arXiv:2305.04239v1 [cs.CV] 7 May 2023

基于高斯径向基函数核的3D跨模态检索实例变异损失

Zhitao Liu*
zl425uestc@gmail.com 电子科技大学 成都，中国

Zengyu Liu*
lzystan2021@163.com 电子科技大学 成都，中国

Jiwei Wei
mathematic6@gmail.com 电子科技大学 成都，中国

Guan Wang
wangguan12621@gmail.com 电子科技大学 成都，中国

Zhenjiang Du
zhenjiang@std.uestc.edu.cn 电子科技大学 成都，中国

Ning Xie†
seanxiening@gmail.com 电子科技大学 中国技术 成都，中国

Hengtao Shen
shenhengtao@hotmail.com 电子科技大学 成都，中国

摘要

三维跨模态检索正受到多媒体领域的广泛关注。该领域的核心在于学习一个联合嵌入空间，以表示不同模态的数据，例如图像、3D点云和多边形网格，从而提取模态不变的判别性特征。因此，跨模态检索方法的性能在很大程度上取决于该嵌入空间的表征能力。现有方法对所有实例一视同仁，对难度不同的实例施加相同的惩罚强度，而忽略了实例之间的差异。这可能导致模糊收敛或陷入局部最优，从而严重损害特征空间的可分性。为了解决这一局限性，我们提出了一种实例变体损失（Instance-Variant Loss），为不同实例分配不同的惩罚强度，以提高空间的可分性。具体而言，我们根据实例的类内距离，为其分配与之正相关的不同惩罚权重。同时，通过为不同模态中属于同一类别的数据学习一个共享权重向量，缩小特征之间的跨模态差异。此外，我们利用高斯径向基函数核评估样本相似度，进一步提出了一种类内损失函数，以最小化同类实例之间的类内距离。在三个三维跨模态数据集上进行的大量实验表明，我们提出的方法优于近期最先进的方法。

ACM计算分类系统概念

• **信息系统 → 检索模型与排序；多媒体与多模态检索。**

关键词

深度度量学习、三维跨模态检索、跨域特征学习

ACM参考文献格式：

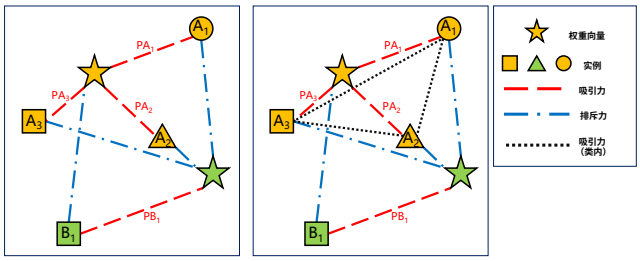
Zhitao Liu、Zengyu Liu、Jiwei Wei、Guan Wang、Zhenjiang Du、Ning Xie 和 Hengtao Shen。2023。基于高斯径向基函数核的3D跨模态检索实例变异损失。载于《请确保输入权利确认邮件中的正确会议名称（*MM'23*）》。ACM，纽约，纽约州，美国，10页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

随着三维模型在 CAD、虚拟现实/增强现实和自动驾驶应用中的日益普及，如何高效、准确地检索三维模型，正受到多模态领域越来越多的关注。作为机器人导航、场景理解、3D建模和动画等众多下游任务的基础，这一领域已广受关注 [5]。三维跨模态检索旨在缩小跨模态差异，并从多模态数据中学习模态不变（最小化类内距离）且具有判别性的特征（最大化类间距离）。与 2D跨模态检索（图文检索 [33, 40],基于草图的图像检索 [22, 26]）相比，三维跨模态检索必须考虑三维模型的表示形式和结构，并利用更多多模态数据，包括图像、点云、网格以及三维模型的多视角灰度图像，作为检索的查询域。

*两位作者对本研究的贡献相同。
†通讯作者

允许免费制作本作品全部或部分内容的数字版或纸质副本，供个人或课堂使用，前提是副本不得用于盈利或商业用途，也不得为此制作或分发，并且副本首页须包含本声明和完整引文。对于本作品中版权归 ACM 以外的其他方所有的内容，必须尊重其版权。注明出处后允许摘录。若要以其他方式复制、再版、发布到服务器或分发给列表，须事先获得明确许可和/或支付费用。请通过 permissions@acm.org 申请许可。*MM'23，2023年10月29日至11月3日，加拿大渥太华© 2023 美国计算机协会。ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00 https://doi.org/XXXXXXX.XXXXXXX



(a) 相等惩罚强度
$PA_1 = PA_2 = PA_3 = PB_1$

(b) 与类内距离正相关的惩罚强度
$PA_3 < PA_2 < PA_1 < PB_1$

图1：采用高斯径向基函数核的实例变体损失（类内损失）用于三维跨模态检索的示意图。形状相同的点来自同一模态。不同颜色表示不同的类别。 PA_1 和 PB_1 表示每个实例的惩罚强度。红色和蓝色虚线的长度分别表示类内距离和类间距离，分别在实例与权重向量之间产生吸引力和排斥力。惩罚强度越大，吸引力和排斥力越强。(a)：所有实例具有相同的惩罚强度；(b)：不同实例具有不同的惩罚强度，且该强度与类内距离呈正相关。类内损失可以使类内实例彼此更加接近。

三维跨模态检索的关键挑战在于缩小三维模型与图像之间显著的跨模态差异。现有跨模态检索方法着重学习公共嵌入空间，以弥合异质性差距，从而能够比较不同模态的特征。公共嵌入空间旨在最小化同一类别内的类内距离，并最大化不同类别之间的类间距离。近年来，研究人员采用对比损失 [9, 13],跨模态中心损失 [11], Softmax交叉熵损失 [12], 以及其他对比学习和度量学习方法，来学习能够表征多模态数据特征的公共嵌入空间。为了最小化类内距离，研究人员通过计算不同类别的类别中心（特征均值），使用跨模态中心损失和交叉熵损失，为所有模态数据学习公共嵌入空间 [11]。也有研究采用对比学习来实现模态间质心对齐，从而减小不同数据模态之间的模态差异 [9]。

尽管这些方法取得了显著进展，但大多数方法仍将所有模态中各类别的所有实例一视同仁，忽视了实例之间的差异，导致收敛模糊和性能不佳 [34]。跨模态检索任务通过优化类内距离与类间距离之间的间隔，提升网络在公共空间中的表征能力。如图 1 所示，不同实例对网络的贡献存在很大差异。对难度不同的实例施加相等的惩罚强度，可能导致收敛模糊或陷入局部最优，严重损害特征空间的可分性。此外，三维模型与图像之间固有的数据表征和结构差异，使得学习可分的嵌入空间充满挑战。这种显著的模态差异使网络难以从数据中学习模态不变特征。因此，三维跨模态任务的检索性能往往会受到收敛模糊的影响。因此，为三维跨模态检索任务提出一种有效的实例级惩罚强度加权损失至关重要。该方法将考虑不同实例对网络的贡献，并为不同实例分配不同的惩罚强度。

为了解决上述问题，我们提出了用于三维跨模态检索的实例变体损失。最直观的动机是为不同实例分配不同的惩罚强度。通过在 Softmax 输出中加入一个与实例类内距离呈正相关的可学习权重系数，我们确保更难以区分的实例（类内距离更大）获得更强的惩罚强度，从而提升网络的有效性。同时，由于惩罚强度与实例的类内距离相关，网络也将能够适应数据分布不平衡和多模态优化不平衡问题。不同于经典的 Softmax 损失，我们为不同实例加入不同的惩罚强度，并为所有模态学习一个通用的权重向量。这种方法将有效缓解模态优化不平衡问题，并减少模态之间的差异。

在 Softmax 损失的约束下，我们可以通过对权重向量和实例特征进行归一化，将所有模态数据映射到共享的单位超球面 [14]上。单位超球面有助于更好地缩小类内距离、扩大类间距离。它还有助于减少不同数据源之间的差异，增强所学数据的语义信息，并提高检索任务的鲁棒性。受高斯径向基核函数 [1], 的启发，我们提出了类内损失，旨在缩小单位超球面上同一类别所有样本之间的距离。与中心损失不同，我们的类内损失无需寻找类别中心。高斯核的本质是衡量样本之间的相似度。在描述相似度的空间中，相似样本可以更好地聚集在一起，进而变得线性可分。使用类内损失可以缩小类内距离，并评估网络学习到的多模态共享单位超球面（嵌入空间）。

