathcal{G}_{\pi,\alpha}, \mathcal{Q}) \tag{6}$$

$M$ 可以是上述定义的任意检索指标，例如 Top-$k$ CMC 或平均精度均值。回填指标$\widetilde{\ }M$ 对应于使用 $M$ 进行评估时回填曲线下的面积。

### 5.2 回填实验

现在，我们通过与三种基线方法 BCT、RACT 和 FCT 进行比较来分析我们的方法。对于 BCT 和 FCT，我们使用了作者提供的官方实现。由于撰写本文时 RACT 的代码尚未公开，我们重新实现了 RACT，并对超参数进行了优化。我们使用不同的随机种子运行了所有实验五次，并报告平均结果。我们使用$l_2$ 距离进行图像检索。

**ImageNet-1k**。 我们将 FastFill 与三种最先进的方法 BCT、RACT 和 FCT 进行比较。对于 RACT，我们采用熵回填；由于 BCT 和 FCT 均未提供回填策略，我们采用随机排序。FastFill 的表现优于这三种方法：CMC-top1 提升超过 3.5%，平均精度均值提升超过 4%（图 3，表 1）。与 BCT 和 RACT 相比，FastFill 在 0% 回填时（+10%）表现显著更好，这得益于变换函数。此外，基于 sigma 的排序显著优于随机排序和 RACT 使用的交叉熵排序。附录 D表明，与 FCT 相比，FastFill 既增加了正向翻转的数量，也减少了负向翻转的数量。

**VGGFace2**。我们在表2 和图3中展示了 VGGFace2 数据集上的结果。FastFill 优于基线方法，也优于将基于熵的回填与 FastFill 变换相结合的消融实验，CMC-top1 提高了 0.5%，平均精度均值提高了 1.3%。绝对提升相对较小，因为 0% 与 100% 回填之间的差距并不大，而变换本身已经实现了较高的兼容性。

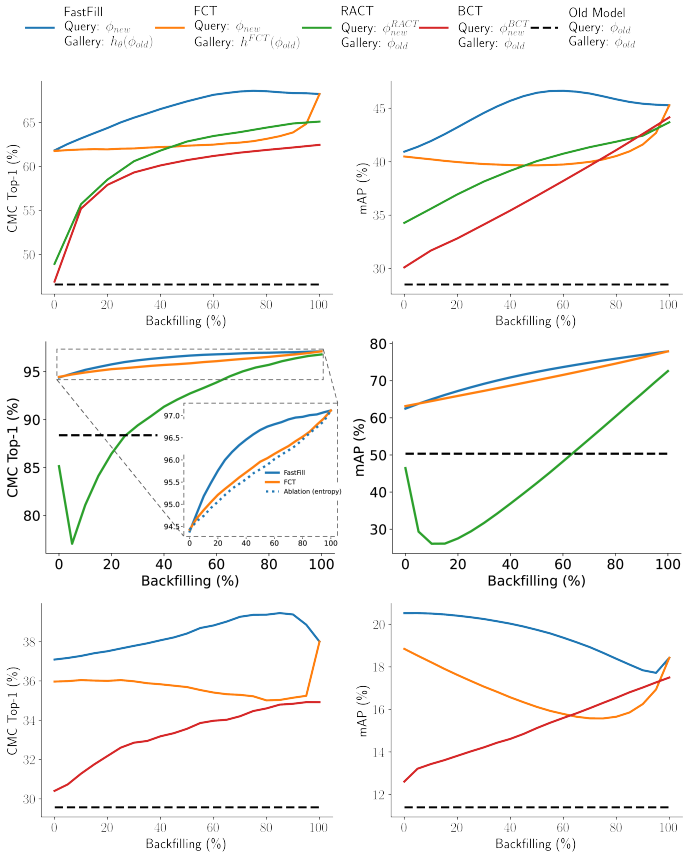



表 1：ImageNet-1k

| 方法 | $\widetilde{\text{top1}}$ | $\widetilde{\text{mAP}}$ |
|---|---|---|
| BCT | 58.92 | 37.02 |
| RACT* | 60.91 | 39.47 |
| FCT | 62.80 | 40.47 |
| FastFill | **66.49** | **44.84** |

表 2：VGGFace2

| 方法 | $\widetilde{\text{top1}}$ | $\widetilde{\text{mAP}}$ |
|---|---|---|
| RACT* | 91.04 | 45.65 |
| FCT | 95.87 | 70.26 |
| FastFill | **96.35** | **71.57** |

表3：Places-365

| 方法 | $\widetilde{\text{top1}}$ | $\widetilde{\text{mAP}}$ |
|---|---|---|
| BCT | 33.3 | 15.15 |
| FCT | 35.73 | 16.81 |
| FastFill | **38.32** | **19.48** |

图 3：在 ImageNet-1k（上）、VGGFace2（中）和 Places-365（下）上的回填结果。查询集使用新模型特征。对于图库集，我们一开始使用（经过变换的）旧特征，并逐步将其替换为新特征。

**Places-365**。我们在 Places-365 数据集上进行了实验。结果汇总于表3和图3中。FastFill 在两项指标上均比采用随机回填的 FCT 高约 2.7%。我们还在附录B.3 中报告了全部三个数据集的 CMC-top5 结果。

### 5.3 更新有偏模型

嵌入模型可能会对其输入域中的某些子群体产生偏差，这是图像检索中的一个关键问题。例如，在人脸识别任务中，模型可能会对特定族裔、性别或人脸属性产生偏差。这通常是由类别不平衡训练数据集导致的。在此，我们展示 FastFill 无需耗费大量资源进行完整回填，即可高效纠正模型偏差。我们使用 VGGFace2-Gender-550 数据集训练了一个有偏旧模型，该数据集包含 50 个女性类别和 500 个男性类别。该旧模型在测试集上对男性身份的表现（90.6%）显著优于对女性身份的表现（68.7%）。我们在一个规模更大且性别均衡的数据集上训练了新模型，其中包含 2183 个女性身份和 2183 个男性身份。新模型仍对男性类别存在轻微偏差（98.3% 对比 97.1%）。应用使用 FastFill 训练的特征对齐方法，可以缓解部分偏差，将性别间的准确率差距缩小至 9.4 %。如图4所示，仅回填 25% 的数据，FastFill

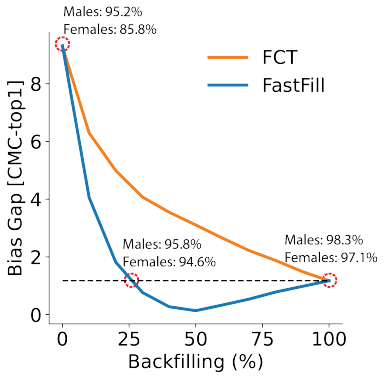



图 4

就能达到与新模型相同的性别间准确率差距；相比之下，使用 FCT 进行随机回填，则需要回填整个图库集，才能将偏差降低到新模型的水平。FastFill 优先回填与女性身份对应的实例，因此，与以往方法相比，能够更快地缩小性别差距（参见附录F）。

## 6 消融研究

这里，我们通过消融研究来突出方法的不同组成部分（图5）。首先，在采用随$\mathcal{L}_{l_2}$机 $\mathcal{L}_{disc}$回$\mathcal{L}_{l_2+disc}$ 填排序时，我们比较不同训练损失（包括有无不确定性）。结果表明， $\mathcal{L}_{l_2+disc}$ 显著改善了回填曲线。训练时引入不确定性对随机回填影响不大。我们还发现，使用 $\sigma$ 进行回填排序能显著提升性能。我们得出结论，FastFill（使用$\mathcal{L}_{l_2+disc}$ 和不确定性进行训练）在随机排序和基于 $\sigma$的排序中都取得了最佳结果。附录$C.3$ 和 $C.4$ 表明，当推理时遇到新类别时，FastFill 尤其有效，因为新类别样本之间的相对置信度（RACT 中使用）所包含的信息较少。