通过采用所提出的实例变体损失和类内损失，网络能够在最小化类内距离的同时，为不同实例分配不同的惩罚强度，从而有效降低因收敛模糊而导致的特征空间不可分性。为验证所提出的实例变体损失和类内损失的有效性，我们将该框架与交叉熵损失函数联合训练，用于三维跨模态检索任务，旨在提取模态不变且具有判别性的特征。我们的方法在三维跨模态和单模态检索任务中显著优于近期最先进的方法。本文的主要贡献可概括如下：

- 我们提出了实例变体损失，它能为不同实例有效分配不同的惩罚强度，通过聚焦于更具挑战性的实例，提升网络的效能，并增强特征空间的可分性。

- 我们提出了基于高斯 RBF 核函数的类内损失，旨在最小化共享嵌入空间中所有实例之间的类内距离。这种方法能够最小化类内距离，并评估网络学习到的多模态共享嵌入空间。

- 所提出的实例变体损失为来自不同模态的数据学习一个共享权重向量，从而有效缓解跨模态差异并提升跨模态检索性能。

- 我们的方法在三个数据集（Pix3D、ModelNet40、MI3DOR）上，针对三维跨模态和单模态检索任务，显著优于近期最先进的方法。这证明了所提出的实例变体损失和类内损失的有效性。

2 相关工作

三维跨模态检索。近年来，人们对三维跨模态检索算法的兴趣日益浓厚，包括基于图像的三维形状检索（IBSR）、基于三维模型的形状检索以及三维跨模态互检索。基于图像的方法将三维形状表示为从该形状的预定义视点采集的一组三维视图 [10, 38, 44]。使用一组二维视图表征的优势在于，它们可以直接利用现有的强大卷积神经网络进行特征提取 [16, 23]，并缩小三维模型与图像之间的域差距。Lin等人[13] 采用对比学习，根据单张图像实现了实例级三维形状检索。迄今为止，IBSR 任务已取得显著进展。然而，这类任务仍存在一些局限，例如查询域仅限于图像，且缺少三维模型的全局空间信息和局部几何信息。基于模型的形状检索通常将三维模型表示为多边形网格 [3, 6] 和点云 [20, 21, 31]。PointNet [20]开发了逐点操作和对称函数，以解决置换不变性问题。DGCNN [31] 提出了一种动态图卷积神经网络，利用K个最近邻点进行EdgeConv。MeshNet [3] 和MeshCNN [6] 通过建模物体网格面的几何关系，设计了直接从网格中学习特征的方法。基于模型的方法的优势在于能够探索三维模型的全局空间信息和局部几何信息，从而获得具有代表性的三维描述子。然而，这些方法主要关注单模态检索，并未进一步探索跨模态检索。

三维跨模态互检索全面考虑了上述方法的优缺点。它们利用 DGCNN、ImageNet 和 MeshNet 等特征提取网络，从点云、图像和多边形网格中提取特征。通过测量共享嵌入空间中的类内距离和类间距离，跨模态互检索能够实现任意两种模态之间的互检索。Jing等人 [11]开创了三维跨模态（互）检索领域，使用跨模态中心损失来缩小模态之间的差异，并最小化样本的类内距离。Chen等人[2] 利用多模态对比原型损失实现了半监督三维跨模态检索。三维跨模态互检索检索具有深远影响。它支持从真实（投影）图像、点云和网格数据中进行跨模态检索。借助三维跨模态（相互）检索技术，研究人员可以快速为三维模型找到对应的纹理材质，并迅速在不同的三维模型表示方法之间进行转换。本文旨在利用所提出的实例变体损失和类内损失，实现真实图像与三维模型之间的跨模态相互检索。

度量学习。度量学习旨在学习一种距离度量函数，使语义相关的实例彼此接近。在以往的文献中，研究人员针对各种任务提出了大量度量学习方法 [25, 30, 42]。Wang 等人[30] 提出了多重相似度损失，用于收集信息丰富的样本对并为其分配权重。Sun 等人[25]引入了圆损失，以对不同的相似度分数进行加权。然而，上述方法是为单模态检索任务开发的，通常无法准确捕捉跨模态组件之间的关系，因为不同模态之间存在模态鸿沟。Frome 等人[4] 尝试将图像和句子投影到公共嵌入空间中，并使用未加权的三元组损失来促使语义相关的实例聚集在一起。Wei 等人[32, 34] 引入了通用加权度量学习框架，根据样本对的相似度分数对信息丰富的样本对进行采样并为其分配适当的权重值。然而，上述方法无法适应类别级跨模态检索。受现有数据集局限性的约束，研究人员难以构建样本对或三元组。因此，在本文中，我们针对三维跨模态检索开发了一种新颖的度量学习方法，该方法可以根据实例的类内距离直接调整其惩罚强度，从而改变其对损失函数的贡献。

3 所提方法

本节将介绍我们提出的基于高斯径向基核函数的实例变体损失和类内损失。第 3.1 小节将阐述问题描述和预备知识。第 3.2 和 3.3 小节将分别介绍实例变体损失和类内损失的数学公式及分析。

3.1 问题陈述与预备知识

假设数据集 S 包含 $n*N*M$ 个实例，其中包含来自 M 种模态的 N 个类别的 n 个实例，第 i个实例 t_i 是一个集合，由模态集合 s_i、语义标签 y_i和权重向量 W_{y_i}组成。形式化地说：

$$\begin{aligned} S &= \{t_i\}_{i=1}^{n*N}, \qquad t_i = (s_i, y_i, Wy_i), \\ s_i &= \{x_i^m\}_{m=1}^{M}, \qquad Wy_i = \{w_{y_i}^m\}_{m=1}^{M}. \end{aligned} \tag{1}$$

通常，不同模态的实例 $\{x_i^1、\ x_i^2、\cdots\ x_i^M\}$ 位于 M不同的表示空间中，因此无法直接测量它们之间的相似度（距离）。三维跨模态检索任务旨在为每种模态 $m \in [1, M]$学习 M投影函数 f_m，其中 $v_i^m = f_m(x_i^m、\ \theta_m)$，$V_{y_i}^m = \{v_1^m、\ v_2^m、\cdots\ v_n^m\}$，$\theta_m$是一个可学习参数。$v_i^m$是公共表示空间中的投影特征，$V_{y_i}^m$是同一类别所有特征的集合。为了获得

最佳检索性能，我们旨在确保同类实例之间的距离，即类内距离 $D(V_{y_i}^m、\ W_{y_i})$，比不同类别实例之间的距离，即类间距离 $D(V_{y_i}^m、\ W_j)$，小一个显著的间隔 λ_0。

$$\begin{aligned} &D(V_{y_i}^m, V_{y_i}^{m'}) + \lambda_0 < D(V_{y_i}^m, V_{y_j}^{m'}), \\ &D(V_{y_i}^m, W_{y_i}) + \lambda_0 < D(V_{y_i}^m, W_j). \end{aligned} \tag{2}$$

为了更好地理解所提出的实例变体损失，我们将首先简要回顾原始 Softmax、A-Softmax [14], 和 AM-Softmax [28]。原始 Softmax损失的公式如下:

$$\begin{aligned} \mathcal{L}_S &= -\frac{1}{n}\frac{1}{N}\sum_{i=1}^{n*N}\log\frac{e^{W_{y_i}\cdot f_i}}{\sum_{j=1}^{N} e^{W_j f_i}} \\ &= -\frac{1}{n}\frac{1}{N}\sum_{i=1}^{n*N}\log\frac{e^{\|W_{y_i}\|\|f_i\|\cos(\theta_{y_i})}}{\sum_{j=1}^{N} e^{\|W_j\|\|f_i\|\cos(\theta_j)}}, \end{aligned} \tag{3}$$

其中，f是最后一个全连接层的输入（f_i表示第i个样本），W_j是最后一个全连接层的第 j列。$W_{y_i}\cdot\ f_i$也称为第 i个样本的目标对数值 [19]。W_j也称为第 j个类别的权重向量。权重向量与特征均值向量（类别中心）之间的关系如 [29]中的图 6 所示。在 A-Softmax 损失中，作者提出对权重向量进行归一化（使$\|W_i\|$为 1），并将目标对数值从 $\|f_i\|\cos(\theta_{y_i})$推广为 $\|f_i\|\psi(\theta_{y_i})$，其中 $\psi(\theta) = (-1)^k\cos(m\theta) - 2k + \lambda\cos(\theta)1+\lambda$。在 AM-Softmax 中，作者提出对权重向量和实例特征（$\|W_i\| = \|f_i\| = 1$）同时进行归一化，而 $\psi(\theta_{y_i}) = \cos(\theta_{y_i}) - \lambda_0$。

3.2 实例变体损失

Softmax损失函数已广泛应用于单模态检索任务，因为它不仅能够确定用于区分不同数据类别的最优平面，还能为每个类别学习理想的权重向量。与类别中心（特征均值）相比，权重向量更适用于涉及困难样本的度量学习问题。然而，与单模态检索任务不同，跨模态检索任务还必须减小跨模态差异。我们创新性地为不同模态的数据学习一个共享权重向量（$Wy_i = w_{y_i}^1 = w_{y_i}^M$）。在权重向量的迭代过程中，它将获取所有模态数据的特征，从而有效解决模态优化中的不平衡问题。本文假设 W_i和 f的范数均归一化为 1，除非另有说明。给定提取特征 $\{V_i^m\}$ $(i \in [1, n*N],\ \ m \in [1, M])$，$f_i^m = V_i^m$。新的跨模态Softmax损失形式如下。

$$\begin{aligned} \mathcal{L}_{NS} &= -\frac{1}{n}\frac{1}{N}\frac{1}{M}\sum_{i=1}^{n*N}\sum_{m=1}^{M}\log\frac{e^{w_{y_i}\cdot f_i^m}}{\sum_{j=1}^{c} e^{w_j f_i^m}} \\ &= -\frac{1}{n}\frac{1}{N}\frac{1}{M}\sum_{i=1}^{n*N}\sum_{m=1}^{M}\log\frac{e^{w_{y_i}\cdot V_i^m}}{\sum_{j=1}^{c} e^{w_j\cdot V_i^m}}. \end{aligned} \tag{4}$$

为简化后续推导，我们可以将公式 4 重写为公式 5。

$$\begin{aligned} \mathcal{L}_{NS'} &= -\frac{1}{n}\frac{1}{N}\frac{1}{M}\sum_{i=1}^{n*N}\sum_{m=1}^{M}\log\frac{e^{\phi}(\theta_{y_i}^m)}{e^{\phi(\theta_{y_i}^m)} + \sum_{j=1,j\neq y_i}^{N} e^{\eta(\theta_j^m)}} \\ &= \frac{1}{n}\frac{1}{N}\frac{1}{M}\sum_{i=1}^{n*N}\sum_{m=1}^{M}\log(1 + \sum_{j=1,j\neq y_i}^{N} e^{\eta(\theta_j^m) - \phi(\theta_{y_i}^m)}). \end{aligned} \tag{5}$$

为确保类内距离小于类间距离，$1 - \cos(\theta_j^m) - (1 - \cos(\theta_{y_i}^m)) \geq \lambda_0$，$\lambda_0 > 0$，其中 $\cos(\theta_{y_i}^m) - \lambda_0 \geq \cos(\theta_j^m)$，其中 $\cos(\theta_{y_i}^m) > \cos(\theta_j^m)$。

因此，我们可以令 $\eta(\theta_j^m) = \|W_j\|\|f_i^m\|\cos(\theta_j^m) = \cos(\theta_j^m)$，以及 $\phi(\theta_{y_i}^m) = \|W_i\|\|f_i^m\|\cos(\theta_{y_i}^m) - \lambda_0 = \cos(\theta_{y_i}^m) - \lambda_0$。

如果网络对所有实例施加相同的惩罚强度，模糊收敛（局部最优）会严重损害特征空间的可分性。为解决这一局限，我们考虑通过允许每个实例根据当前优化状态按自身节奏学习，来增强优化灵活性。实例变体损失的形式如下。

$$\begin{aligned} \mathcal{L}_{IV} = \frac{1}{n}\frac{1}{N}\frac{1}{M}\sum_{i=1}^{n*N}\sum_{m=1}^{M}\frac{\sum_{j=1,j\neq y_i}^{N} e^{\eta(\theta_j^m) - \phi(\theta_{y_i}^m)}}{1 + \sum_{j=1,j\neq y_i}^{N} e^{\eta(\theta_j^m) - \phi(\theta_{y_i}^m)}} \\ \times\log(1 + \sum_{j=1,j\neq y_i}^{N} e^{\eta(\theta_j^m) - \phi(\theta_{y_i}^m)}). \end{aligned} \tag{6}$$

令 $\Gamma = \sum_{j=1,\ \ j\neq y_i}^{N} e^{\eta}(\theta_j^m) - \phi(\theta_{y_i}^m)$，

$$\begin{aligned} \mathcal{L}_{IV} &= \frac{1}{n}\frac{1}{N}\frac{1}{M}\sum_{i=1}^{n*N}\sum_{m=1}^{M}\frac{\Gamma}{1+\Gamma}\times log(1+\Gamma) \\ &= \frac{1}{n}\frac{1}{N}\frac{1}{M}\sum_{i=1}^{n*N}\sum_{m=1}^{M} log(1+\Gamma)^{\frac{\Gamma}{1+\Gamma}} \\ \mathcal{L}_{NS'} &= \frac{1}{n}\frac{1}{N}\frac{1}{M}\sum_{i=1}^{n*N}\sum_{m=1}^{M} log(1+\Gamma). \end{aligned} \tag{7}$$

在输入数据和优化器等影响损失收敛的因素相同的条件下，$L_{IV} < L_{NS'}$。对 Γ 关于 $\theta_{y_i}^m(\theta_{y_i}^m \in [0, \pi])$ 求偏导，并对 L_{IV} 和 $L_{NS'}$ 关于 Γ、$\frac{\partial\Gamma}{\partial\theta_{y_i}^m} < 0$、$d\mathcal{L}_{d\Gamma}^{IV} > 0$，$d\mathcal{L}_{d\Gamma}^{NS'} > 0$ 求导。因此，我们可以得到，L_{IV} 和 $L_{NS'}$ 关于 $\theta_{y_i}^m$ 均单调递减。假设 L_{IV} 和 $L_{NS'}$ 被优化到相同的值，并且所有训练特征都能被完美分类（与 L-softmax 的假设 [15] 一致）。

$$\begin{aligned} &\because \mathcal{L}_{IV} < \mathcal{L}_{NS'}, \quad \frac{d\mathcal{L}_{IV}}{d\Gamma} > 0, \quad \frac{d\mathcal{L}_{NS'}}{d\Gamma} > 0, \\ &\quad min(\mathcal{L}_{IV}) = min(\mathcal{L}_{NS'}). \\ &\therefore \Gamma_{IV} > \Gamma_{NS'}. \qquad \because \frac{\partial\Gamma}{\partial\theta_{y_i}^m} < 0. \\ &\therefore \{\theta_{y_i}^m\}_{IV} < \{\theta_{y_i}^m\}_{NS'}. \end{aligned} \tag{8}$$