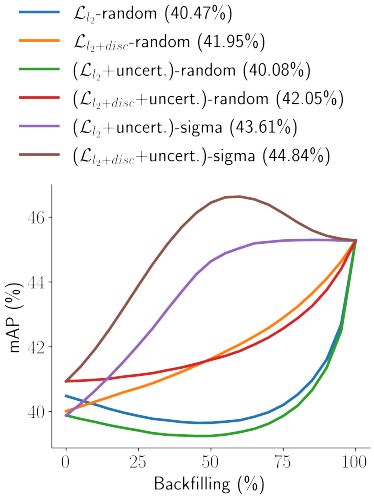



图 5：ImageNet 上的消融研究：我们比较了不同训练损失和回填排序对回填曲线的影响。图例括号中报告的是平均精度均值。

## 7 结论

模型兼容性是许多大规模检索系统中的关键问题，也是更新嵌入模型时常见的障碍。在这项工作中，我们提出了 FastFill：一种在线模型更新策略，可高效弥合以往模型兼容性研究中的准确率差距。FastFill 的核心思路是执行部分回填：使用新模型，我们重新计算大型图库集中的一小部分图像。为提高部分回填过程的效率，我们为对齐模型提出了一种新的训练目标，该模型将旧模型特征映射到新模型特征；此外，我们还提出了一种图库图像的排序方式，用于执行回填。为获得该排序，我们使用不确定性估计来训练对齐模型，并利用预测的不确定性对图库集中的图像进行排序。

我们证明，与以往工作相比，FastFill 在多个数据集上表现更优。更重要的是，我们展示了利用 FastFill 可以高效修正存在偏差的旧模型，而无需进行完整回填。

## 8 伦理声明

我们认为，FastFill 能够通过降低模型更新的相关成本，对现实世界应用产生积极影响，从而实现频繁的模型更新。

基于机器学习的模型通常并非完全无偏；当我们意识到某项局限性和/或偏差时，就需要定期进行更新。如论文所示，当我们拥有一个偏差少于现有模型的新模型时，FastFill 能够更快速、更低成本地达到新模型的公平性水平。这可以促使人们频繁更新模型，以解决与代表性不足群体相关的局限性。

与许多机器学习应用一样，尤其是用于人脸识别的应用，FastFill 也可能被滥用。人脸识别软件可能被用于许多不道德的场景，因此，提升其性能既可能带来显著的益处，也可能带来相应的负面影响。

## 9 局限性

我们提出的方法没有提供一种机制来决定需要回填多少比例的数据才能实现提前停止。了解所需的回填量可以节省额外的计算。

在一些现实应用中，下游任务可能会以比本工作所考虑的基于最近邻的检索更复杂的方式使用图像表征。例如，表征可以作为输入传递给另一个经过训练的模型。因此，从旧特征更新到新特征时，我们还需要重新训练下游模型，使其与新的特征空间兼容。不过，这并非 FastFill 特有的局限，而是基于变换的更新方法普遍存在的局限。

## 参考文献

Mateusz Budnik 和 Yannis Avrithis. 用于知识迁移的非对称度量学习. 载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 8228–8238 页，2021年。

Qiong Cao、Li Shen、Weidi Xie、Omkar M Parkhi 和 Andrew Zisserman. Vggface2：一个用于识别不同姿态和年龄人脸的数据集. 载于*2018*年第*13*届*IEEE*国际自动人脸与手势识别会议（*FG 2018*），第 67–74 页。IEEE，2018年。

Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey Hinton。视觉表征对比学习的一个简单框架。发表于国际机器学习会议，第 1597–1607 页。PMLR，2020年。

Jiankang Deng、Jia Guo、Niannan Xue 和 Stefanos Zafeiriou。Arcface：用于深度人脸识别的加性角度间隔损失。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 4690–4699 页，2019年。

Rahul Duggal、Hao Zhou、Shuo Yang、Yuanjun Xiong、Wei Xia、Zhuowen Tu 和 Stefano Soatto。兼容性感知异构视觉搜索。发表于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 10723–10732 页，2021年。

Jianping Gou、Baosheng Yu、Stephen J Maybank 和 Dacheng Tao。知识蒸馏：综述。《国际计算机视觉杂志》，129(6)：1789–1819，2021年。

Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。《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。

Geoffrey Hinton、Oriol Vinyals、Jeff Dean 等。《蒸馏神经网络中的知识》。*arXiv* 预印本 *arXiv:1503.02531*，2(7)，2015年。

Weihua Hu、Rajas Bansal、Kaidi Cao、Nikhil Rao、Karthik Subbian 和 Jure Leskovec。学习向后兼容嵌入。载于第*28*届*ACM SIGKDD*知识发现与数据挖掘大会论文集，2022年。Ahmet Iscen、Jeffrey Zhang、Svetlana Lazebnik 和 Cordelia Schmid。通过特征适应实现内存高效的增量学习。载于欧洲计算机视觉会议，第699–715页。施普林格，2020年。Alex Kendall 和 Yarin Gal。计算机视觉的贝叶斯深度学习需要哪些不确定性？神经信息处理系统进展，第30卷，2017年。Alex Kendall、Yarin Gal 和 Roberto Cipolla。利用不确定性对场景几何与语义的损失进行加权的多任务学习。载于*IEEE*计算机视觉与模式识别会议论文集，第7482–7491页，2018年。Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，2014年。Yixuan Li、Jason Yosinski、Jeff Clune、Hod Lipson 和 John Hopcroft。收敛学习：不同的神经网络会学习到相同的表征吗？载于特征提取：现代问题与挑战，第196–212页。PMLR，2015年。Zhizhong Li 和 Derek Hoiem。无遗忘学习。*IEEE*模式分析与机器智能汇刊，40(12):2935–2947，2017年。Qiang Meng、Chixiang Zhang、Xiaoqiang Xu 和 Feng Zhou。学习兼容嵌入。载于*IEEE/CVF*国际计算机视觉会议论文集，第9939–9948页，2021年。英伟达。用于 PyTorch 的 Resnet50v1.5，2019年。网址：
h
t
t
p
s://
g
i
t
h
u
b.
c
o
m/
N
V
I
D
I
A/
DeepLearningExamples/tree/master/PyTorch/Classification/ConvNets/resnet50v1.5。
N Martinez、M Bertran、A Oesterling 和 G Sapiro。分布鲁棒的群体向后兼容性。载于*NeurIPS 2021 DistShift*研讨会，2021年。German I Parisi、Ronald Kemker、Jose L Part、Christopher Kanan 和 Stefan Wermter。利用神经网络实现持续终身学习：综述。神经网络，113:54–71，2019年。Vivek Ramanujan、Pavan Kumar Anasosalu Vasu、Ali Farhadi、Oncel Tuzel 和 Hadi Pouransari。面向大规模嵌入检索系统的向前兼容训练。*IEEE*计算机视觉与模式识别会议论文集，2022年。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。《国际计算机视觉杂志》，115(3):211–252，2015年。Yantao Shen、Yuanjun Xiong、Wei Xia 和 Stefano Soatto。迈向向后兼容的表征学习。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第6368–6377页，2020年。Megha Srivastava、Besmira Nushi、Ece Kamar、Shital Shah 和 Eric Horvitz。机器学习系统中向后兼容性的实证分析。载于第*26*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集，第3272–3280页，2020年。Shupeng Su、Binjie Zhang、Yixiao Ge、Xuyuan Xu、Yexin Wang、Chun Yuan 和 Ying Shan。在图像检索中通过双向兼容训练实现保护隐私的模型升级。*arXiv* 预印本 *arXiv:2204.13919*，2022年。Christian Szegedy、Vincent Vanhoucke、Sergey Ioffe、Jon Shlens 和 Zbigniew Wojna。重新思考用于计算机视觉的 Inception架构。载于*IEEE*计算机视觉与模式识别会议论文集，第2818–2826页，2016年。