当 L_{IV} 和 $L_{NS'}$ 优化到相同的值，且所有训练特征都能被完美分类时，L_{IV} 的 $\theta_{y_i}^m$ 将会是

小于 $L_{NS'}$，这意味着 L_{IV} 的性能将优于 $L_{NS'}$（证明见附录）。在训练优化过程中，当一个实例难以区分时（类内距离较大或类间距离较小），它将受到更强的惩罚。与传统的 Softmax 损失相比，我们引入了困难负样本挖掘的概念。此外，我们综合评估类内距离和类间距离，以定义困难样本，从而有效提高特征空间的可分性。我们还可以使用超参数 τ 来缩放 $\frac{\Gamma}{1+\Gamma}$，以适应不同数据集中的不同数据分布。同时，研究人员只需修改 $\phi(\theta_{y_i}^m)$ 和 $\eta(\theta_j^m)$，即可提高网络的检索性能。所提出的实例变体损失可以表示为以下形式。

$$\begin{aligned}\mathcal{L}_{IV} &= \frac{1}{n}\frac{1}{N}\frac{1}{M}\sum_{i=1}^{n*N}\sum_{m=1}^{M}\left(\frac{\Gamma}{1+\Gamma}\right)^{\tau}\times log(1+\Gamma).\\ \Gamma &= \sum_{j=1,j\neq y_i}^{N} e^{\eta(\theta_j^m)-\phi(\theta_{y_i}^m)}\\ \eta(\theta_j^m) &= \cos(\theta_j^m)/\omega,\quad \phi(\theta_{y_i}^m) = (\cos(\theta_{y_i}^m)-\lambda_0)/\omega,\end{aligned} \tag{9}$$

其中，λ_0 和 τ 是超参数，ω 是用于Softmax的温度系数 [8] 。

3.3 高斯径向基函数类内损失

在Softmax损失的约束下，我们可以通过对权重向量和实例特征进行归一化，将所有模态数据映射到共享的单位超球面上。我们的目标是使类内度量在渐近意义上正确，并在点数有限时具有经验上的合理性。此外，高斯径向基函数核的本质是衡量样本之间的相似度。相似的样本可以在能够描述其相似度的空间中更好地聚集在一起，进而实现线性可分。为此，我们采用高斯径向基函数（RBF）核。

$$\mathcal{K}_t(x,x') \triangleq e^{-t\|x-x'\|_2^2} = e^{2t\cdot x^\top x'-2t},\quad t>0, \tag{10}$$

并将类内损失定义为平均成对高斯径向基函数的负对数似然：

$$\begin{aligned}\mathcal{L}_{IC} &= -\frac{1}{N}\sum_{y_i=1}^{N}\frac{1}{n*M}\log\Big(\sum_{i=1,j=1,i\neq j}^{n*M}\mathcal{K}_t(x,x')\Big)\\ &= -\frac{1}{N}\sum_{y_i=1}^{N}\frac{1}{n*M}\log\Big(\sum_{i=1,j=1,i\neq j}^{n*M}e^{-t\left\|x_i^{y_i}-x_j^{y_i}\right\|_2^2}\Big).\end{aligned} \tag{11}$$

该损失函数的目标是最小化不同模态中任意两个同类实例之间的距离，从而减小同一类别所有数据的类内距离。与现有的类内损失函数相比 [11, 37],，我们的类内损失直接最小化两个实例之间的距离，无需学习类别中心，从而降低网络复杂度和跨模态差异。我们将在附录中进一步讨论其数学特性。

4 实验

数据集。为了验证我们提出的方法，我们在三个数据集上进行了实验：ModelNet40 [36], MI3DOR [43],Pix3D [24],以及包含四个类别的 Pix3D（Pix3D-4，是由 [13]创建的 Pix3D 数据集子集）。ModelNet40 数据集是一个三维物体基准数据集，包含 12,311 个属于 40 个不同类别的 CAD 模型，其中 9,843 个用于训练，2,468 个用于测试。该数据集提供三种模态：图像、点云和网格。MI3DOR 是一个用于二维到三维任务的大规模数据集，包含来自 21 个类别的 21,000 张图像和 7,690 个模型，其中 3,842 个用于训练，3,848 个用于测试。Pix3D 是一个包含真实图像和真实标注模型的大规模数据集，具有精确的二维三维对齐，包含来自 9 个类别的 395 个模型和 16,913 张图像，其中 313 个用于训练，82 个用于测试。Pix3D-4 仅选取包含 300 个以上未遮挡且未截断样本的类别。Pix3D-4 包含来自 4 个类别（床、椅子、沙发、桌子）的 322 个模型，其中 257 个用于训练，65 个用于测试。MI3DOR 和 Pix3D 数据集都只有三维模型（OBJ格式的复杂网格）。因此，我们对 MI3DOR 和 Pix3D 数据集中的三维模型进行采样，分别为每个数据集生成包含 1024 个面的网格数据和包含 2048 个点的点云数据。

4.1 实现细节

我们提出了一种端到端的跨模态检索任务框架，该框架基于所提出的实例变体损失、类内损失和交叉熵损失。图 2 展示了三维跨模态检索任务框架的概览。对于二维图像特征提取，我们采用 ResNet18 [7]作为骨干网络，该网络包含四个卷积块，均使用 3×3 卷积核，卷积核数量分别为 64、128、256 和 512。我们采用 DGCNN 作为骨干网络来提取点云特征。DGCNN [31]包含四个 EdgeConv 块，卷积核数量分别设为 64、64、64 和 128。我们使用 MeshNet [3] 提取网格特征。共享的多模态特征编码器由两个全连接层组成，其维度为 512、256、C（C 为类别数）。我们使用所提出的三种损失函数训练网络，使其联合学习具有判别性和模态不变的特征：$L = L_{IV} + L_{IC} + L_{CE}$.

训练细节。我们使用 PyTorch [17]实现了网络。对于所有三个数据集，网络均采用 SGD 优化器进行训练，学习率为 0.01。对于 ModelNet40 和 MI3DOR，学习率每 20,000 次迭代降低 90%；对于 Pix3D，则每 4,000 次迭代降低 90%。我们为所有三个数据集设置了温度参数 $\omega = \frac{1}{30}$ 和 $\lambda_0 = 0.35$，并分别为 ModelNet40、MI3DOR 和 Pix3D 设置了 $\tau = 0.1$、$\tau = 0.1$和 $\tau = 8$。由于我们是首个使用 MI3DOR 和 Pix3D 等真实世界数据集开展三维跨模态检索任务研究的工作，因此缺乏可用于比较的基线方法。因此，我们选择跨模态中心损失（简称 CMCL） [11]作为基线，并在上述数据集上进行了相关实验。CMCL 在 MI3DOR 和 Pix3D 数据集上的训练细节与在 ModelNet40 上相同。

评估指标。所有实验的评估结果均采用平均精度均值（mAP）分数呈现，这是跨模态检索任务的一项经典性能评估标准 [2, 11, 39]。检索任务的平均精度均值用于衡量检索到的数据是否与查询属于同一类别（相关）或不属于同一类别（不相关）。给定一个查询和一组对应的 R

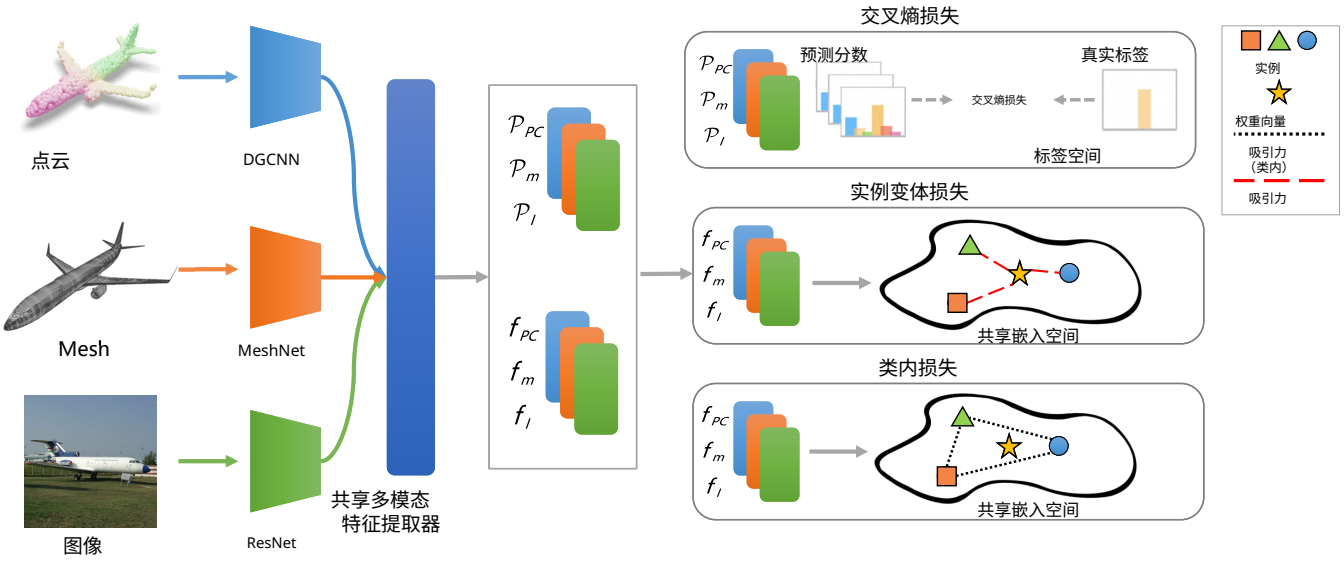


图 2：我们为 3D 跨模态检索任务提出的框架概览。在标签空间中采用交叉熵损失。在共享嵌入空间中，利用实例变体损失学习具有判别性且模态不变的特征。同时，类内损失专注于学习模态不变特征，进一步提升了我们所提出框架的性能。不同的颜色和形状表示来自同一类别不同模态的实例。

表1：ModelNet40、MI3DOR、Pix3D-9（原始 Pix3D）和 Pix3D-4 数据集上三维单模态与跨模态检索任务的性能（以平均精度均值计）。与当前最先进方法 CMCL [11]进行比较。当目标或源来自图像域时，结果针对多视图图像（仅适用于 ModelNet40）进行报告：1、2、4 个视图分别记为 v1、v2 和 v4。

数据集		ModelNet40						MI3DOR		Pix3D-9		Pix3D-4	
		mAP-v1		mAP-v2		mAP-v4		mAP		mAP		mAP	
源	目标	CMCL	Ours	CMCL	Ours	CMCL DSCMR [41]	Ours	CMCL	Ours	CMCL	Ours	CMCL	Ours
图像	图像	82.06	**83.55**	86.00	**88.13**	90.23	82.31	**90.50**	75.26	**77.24**	72.97	**77.99**	85.97	**91.38**
图像	Mesh	85.58	**86.20**	87.31	**88.64**	89.59	77.30	**89.94**	78.24	**79.78**	73.79	**81.64**	86.78	**90.86**
图像	点	85.23	**85.77**	86.79	**88.16**	89.04	74.33	**89.48**	79.69	**80.15**	74.48	**80.15**	84.16	**91.43**
Mesh	图像	83.58	**85.21**	85.96	**87.76**	88.11	76.18	**88.84**	79.14	**82.14**	87.50	**92.83**	94.74	**94.81**
点	图像	82.29	**85.23**	85.18	**87.87**	87.11	73.74	**89.03**	79.40	**81.82**	74.54	**83.72**	86.22	**93.09**
Mesh	Mesh	88.51	**89.30**	——	——	——	74.84	——	85.59	**87.05**	83.36	**90.75**	90.54	**94.72**
Mesh	点	87.37	**88.14**	——	——	——	70.21	——	85.74	**86.59**	88.53	**89.05**	95.22	**95.77**
点	点	87.04	**88.51**	——	——	——	70.80	——	86.32	**86.63**	72.58	**81.47**	87.77	**92.23**
点	Mesh	87.58	**89.04**	——	——	——	71.59	——	85.91	**86.85**	85.30	**89.16**	92.43	**92.54**
Mean		85.47	**86.77**	86.97	**88.39**	88.29	74.59	**89.20**	81.70	**83.14**	79.23	**85.20**	89.31	**92.98**

检索到的数据（排名前 R 的数据）中，平均精度定义为：

$$AP = \frac{1}{T}\sum_{r=1}^{R} P_r \times \delta(r), \tag{12}$$

其中， T 表示检索集合中的相关项数量， P_r表示前 r 个检索项的精确率， $\delta(r)$是一个指示函数：当第 r个检索项相关时，其值为 1（此处“相关”是指属于查询所属的类别）。平均精度均值（MAP）可通过对平均精度（AP）值求均值来计算。

4.2 三维跨模态检索任务

为了评估所提出损失函数的有效性，我们在 ModelNet40、MI3DOR 和 Pix3D 数据集上，针对图像、点云和网格三种不同模态开展实验。为了全面考察学习到的特征的质量，我们执行了两种检索任务：单模态检索和跨域检索。我们的方法在三维单模态和跨模态检索任务上的性能如表1所示。由于跨模态中心损失（CMCL）仅针对三维跨模态检索提出，且其仅在 ModelNet 数据集上进行了实验，因此我们在 MI3DOR 和 Pix3D 数据集上复现 CMCL 方法，以评估其在真实图像上的检索性能。我们

表2：在ModelNet40数据集上对损失函数进行的消融研究。图像视图数固定为1。所有实验使用相同的训练轮数，批次大小固定为128。

Loss	$\mathcal{L}_{CE}$	$\mathcal{L}_{IV}$	$L_{NS'}$	L_{IV} +L_{IC}	L_{CE} +L_{IV}	L_{CE} +L_{IV} +L_{IC}	L_{CE} +L_{IC}
图像到图像	77.65	80.55	79.73	82.95	80.97	**83.55**	81.61
图像到网格	82.26	84.84	84.54	85.82	84.90	**86.20**	83.05
图像到点	76.91	84.09	83.86	84.86	83.83	**85.77**	81.17
网格到网格	86.78	88.56	83.86	88.66	89.28	**89.30**	84.60
网格到图像	82.13	83.19	82.67	84.93	83.97	**85.21**	82.29
网格到点	81.35	87.30	87.23	86.75	87.77	**88.14**	81.57
点到点	71.97	87.70	87.24	87.52	87.57	**88.51**	79.95
点到图像	75.56	82.68	81.93	84.64	83.34	**85.23**	80.44
点到网格	80.83	88.03	87.88	88.39	88.32	**89.04**	81.96
Mean	79.49	85.22	84.84	86.06	85.55	**86.77**	81.85

提出的联合训练方法在所有检索任务和数据集上均显著优于最先进的方法。特别是，当输入的检索数据为真实图像时，相较于 CMCL 方法，我们的检索方法展现出更显著的性能优势。这表明我们的方法能够有效处理包含棘手负样本的真实数据集。这得益于我们的实例变体损失为不同实例分配不同的惩罚强度，从而改善了空间分隔效果。在所有数据集的各个检索对上，我们的方法均取得了显著更好的性能，展现了所提方法强大的泛化能力。

4.3 损失函数的影响

我们提出的损失函数包含以下三个组成部分：标签空间中每种模态的交叉熵损失，记为 L_{CE}；共享嵌入空间中的实例变体损失，记为 L_{IV}；以及类内损失，记为 L_{IC}。我们还提及不带实例变体权重的 Softmax损失，记为 $L_{NS'}$。为进一步研究各组成部分的影响，我们评估了损失函数的不同组合，包括分别使用 L_{CE}、L_{IV} 和 $L_{NS'}$ 进行优化；联合使用 L_{CE}&L_{IV}、L_{CE}&L_{IC}、L_{IV}&L_{IC}进行优化；以及联合使用 L_{CE}、L_{IC} 和 L_{IV}进行优化。这五个网络采用相同的设置和超参数进行训练，其性能如表2所示。如表2所示：A) L_{CE}、L_{IC} 和 L_{IV} 的组合在所有跨模态和单模态检索任务中均取得最佳性能。B) 作为基线，仅使用交叉熵损失也能取得相对较高的平均精度均值，这是因为三种模态共享一个头，迫使网络在公共空间中为同一类别的不同模态学习相似的表示。C) 实例变体损失可以单独使用，并取得相当不错的检索结果。与类内损失结合时，某些任务的检索性能超过了 CMCL 方法。D) L_{IC} 可以提升单模态和跨模态检索的性能。E) L_{IV} 优于 $L_{NS'}$。