Frederik Träuble、Julius Von Kügelgen、Matthäus Kleindessner、Francesco Locatello、Bernhard Schölkopf 和 Peter Gehler。向后兼容的预测更新：一种概率方法。神经信息处理系统进展，34：116–128，2021年。

Chien-Yi Wang、Ya-Liang Chang、Shang-Ta Yang、Dong Chen 和 Shang-Hong Lai。面向跨模型兼容性的统一表征学习。*arXiv*预印本 *arXiv:2008.04821*，2020年。

Sijie Yan、Yuanjun Xiong、Kaustav Kundu、Shuo Yang、Siqi Deng、Meng Wang、Wei Xia 和 Stefano Soatto。正一致训练：迈向无回归模型更新。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 14299–14308 页，2021年。

Friedemann Zenke、Ben Poole 和 Surya Ganguli。通过突触智能实现持续学习。载于国际机器学习会议，第 3987–3995 页。PMLR，2017年。

Binjie Zhang、Yixiao Ge、Yantao Shen、Yu Li、Chun Yuan、Xuyuan Xu、Yexin Wang 和 Ying Shan。采用无回归兼容训练进行图像检索热刷新模型升级。载于国际学习表征会议，2021年。

Yue Zhao、Yantao Shen、Yuanjun Xiong、Shuo Yang、Wei Xia、Zhuowen Tu、Bernt Shiele 和 Stefano Soatto。Elodi：面向正一致训练的集成逻辑值差抑制。*arXiv* 预印本 *arXiv:2205.06265*，2022年。

Bolei Zhou、Agata Lapedriza、Jianxiong Xiao、Antonio Torralba 和 Aude Oliva。使用 Places数据库学习用于场景识别的深层特征。神经信息处理系统进展，27，2014。

1. 附录 A 介绍了正文实验所用的实验设置。我们详细说明了数据集设置（A.1）以及正文采用的训练设置（A.2）。2. 附录B报告了更多实验结果，包括：针对架构变更的实验（B.1）、使用侧信息时 FastFill 和 FCT 性能的分析（B.2），以及主要实验的前五检索结果（B.3）。3. 附录 C 重点介绍了更多消融研究。我们首先扩展正文中在 ImageNet-1k 上进行的消融实验（C.1），随后聚焦于 Places-365数据集（C.2）。然后，我们展示了本方法在推理时遇到此前未见的新类别时的一项主要优势；首先在 ImageNet 上展示这一点（C.3），随后进一步开展一项消融研究，在 VGGFace2 上将基于熵/置信度的回填与 FastFill 和 FCT 相结合（C.4）。4. 附录 D 详细分析了不同方法沿回填曲线产生的正向和负向翻转。5. 附录 E 更详细地分析了由我们的方法计算得出的 $\sigma^2$ 与各种损失函数之间的相关性。6. 附录 F 更详细地分析了 VGGFace2-Gender数据集中哪些类别最先被回填。

# A 实验设置

## A.1 数据集设置

**ImageNet-1k (21).** ImageNet-1k 是一个大规模图像识别数据集，也是 ImageNet 中使用最广泛的子集 [ILSVRC]。它曾用于 ILSVRC 2012 图像识别挑战赛。该数据集包含 1000 个物体类别，共有 1,281,167 张训练图像和 50,000 张验证图像。训练数据集相对均衡，每个类别约有 1.2k 张图像。验证集的类别与训练集相同，每个类别恰好有 500 张图像。
我们使用 ImageNet-1k 中前 500 个类别组成的子集来训练旧模型，并将其称为 ImageNet-500。值得注意的是，ImageNet-1k 的前半部分包含的类别比后半部分更容易。

**Places-365 (32).** Places-365 是一个大规模场景识别数据集，涵盖 365 个类别，包含 180 万张训练图像和 36,500 张验证图像。训练集的每个类别包含 3068 至 5000 张图像，验证集则每个类别恰好包含 100 张图像。我们使用训练集中的前 182 个类别创建一个较小的子集，并将其称为 Places-182。

**VGGFace2（2）**。VGGFace2 是一个大规模人脸识别数据集。其训练数据集由 8631 个类别组成，每个类别代表一个不同的人，共包含 314 万张图像。测试集包含 17 万张图像，对应 500 个身份，这些身份与训练集中的身份不同。训练集略有不均衡，每个类别平均有 362.6 张图像，最少 57 张，最多 843 张。测试集每个类别恰好包含 500 张图像。在测试时，我们会从测试集中每个类别随机抽取 50 张图像，并固定使用这些图像。
该数据集还为训练数据集中少量图像提供了属性列表。我们知道 30,000 张图像的性别，这些图像对应 5300 个不同身份。我们按如下方式创建性别已知图像的训练集和验证集：剔除训练集中所有性别未知的图像。随机抽取 250 个女性身份和 250 个男性身份，组成新的验证集。由于剩余男性类别多于女性类别，我们从剩余男性身份中抽取一个子集，以创建平衡的训练数据集（女性和男性类别各 2183 个，每类约有 800,000 张图像；具体数量取决于随机种子）。我们将这个新子集称为 **VGGFace2-Gender**。接着，我们从训练集中抽取 500 个男性身份和 50 个女性身份，创建一个不平衡子集，并将其命名为 VGGFace2-Gender-550。我们使用该数据集训练一个有偏旧模型。

A.2 训练设置

ImageNet-1k。我们训练两个 ResNet50 模型（7）：在 ImageNet-500 上训练旧模型（$\phi_{old}$），在 ImageNet-1k 上训练新模型（$\phi_{new}$）。我们将两个模型的嵌入维度均设为 128。我们使用 ResNetV1.5 Nvidia（17）中的超参数训练两个模型（SGD 优化器、训练轮数=100、批次大小=1024、学习率=1.024、权重衰减=$3.0517578125 * 10^{-5}$、动量=0.875，以及带 5 个训练轮次线性预热的余弦学习率衰减）。对于我们的方法，我们在 ImageNet-1k 上联合训练变换模型 $h$和不确定性预测器 $\psi$，使用上文定义的目标函数（5）。变换函数 $h$ 的 MLP 架构与（20）相同，训练设置也相同（Adam Kingma 和 Ba（13）、训练轮数=80、学习率=$5 * 10^{-4}$、带 5 个训练轮次线性预热的余弦学习率衰减，并在 40 个训练轮次后冻结 BatchNorm 层）。对于不确定性估计器 $\psi$，我们只需使用一个线性层，将变换后的嵌入空间映射为单个实数值。对于 FCT，我们严格按照其论文和官方代码库（https://github. com/apple/ml-fct）中所述，在不使用辅助信息的情况下训练其变换。对于 BCT，我们使用官方代码实现（https://github.com/YantaoShen/openBCT）训练一个新模型，并采用论文中提到的超参数。不过，为了能与其他三个模型直接比较，我们将嵌入维度设为 128。对于 RACT，由于没有发布官方代码，我们根据论文中提供的信息重新实现该方法。我们按照论文中的定义设置超参数（$\lambda = 1$ 和 $\tau = 0.5$），并通过优化学习率来获得最佳性能。我们比较 RACT 使用的学习率计划与 ResNet50V1.5 Nvidia（17）使用的学习率计划，并分别对两者优化学习率（我们尝试了 [1e-5, 1] 的 10 的幂次方学习率）。