4.4 批量大小的影响

很少有研究者探讨权重向量与类别中心 [28, 29] 之间的差异。研究者比较了完全优化前后权重向量与类别中心分布的差异，得出结论：它们的分布重叠于

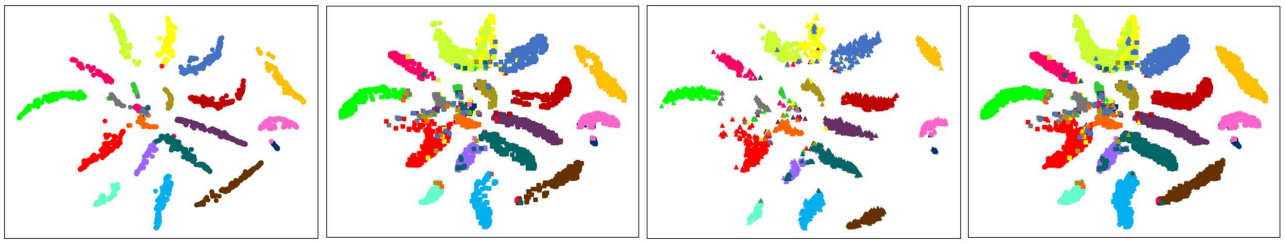
表3：在 ModelNet40 数据集上针对批次大小的消融研究。图像的视图数量固定为 1。

批次大小	32	64	96	128
图像到图像	81.78	83.34	83.45	**83.55**
图像到网格	85.16	**86.23**	86.18	86.20
图像到点	83.81	85.29	85.51	**85.77**
网格到网格	89.57	**89.59**	89.47	89.30
网格到图像	84.91	**85.49**	85.33	85.21
网格到点	87.39	87.95	87.89	**88.14**
点到点	86.77	87.49	88.02	**88.51**
点到图像	83.93	84.84	85.11	**85.23**
点到网格	88.07	88.61	88.83	**89.04**
Mean	85.71	86.54	86.64	**86.77**

表 4：MI3DOR 与 Pix3D 之间三维单模态和跨模态未见检索任务的性能。

未见	MI3DOR：图像->视图		MI3DOR->Pix3D-9	
方法	CMCL	Ours	CMCL	Ours
图像到图像	24.77	**30.24**	54.15	**54.29**
图像到网格	25.82	**31.69**	**53.03**	48.27
图像到点	25.92	**31.29**	51.14	**53.08**
网格到网格	85.59	**87.05**	76.72	**76.75**
网格到图像	22.45	**25.28**	**46.32**	42.22
网格到点	85.71	**86.59**	48.24	**48.91**
点到点	86.19	**86.63**	77.75	**80.32**
点到图像	21.81	**24.16**	46.07	**48.55**
点到网格	85.80	**86.85**	64.17	**64.57**
Mean	51.56	**54.42**	**57.51**	57.44

最优网络优化 [29]。已有研究证明，批量大小会影响中心损失的检索性能 [11, 35] 。因此，我们进一步研究了批量大小对权重向量计算和类内损失的影响。为分析批次大小对性能的影响，我们在 ModelNet40 数据集上使用不同的批次大小（32、64、96、128）进行了实验。所有网络均使用相同的训练轮数和超参数进行训练。如表3所示，改变批量大小并不会显著



(a) 图像特征 (b) 点云特征 (c) 网格特征 (d) 所有模态

图3：使用 t-SNE 方法对 MI3DOR 数据集中的测试数据进行可视化 [27]。来自同一类别的实例以相同颜色呈现。

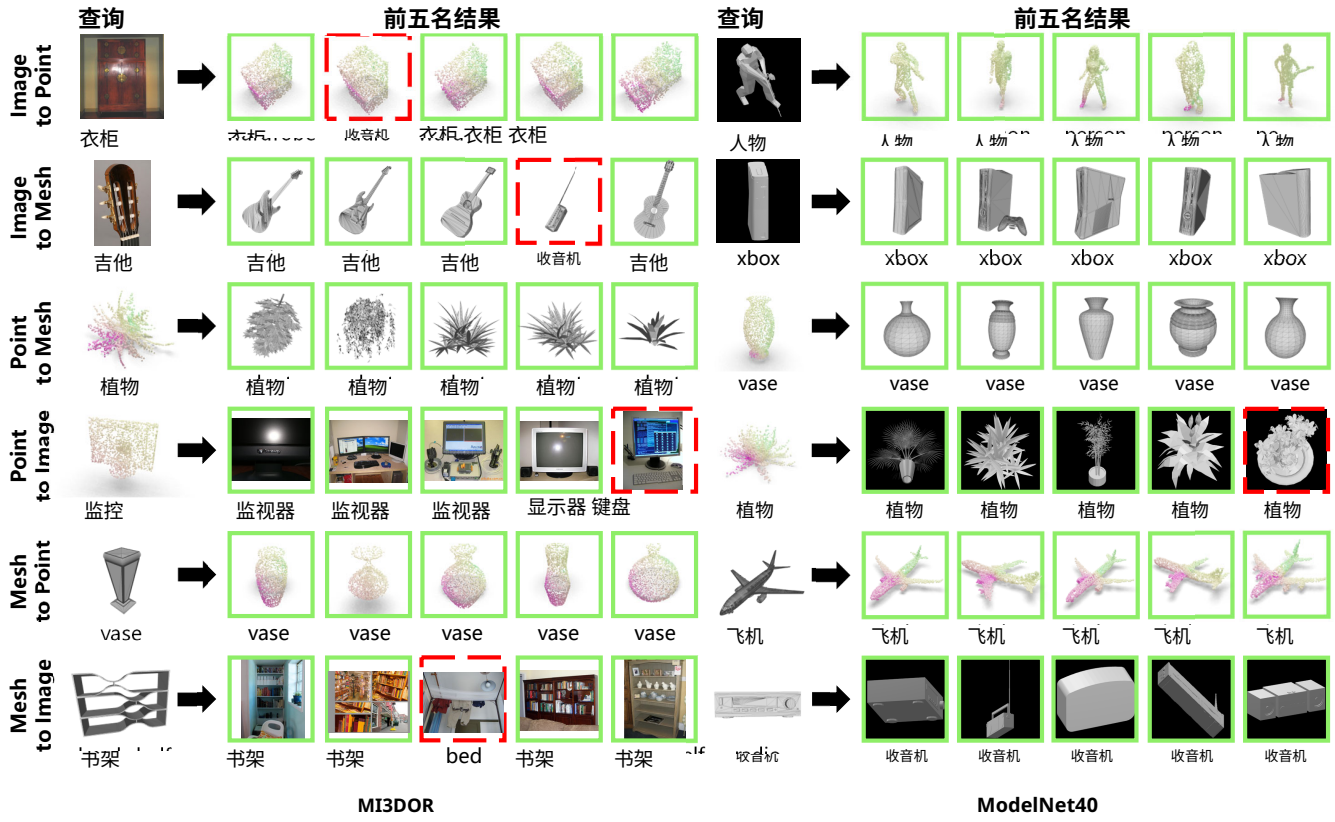


图4：ModelNet40 和 MI3DOR 数据集上的前5名检索结果。绿色边界框表示图像属于同一类别 es 表示查询，而红色虚线边界框表示错误匹配。

影响检索性能，这表明权重向量和类内损失不会受到批次内数据量变化的影响。这也意味着，本文提出的损失函数只需较少资源，就能取得相当的检索结果。同时，我们注意到，使用较小的批次大小时，部分检索任务的性能有所提升。我们推测，主要原因是超参数选择和模态不平衡优化问题导致特征提取不够完善 [18]。

4.5 三维跨模态未见检索任务

考虑到跨模态检索在现实世界中的应用，我们可能会遇到训练数据集中未出现过的新数据域或分布。因此，我们使用 MI3DOR 数据集中的真实图像来训练检索网络，并使用 MI3DOR 和 Pix3D 数据集中的视图进行测试。选择 MI3DOR 数据集作为基准数据集的主要原因是，它既包含与模型对应的真实图像，也包含模型的多视图图像；同时，与 Pix3D 相比，该数据集中的数据分布更加均衡，因此我们无需担心