对于检索实验，我们将图库集和查询集都设为 ImageNet-1k 的验证集。对于 BCT 和 FCT，我们使用随机回填，因为这两篇论文都未包含回填排序。对于 RACT，我们同时尝试随机回填和熵回填，并选择效果更好的一种。

**Places-365**。我们采用与 ImageNet 实验类似的设置：训练两个嵌入大小为 512 的 ResNet50 模型：旧模型使用 Places-182 进行训练，新模型使用 Places-365 进行训练。我们使用 Places-365 验证数据集进行检索评估。

**VGGFace2**。对于 VGGFace2 数据集，我们在 VGGFace2-863 上训练一个较小的 ResNet18 模型，并在 VGGFace2-8631 上训练一个较大的 ResNet50 模型。与另外两个数据集不同，我们使用 ArcFace 目标函数（4）来训练这两个模型和变换。我们采用 0.5 的间隔和 64 的缩放值，如（4）所提出。学习率计划与 ImageNet 相同。与另外两个数据集不同，我们对特征进行归一化。

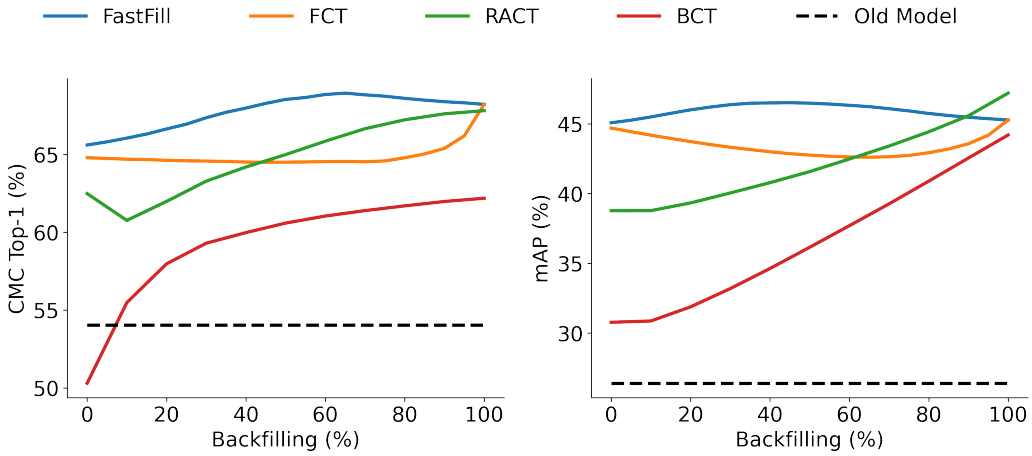
## B 更多实验结果

### B.1 架构变更实验

我们分析了旧模型与新模型架构不同时 FastFill 与现有方法的表现对比。我们在 ImageNet 上进行实验，其中旧模型是 ResNet-18，新模型是 ResNet-50。两者均在相同的完整 ImageNet-1k 训练集上训练。与之前一样，我们使用 5 个随机种子重复实验，并在表 4和图 6中展示结果。总体趋势与原论文中的其他设置相同。FastFill 的表现显著优于所有基线方法。

表 4：架构变更的 ImageNet-1k 实验：旧模型是 ResNet-18，新模型是 ResNet-50。

| 方法 | $\widetilde{\text{top1}}$ | $\widetilde{\text{mAP}}$ |
|---|---|---|
| 旧模型 | 54.04 | 26.41 |
| FastFill | **67.77** | **45.94** |
| FCT | 64.91 | 43.43 |
| RACT | 64.80 | 42.04 |
| BCT | 59.26 | 36.56 |




Figu图 6：我们在 ImageNet 上进行实验，以研究架构变更的性能。旧模型为 ResNet-18，新模型为 ResNet-50；两者均在 ImageNet-1k 上训练。

### B.2 侧信息实验

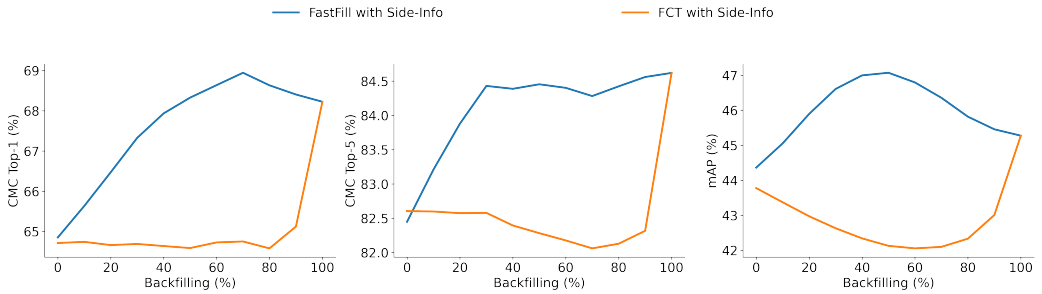



图 7：我们将侧信息添加到使用 FastFill（蓝色）或 FCT（橙色）训练的变换函数中。FastFill 在三个指标上均优于 FCT：CMC-top1（左）、CMC-top5（中）和平均精度均值（右）。

现在，我们使用侧信息训练 FCT 变换模型（$h$），以及 FastFill 变换模型和不确定性模型（$h$ 和 $\psi$）。对于侧信息模型，我们使用 SimCLR 在 ImageNet-500 上训练一个 ResNet50 模型（3）。除此之外，我们采用与论文主体中的 ImageNet 实验相同的设置。图 7 显示，两种方法在使用侧信息后性能都有所提升。但 FastFill 在三个不同指标上仍显著优于 FCT：CMC-top1（67.81% 对 64.72%）、CMC-top5（84.23% 对 82.35%）和平均精度均值（46.23% 对 42.55%）。

### B.3 CMC-TOP5

我们使用 CMC-top5 检索指标进行实验。如图 8所示，在使用 CMC-top5 评估检索性能时，FastFill 在 ImageNet-1k（+2.8%）、VGGFace2（0.1%）和 Places-365（+2.1%）上的表现均优于所有基线方法。