测试所用网络的无偏性。如表 4 所示，我们的网络比 C MCL 方法具有更强的泛化能力。然而，它未能解决未见域检索这一挑战，因为网络的性能显著下降。我们将在附录中讨论更多有关未见域检索的实验结果。

4.6 定性可视化

t-SNE 特征嵌入可视化。图 3 展示了每种模态各自的不同聚类，凸显了所提方法在区分不同类别样本方面的有效性。此外，跨模态的组合特征证实，学习到的公共空间能够捕捉模态不变的表征。

三维跨模态检索可视化。图 4 展示了从 ModelNet40 和 MI3DOR 数据集中选取的六个查询的跨模态检索样本。余弦距离使用每个查询的归一化特征来衡量不同模态间的数据相似度。图中显示，尽管模态不同，外观相似的实例在特征空间中的距离更近，这证明网络学到了模态不变特征。更多实验结果将在附录中展示。

5 结论

本文为三维跨模态检索提出了实例变体损失和类内损失。实例变体损失能够为不同实例有效分配不同的惩罚强度，通过关注更具挑战性的实例，提升网络的有效性，并增强特征空间的可分性。基于高斯径向基函数核的类内损失，旨在最小化共享嵌入空间中所有实例之间的类内距离。该方法能够最小化类内距离，并评估网络学习到的共享嵌入空间。所提出的实例变体损失为来自不同模态的数据学习一个共享权重向量，从而有效缓解跨模态差异，提升跨模态检索性能。我们在三维跨模态检索任务上进行了大量实验。与当前先进方法相比，所提出的框架在 ModelNet40、MI3DOR 和 Pix3D 数据集上显著优于它们。未来，我们将进一步探索潜在技术，以提升所提出损失的鲁棒性和有效性，并增强对未见数据的检索性能。

参考文献

[1] Sergiy V Borodachov、Douglas P Hardin 和 Edward B Saff。2019年。可求 *rectifiable* 集上的离散能量。第 3 卷。施普林格。[2] Zhimin Chen、Longlong Jing、Yang Liang、Yingli Tian 和 Bing Li。2021年。三维物体的多模态半监督学习。载于：第 *32* 届英国机器视觉大会 *2021*（*BMVC 2021*），线上举行，*2021* 年 *11* 月 *22—25* 日。BMVA Press，第 381 页。[3] Yutong Feng、Yifan Feng、Haoxuan You、Xibin Zhao 和 Yue Gao。2019年。MeshNet：用于三维形状表示的网格神经网络。载于：第三十三届 *AAAI* 人工智能大会（*AAAI 2019*）、第三十一届人工智能创新应用大会（*IAAI 2019*）、第九届 *AAAI* 人工智能教育进展研讨会（*EAAI 2019*），美国夏威夷州檀香山，*2019* 年 *1* 月 *27* 日至 *2* 月 *1* 日。AAAI Press，第 8279–8286 页。[4] Andrea Frome、Gregory S. Corrado、Jonathon Shlens、Samy Bengio、Jeffrey Dean、Marc'Aurelio Ranzato 和 Tomás Mikolov。2013年。DeViSE：一种深度视觉语义嵌入模型。载于：神经信息处理系统进展 *26*：第 *27* 届神经信息处理系统年度大会 *2013*。会议论文集，会议于 *2013* 年 *12* 月 *5—8* 日在美国内华达州太浩湖举行。第 2121–2129 页。

[5] Xian-Feng Han、Hamid Laga 和 Mohammed Bennamoun。2019。基于图像的三维物体重建：深度学习时代的最新进展与趋势。*IEEE*模式分析与机器智能汇刊43, 5（2019）。[6] Rana Hanocka、Amir Hertz、Noa Fish、Raja Giryes、Shachar Fleishman 和 Daniel Cohen-Or。2019。MeshCNN：一种带边的网络。*ACM*图形学汇刊38, 4（2019），90:1–90:12。[7] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。2016。用于图像识别的深度残差学习。载于*2016 IEEE*计算机视觉与模式识别会议（*CVPR 2016*），美国内华达州拉斯维加斯，*2016*年*6*月*27*日至*30*日。IEEE计算机学会，770–778。[8] Geoffrey Hinton、Oriol Vinyals 和 Jeff Dean。2015。在神经网络中蒸馏知识。arXiv: 1503 02531 [ ]。统计机器学习 Nian Hu、Heyu [9] Zhou、Xiangdong Huang、Xuanya Li 和 An-An Liu。2022。用于二维图像到三维形状检索的选择性伪标签特征变换框架。*IEEE*视频技术电路与系统汇刊32, 11（2022），8010–8021。[10] Jingjia Huang、Wei Yan、Ge Li、Thomas H. Li 和 Shan Liu。2022。学习用于多视图三维物体识别的解耦表示。*IEEE*《电路与系统视频技术汇刊》32, 2（2022），646–659。[11] Longlong Jing、Elahe Vahdani、Jiaxing Tan 和 Yingli Tian。2021。用于三维跨模态检索的跨模态中心损失。载于*IEEE*计算机视觉与模式识别会议（*CVPR 2021*），虚拟会议，*2021*年*6*月*19*日至*25*日。计算机视觉基金会 / IEEE，3142–3151。[12] Shuang Liang、Weidong Dai 和 Yichen Wei。2021。用于抗噪声的基于草图三维形状检索的不确定性学习。*IEEE*图像处理汇刊30（2021），8632–8643。[13] Ming-Xian Lin、Jie Yang、He Wang、Yu-Kun Lai、Rongfei Jia、Binqiang Zhao 和 Lin Gao。2021。通过跨模态实例与类别对比学习进行单幅图像三维形状检索。载于*2021 IEEE/CVF*国际计算机视觉大会（*ICCV 2021*），加拿大魁北克省蒙特利尔，*2021*年*10*月*10*日至*17*日。IEEE，11385–11395。[14] Weiyang Liu、Yandong Wen、Zhiding Yu、Ming Li、Bhiksha Raj 和 Le Song。2017。SphereFace：用于人脸识别的深度超球面嵌入。载于*2017 IEEE*计算机视觉与模式识别会议（*CVPR 2017*），美国夏威夷州檀香山，*2017*年*7*月*21*日至*26*日。IEEE计算机学会，6738–6746。[15] Weiyang Liu、Yandong Wen、Zhiding Yu 和 Meng Yang。2016。用于卷积神经网络的大间隔Softmax损失。载于第*33*届国际机器学习大会论文集，*ICML 2016*，美国纽约市，*2016*年*6*月*19*日至*24*日（*JMLR*研讨会与会议论文集，第*48*卷）。JMLR.org，507–516。[16] Zhenguang Liu、Shuang Wu、Shuyuan Jin、Shouling Ji、Qi Liu、Shijian Lu 和 Li Cheng。2023。研究用于三维运动预测的姿态表示与运动上下文建模。*IEEE*模式分析与机器智能汇刊45, 1（2023），681–697。[17] Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga、Alban Desmaison、Andreas Köpf、Edward Z. Yang、Zachary DeVito、Martin Raison、Alykhan Tejani、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chintala。2019。PyTorch：一种命令式风格的高性能深度学习库。载于神经信息处理系统进展*32*：*2019*年神经信息处理系统年会（*NeurIPS 2019*），*2019*年*12*月*8*日至*14*日，加拿大不列颠哥伦比亚省温哥华。8024–8035。[18] Xiaokang Peng、Yake Wei、Andong Deng、Dong Wang 和 Di Hu。2022。通过即时梯度调制实现平衡多模态学习。载于*IEEE/CVF*计算机视觉与模式识别会议（*CVPR 2022*），美国路易斯安那州新奥尔良，*2022*年*6*月*18*日至*24*日。IEEE，8228–8237。[19] Gabriel Pereyra、George Tucker、Jan Chorowski、Lukasz Kaiser 和 Geoffrey E. Hinton。2017。通过惩罚置信输出分布对神经网络进行正则化。载于第*5*届国际学习表征会议（*ICLR 2017*），法国土伦，*2017*年*4*月*24*日至*26*日，研讨会论文集。OpenReview.net。[20] Charles Ruizhongtai Qi、Hao Su、Kaichun Mo 和 Leonidas J. Guibas。2017。PointNet：在点集上进行深度学习以实现三维分类与分割。载于*2017 IEEE*计算机视觉与模式识别会议（*CVPR 2017*），美国夏威夷州檀香山，*2017*年*7*月*21*日至*26*日。IEEE计算机学会，77–85。[21] Charles Ruizhongtai Qi、Li Yi、Hao Su 和 Leonidas J. Guibas。2017。PointNet++：在度量空间中的点集上进行深度层次特征学习。载于神经信息处理系统进展*30*：*2017*年神经信息处理系统年会，*2017*年*12*月*4*日至*9*日，美国加利福尼亚州长滩。5099–5108。[22] Aneeshan Sain、Ayan Kumar Bhunia、Yongxin Yang、Tao Xiang 和 Yi-Zhe Song。2021。StyleMeUp：迈向风格无关的基于草图图像检索。载于*IEEE*计算机视觉与模式识别会议（*CVPR 2021*），虚拟会议，*2021*年*6*月*19*日至*25*日。计算机视觉基金会 / IEEE，8504–8513。[23] Hang Su、Subhransu Maji、Evangelos Kalogerakis 和 Erik G. Learned-Miller。2015。用于三维形状识别的多视图卷积神经网络。载于*2015 IEEE*国际计算机视觉大会（*ICCV 2015*），智利圣地亚哥，*2015*年*12*月*7*日至*13*日。IEEE计算机学会，945–953。