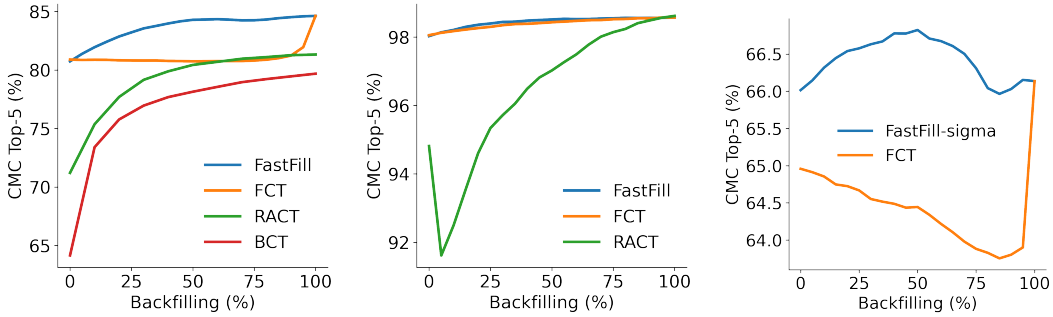



图 8：我们使用 CMC-top5 检索指标在 (a) ImageNet-1k、(b) VG-GFace2 和 (c) Places-365 上进行实验。

## C 更多消融研究

### C.1 在 ImageNet-1k 上的消融实验

我们在图 6和表 9中，在主论文第 5节所提供的消融研究基础上，加入 CMC-top1 指标进行了进一步分析。具体而言，我们展示了在使用 $\mathcal{L}_{disc}$ 训练时，无论是否考虑不确定性，在回填率为 0% 以及整个回填过程中（以平均精度均值衡量），兼容性性能都较差。不过，将其加入 $\mathcal{L}_{l_2}$ 以获得 $\mathcal{L}_{l_2+disc}$，能提升随机回填和基于 $\sigma^2$ 的回填在平均精度均值和 CMC-top1 方面的结果。

表 5：消融研究：我们比较了不同训练损失和回填排序对回填曲线的影响。

| 方法 | $\widetilde{\text{top1}}$ | $\widetilde{\text{mAP}}$ |
|---|---|---|
| ($\mathcal{L}_{l_2}$)-随机 | 62.80 | 40.47 |
| ($\mathcal{L}_{disc}$)-随机 | 65.51 | 39.45 |
| ($\mathcal{L}_{l_2+disc}$)-随机 | 65.06 | 41.95 |
| ($\mathcal{L}_{l_2}$+不确定性)-随机 | 62.58 | 40.08 |
| ($\mathcal{L}_{l_2+disc}$+不确定性)-随机 | 64.51 | 42.05 |
| ($\mathcal{L}_{l_2}$+不确定性)-sigma | 65.76 | 43.61 |
| ($\mathcal{L}_{disc}$+不确定性)-sigma | 55.21 | 27.59 |
| ($\mathcal{L}_{l_2+disc}$+不确定性)-sigma | **66.49** | **44.84** |

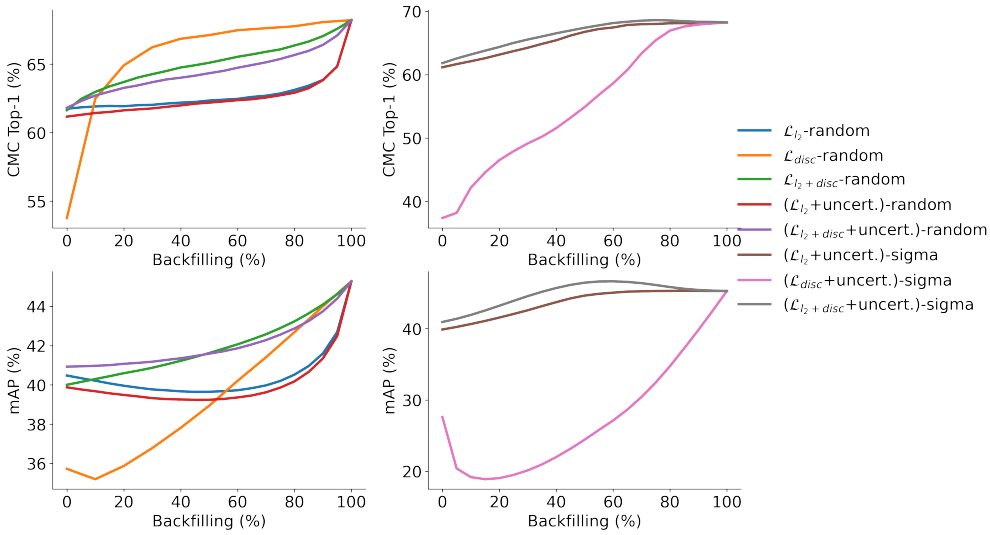



图9：消融研究：我们比较了不同训练损失和回填排序对回填曲线的影响。

### C.2 Places-365 消融实验

我们在图 10 和表6 中对 Places-365数据集进行了消融研究。结论与 ImageNet 消融实验类似：FastFill 使用替代的不确定性度量后，其消融变体表现有所提升，并且显著优于此前的 FCT。

表 6：Places-365 上的消融研究：我们比较了不同训练损失和回填排序对回填曲线的影响。

| 方法 | 变换 | 不确定性 Es-估计 | $\widetilde{\text{top1}}$ | $\widetilde{\text{mAP}}$ |
|---|---|---|---|---|
| 旧模型 | ✗ | ✗ | 29.57 | 11.39 |
| FCT | ✓ | ✗ | 37.73 | 16.81 |
| FastFill | ✓ | 贝叶斯 | **38.32** | **19.48** |
| FastFill消融 | ✓ | 熵 | 38.12 | 18.6 |

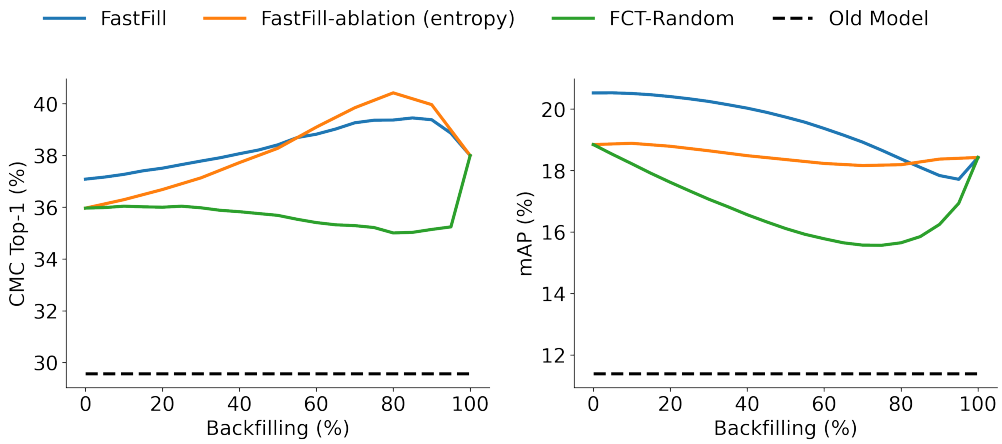



图 10：Places-365 上的消融研究：我们比较不同训练损失和回填顺序对回填曲线的影响。

### C.3 从 ImageNet-250 到 ImageNet-500 的消融实验

我们进一步进行了消融研究，比较了采用不同不确定性估计所提供的不同回填排序时的评估设置。我们开展了两类实验：一类实验中，训练和评估使用相同的类别；另一类实验中，我们评估模型在此前未见类别上的兼容性。结果如图 11 和表 7所示。在所有实验中，我们都使用 ImageNet-250 训练旧模型，并使用 ImageNet-500 训练新模型。在第一组实验（[500: ]）中，查询和图库均使用前 500 个类别。这是**训练集与测试集类别相同**的设置。在第二组实验（[500: ]）中，查询和图库均使用后 500 个类别。这是**训练集与测试集类别不相交**的设置。最后，查询和图库均使用全部 1000 个类别（[: 1000]）。在这种设置下，**训练集与测试集类别有一半**重叠。

以下结果表明，在存在新类别时，贝叶斯估计（FastFill）能进一步提升准确率。例如，与其消融版本相比，FastFill 在前 500 个类别 [:500], 上评估时，top-1 提升了 +1%；而在训练期间未见的后 500 个类别 [500:] 上评估时，top-1 提升了 +2.7%。此外，在未见类别上评估时，采用基于 logits 的不确定性估计的 FastFill 消融版本，其结果与完全不进行不确定性估计的情况（如 FCT）相近。

我们注意到，对于大多数真实世界的图像检索任务（例如人脸识别或零样本检索），测试时样本（图库集和查询集中的样本）来自训练时未出现过的新类别。像 RACT 一样，使用训练时分类头来衡量不确定性（30）在实际应用中有很大局限性，因为新类别样本之间的相对置信度或熵往往不那么有参考价值。相比之下，我们提出的基于特征对齐的不确定性估计方法专为类别无关场景设计，在此前未见过的类别上也能表现良好。这是我们方法的主要贡献和创新之一。

表 7：ImageNet 上的消融研究：我们比较了不同训练损失和回填排序对回填曲线的影响。旧模型在 ImageNet-250 上训练，新模型在 ImageNet-500 上训练。

| 方法 | 变换 | 不确定性估计 | $\widetilde{\text{top1}}$ [: 500] | $\widetilde{\text{top1}}$ [500 :] | $\widetilde{\text{top1}}$ [: 1000] | $\widetilde{\text{mAP}}$ [: 500] | $\widetilde{\text{mAP}}$ [500 :] | $\widetilde{\text{mAP}}$ [: 1000] |
|---|---|---|---|---|---|---|---|---|
| 旧模型 | ✗ | ✗ | 50.7 | 12.41 | 29.32 | 29.13 | 2.47 | 14.9 |
| FastFill | ✓ | 贝叶斯 | **75.27** | **17.9** | **43.0** | **53.88** | **4.68** | **26.44** |
| FastFill消融 | ✓ | 熵 | 74.26 | 15.22 | 40.78 | 53.48 | 4.13 | 25.74 |
| FastFill消融 | ✓ | 边际置信度 | 74.21 | 15.17 | 40.66 | 53.21 | 4.01 | 25.56 |
| FastFill消融 | ✓ | 最低置信度 | 74.28 | 15.18 | 40.74 | 53.38 | 4.07 | 25.65 |
| FCT | ✓ | ✗ | 71.15 | 14.94 | 38.77 | 48.61 | 3.7 | 23.4 |
| RACT | ✗ | 熵 | 66.88 | 13.65 | 36.65 | 41.6 | 3.3 | 20.4 |

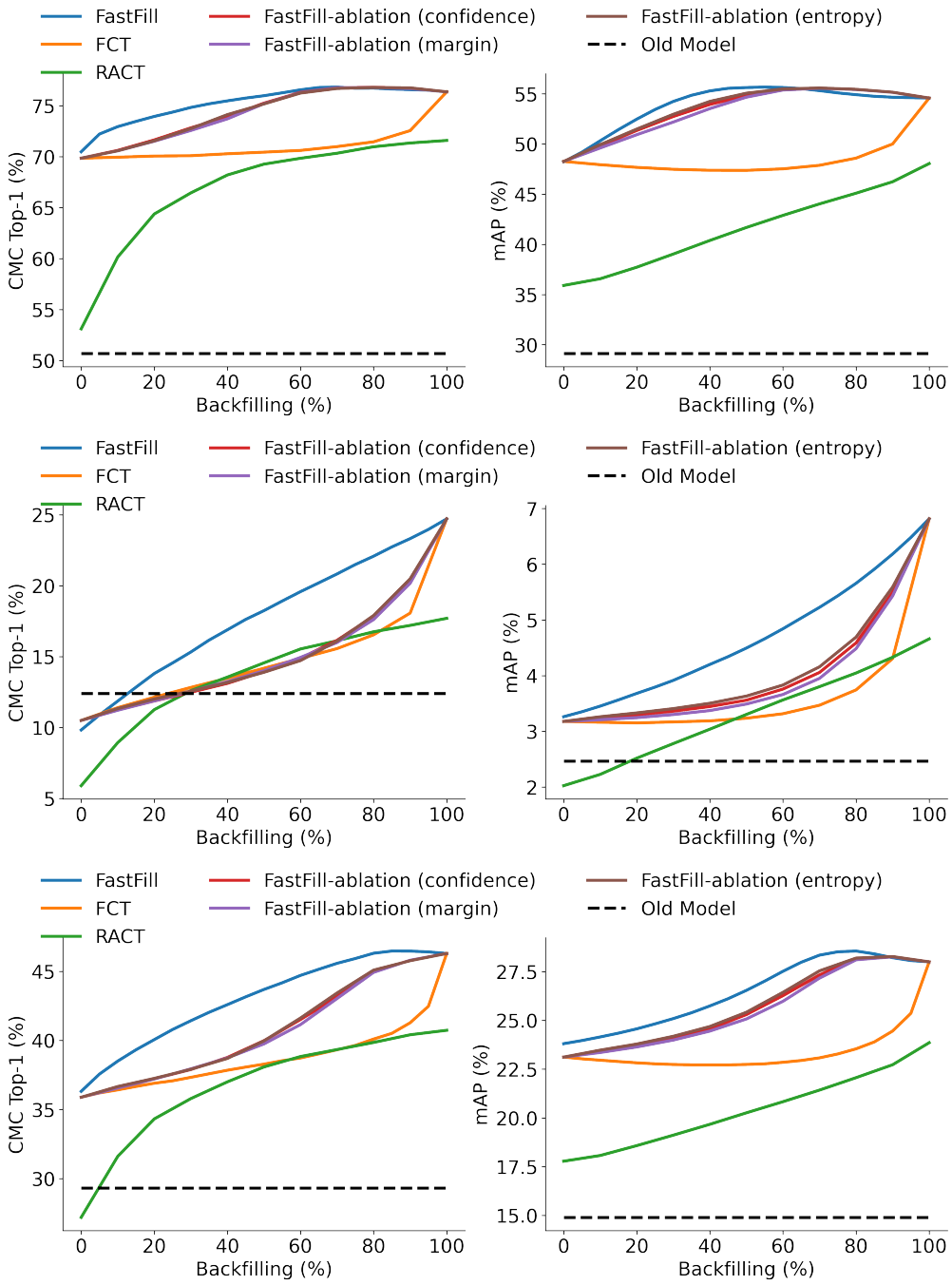



图 11：ImageNet 上的消融研究：我们比较了不同训练损失和回填排序对回填曲线的影响。**顶部：** [500: ] 查询和图库均使用前 500 个类别。这是相同的训练-测试类别设置。**中部：** [500: ] 查询和图库均使用后 500 个类别。这是不相交的训练-测试类别设置。**底部：** [: 1000]查询和图库均使用全部 1000 个类别。这种设置中，训练和测试类别有一半重叠。这些结果表明，在推理时遇到新的、此前未见过的类别时，FastFill 尤其强大；而这在许多现实生活中的检索场景中十分常见。

### C.4 VGGFACE2 上的消融实验

我们在 VGGFace2 上使用基于置信度和熵的回填方法（由 Zhang 等人提出（30）），并结合我们提出的特征对齐损失和 FCT 基线开展实验。实验结果如图12所示。

我们发现，在 FastFill 和 FCT 基线上进行基于熵/置信度的消融实验，其表现与随机回填相差无几，而我们提出的不确定性估计则显著提升了它们的性能。此外，FastFill 的变体使用了一个从旧特征空间到新特征空间的变换函数，与 RACT 直接强制新旧模型兼容相比，显著提升了模型更新的准确率（Top-1 指标提升超过 5%，平均精度均值指标提升 25%）。我们注意到，Zhang 等人（30）提出的不确定性估计指标（置信度和熵）直观且易于使用，但依赖于训练时分类头。对于大多数现实世界中的图像检索任务（例如人脸识别或零样本检索），测试时样本（来自图库集和查询集）所属的类别与训练时见过的类别不同。像 Zhang 等人（30）那样使用训练时分类头来衡量不确定性，在实践中存在重大局限，因为新类别样本之间的相对置信度/熵往往无法提供有效信息。相比之下，我们提出的基于特征对齐的不确定性估计方法专为类别无关场景设计，在此前未见过的类别上也表现良好。这是我们方法的主要贡献/创新之一。

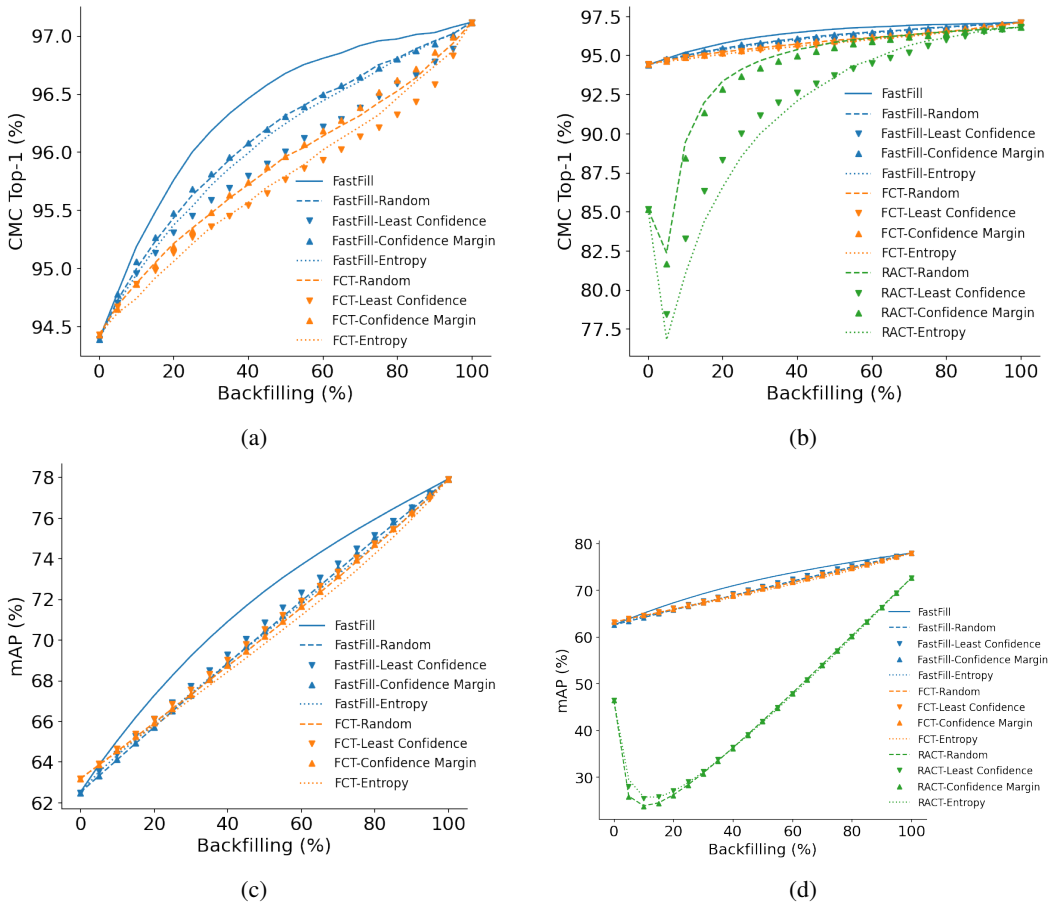



图 12：我们在 VGGFace2 上使用基于置信度和熵的回填方法（由 Zhang 等人提出（30）），并结合我们提出的特征对齐损失和 FCT 基线开展实验。我们分别使用 CMC-top1（a、b）和平均精度均值（c、d）指标比较不同方法。在（b）和（d）中，我们还与 RACT 进行了比较。

D 正向与负向翻转分析

我们分析了 FastFill 和 FCT 的正向和负向翻转。负向翻转指某张图像在回填比例为 0% 时的跨模型检索中被正确分类，但使用 $\alpha \in (0, 1]$ 的回填量后却被错误分类。相反，正向翻转指某张图像最初被错误分类，但在一定程度的（部分）回填后被正确标记。在某个回填比例下的检索性能，取决于回填比例为 0% 时的原始跨模型检索性能，再加上正向翻转数量减去负向翻转数量。我们表明，FastFill 的表现优于 FCT：它在回填量有限时增加了正向翻转数量，同时在整个回填曲线过程中造成的负向翻转更少。值得注意的是，尽管在 100% 回填时 FCT 和 FastFill 使用相同的模型，但翻转数量并不相同，因为翻转数量还取决于回填比例为 0% 时的跨模型性能，而两种方法在这一点上的性能不同。

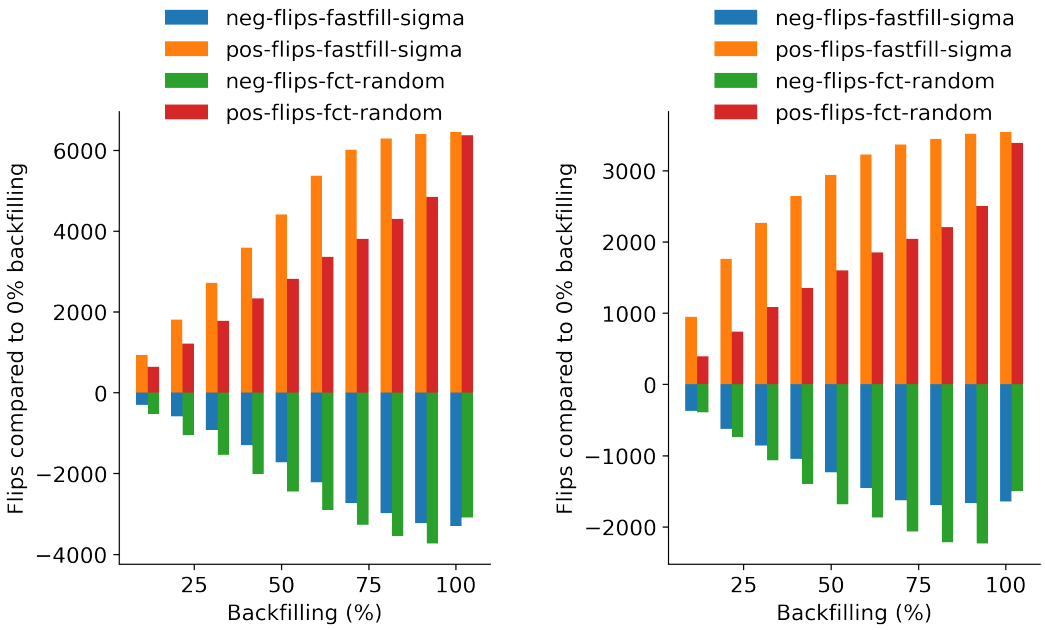



图 13：ImageNet 上的正向与负向翻转分析。我们分别使用 (a) CMC-top1 和 (b) top-5 检索，在 ImageNet 上比较 FastFill（蓝色和橙色）与 FCT（绿色和红色）。目标是在尽可能减少回填的同时，尽量增加正向翻转的数量（橙色和红色），并减少负向翻转的数量（蓝色和绿色）。

## E SIGMA 分析

在上述论文的主要内容分析基础上，我们展示了 $\sigma^2$ 与不同损失函数（$\mathcal{L}_{l_2+disc}(\boldsymbol{x}_i)$、 $\mathcal{L}_{l_2}(\boldsymbol{x}_i)$ 和 $\mathcal{L}_{disc}(\boldsymbol{x}_i)$）及其所产生的排序的相关程度。我们的预测 $\sigma^2$ 与 $\mathcal{L}_{l_2+disc}(\boldsymbol{x}_i)$ 的相关性最高（两种排序的 Kendall-Tau 相关性为 0.67），这并不意外，因为模型经过训练，目标就是预测该损失。不过，它对 $\mathcal{L}_{l_2}(\boldsymbol{x}_i)$（Kendall-Tau 相关性为 0.65）和 $\mathcal{L}_{disc}(\boldsymbol{x}_i)$（Kendall-Tau 相关性为 0.63）的预测也很准确。

当我们训练不确定性估计器来预测 $\mathcal{L}_{l_2}(\boldsymbol{x}_i)$ 时，与 $\mathcal{L}_{l_2}(\boldsymbol{x}_i)$ 所产生排序的相关性（Kendall-Tau=0.73）会提高，但与此同时，与基于 $\mathcal{L}_{disc}(\boldsymbol{x}_i)$ 的排序的相关性则显著降低（Kendall-Tau=0.56）。

当我们使用 $\mathcal{L}_{disc}(\boldsymbol{x}_i)$ 进行训练时，仍能观察到其与 $\mathcal{L}_{disc}(\boldsymbol{x}_i)$ 所确定的排序之间存在较强的相关性（Kendall-Tau=0.62），但与 $\mathcal{L}_{l_2}(\boldsymbol{x}_i)$ 的排序相关性要弱得多（Kendall-Tau=0.34）。

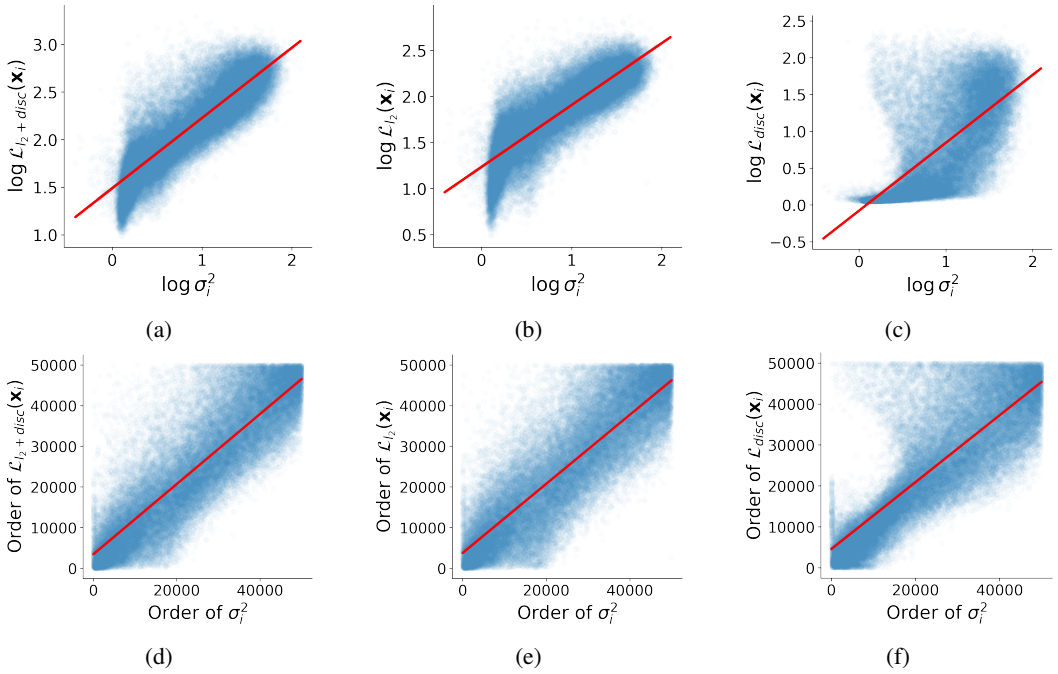
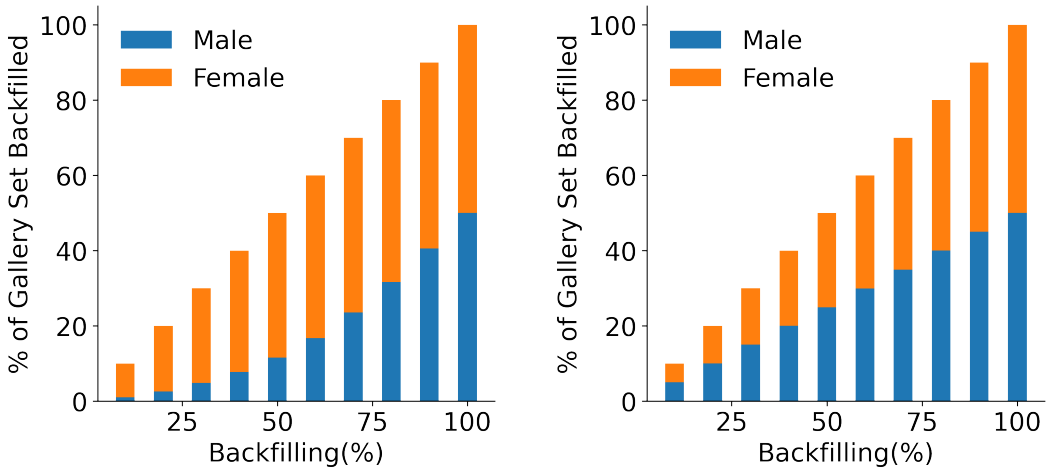



图 14：a、b、c）我们展示了预测的 $\log \sigma_i^2$、 $\psi_\vartheta$ 在图库图像上的输出，以及 $\log \mathcal{L}_{l_2+disc}(\boldsymbol{x}_i)$、$\log \mathcal{L}_{l_2}(\boldsymbol{x}_i)$ 和 $\log \mathcal{L}_{disc}(\boldsymbol{x}_i)$ 分别具有良好的相关性。c）我们分别按照 $\sigma_i^2$ 和 $\mathcal{L}_{l_2+disc}(\boldsymbol{x}_i)$、 $\mathcal{L}_{l_2}(\boldsymbol{x}_i)$、$\mathcal{L}_{disc}(\boldsymbol{x}_i)$ 对 $\boldsymbol{x}_i$ 排序一次，从而得到两组排序。这里，对于每个 $\boldsymbol{x}_i$，我们绘制其在第一组排序中的顺序与在第二组排序中的顺序，并观察到很强的相关性（Kendall-Tau 相关性=0.67、0.65 和 0.63）。

F 回填顺序分析 - VGGFACE2-GENDER

现在，我们使用 FastFill 方法在 VGGFace2-性别数据集上分析哪些类别会最先被回填。我们将类别分为少数群体（女性）和多数群体（男性）。如图所示15 FastFill 会优先回填少数类别，这解释了主论文中第 5.3 节 所观察到的行为。




(a) (b)

图 15：我们使用 (a) FastFill 和 (b) FCT-Random 比较少数群体（女性）和多数群体（男性）类别的回填顺序。FastFill 优先处理少数群体，并首先回填女性图库图像。因此，FastFill 仅在回填 ∼ 25% 后，就达到了新模型的准确率差距。