[24] Xingyuan Sun、Jiajun Wu、Xiuming Zhang、Zhoutong Zhang、Chengkai Zhang、Tianfan Xue、Joshua B. Tenenbaum 和 William T. Freeman。2018。Pix3D：单图像三维形状建模的数据集与方法。载于《2018年IEEE计算机视觉与模式识别会议》，CVPR 2018，美国犹他州盐湖城，2018年6月18日至22日。计算机视觉基金会 / IEEE计算机学会，2974–2983。[25] Yifan Sun、Changmao Cheng、Yuhan Zhang、Chi Zhang、Liang Zheng、Zhongdao Wang 和 Yichen Wei。2020。圆损失：成对相似度优化的统一视角。载于《2020年IEEE/CVF计算机视觉与模式识别会议》，CVPR 2020，美国华盛顿州西雅图，2020年6月13日至19日。计算机视觉基金会 / IEEE，6397–6406。[26] Jialin Tian、Xing Xu、Zheng Wang、Fumin Shen 和 Xin Liu。2021。用于零样本草图图像检索的关系保持知识蒸馏。载于MM '21 24, 2021 20 5473 5481：ACM多媒体会议，线上活动，中国，October - 。ACM，–。[27] Laurens Van der Maaten 和 Geoffrey Hinton。2008。使用 t-SNE 可视化数据。机器学习研究期刊 9, 11 (2008)。[28] Feng Wang、Jian Cheng、Weiyang Liu 和 Haijun Liu。2018。用于人脸验证的加性间隔 Softmax。IEEE信号处理快报 25, 7 (2018)，926–930。[29] Feng Wang、Xiang Xiang、Jian Cheng 和 Alan Loddon Yuille。2017。NormFace：用于人脸验证的 L2超球面嵌入。载于《2017年ACM多媒体会议论文集》，MM 2017，27, 2017, 23，美国加利福尼亚州山景城，October - Qiong Liu、Rainer Lienhart、Haohong Wang、Sheng-Wei "Kuan-Ta" Chen、Susanne Boll、Yi-Ping Phoebe Chen、Gerald Friedland、Jia Li 和 Shuicheng Yan（编）。ACM，1041–1049。[30] Xun Wang、Xintong Han、Weilin Huang、Dengke Dong 和 Matthew R. Scott。2019。用于深度度量学习的多重相似度损失与通用成对加权。载于《IEEE计算机视觉与模式识别会议》，CVPR 2019，美国加利福尼亚州长滩，2019年6月16日至20日。计算机视觉基金会 / IEEE，5022–5030。[31] Yue Wang、Yongbin Sun、Ziwei Liu、Sanjay E. Sarma、Michael M. Bronstein 和 Justin M. Solomon。2019。用于点云学习的动态图卷积神经网络。ACM图形学汇刊 38, 5 (2019)，146:1–146:12。[32] Jiwei Wei、Xing Xu、Zheng Wang 和 Guoqing Wang。2021。用于跨模态匹配的元自步学习。载于MM '21：ACM多媒体会议，线上活动，中国，2021年10月20日至24日。ACM，3835–3843。[33] Jiwei Wei、Yang Yang、Xing Xu、Jingkuan Song、Guoqing Wang 和 Heng Tao Shen。2023。少即是多：用于跨模态匹配的指数损失。IEEE视频技术电路与系统汇刊 (2023)，1–1。[34] Jiwei Wei、Yang Yang、Xing Xu、Xiaofeng Zhu 和 Heng Tao Shen。2022。用于跨模态检索的通用加权度量学习。IEEE模式分析与机器智能汇刊 44, 10 (2022)，6534–6545。[35] Yandong Wen、Kaipeng Zhang、Zhifeng Li 和 Yu Qiao。2016。一种深度人脸识别的判别性特征学习方法。载于《计算机视觉——ECCV 2016——第14届欧洲会议》，荷兰阿姆斯特丹，2016年10月11日至14日，论文集，第VII部分（计算机科学讲义，第9911卷）。施普林格，499–515。[36] Zhirong Wu、Shuran Song、Aditya Khosla、Fisher Yu、Linguang Zhang、Xiaoou Tang 和 Jianxiong Xiao。2015。3D ShapeNets：体积形状的深度表征。载于《IEEE计算机视觉与模式识别会议》，CVPR 2015，美国马萨诸塞州波士顿，2015年6月7日至12日。IEEE计算机学会，1912–1920。[37] Degui Xiao、Jiazhi Li、Jianfang Li、Shiping Dong 和 Tao Lu。2022。IHEM损失：用于鲁棒人脸识别的类内困难样本挖掘损失。IEEE《电路与系统视频技术汇刊》32, 11 (2022)，7821–7831。[38] Yongzhe Xu、Jiangchuan Hu、Kanoksak Wattanachote、Kun Zeng 和 Yongyi Gong。2020。通过最佳视角选择和跨域相似度度量实现基于草图的形状检索。IEEE多媒体汇刊 22, 11 (2020)，2950–2962。[39] Zhixiong Zeng、Ying Sun 和 Wenji Mao。2021。MCCN：用于大规模跨模态检索的多模态协同聚类网络。载于MM '21：ACM多媒体会议，线上活动，中国，2021年10月20日至24日。ACM，5427–5435。[40] Kun Zhang、Zhendong Mao、Quan Wang 和 Yongdong Zhang。2022。用于图文匹配的负样本感知注意力框架。载于《 IEEE/CVF计算机视觉与模式识别会议》，CVPR 2022，美国路易斯安那州新奥尔良，2022年6月18日至24日。IEEE，15640–15649。[41] Liangli Zhen、Peng Hu、Xu Wang 和 Dezhong Peng。2019。深度监督跨模态检索。载于《IEEE计算机视觉与模式识别会议》，CVPR 2019，美国加利福尼亚州长滩，2019年6月16日至20日。计算机视觉基金会 / IEEE，10394–10403。[42] Wenzhao Zheng、Jiwen Lu 和 Jie Zhou。2020。通过自适应可学习评估实现深度度量学习。载于《2020年IEEE/CVF计算机视觉与模式识别会议》，CVPR 2020，美国华盛顿州西雅图，2020年6月13日至19日。计算机视觉基金会 / IEEE，2957–2966。[43] Heyu Zhou 、 An-An Liu 和 Weizhi Nie。2019。用于基于二维图像的三维物体检索的双层嵌入对齐网络。载于《第27届ACM多媒体国际会议论文集》，MM 2019，25, 2019 21 1667 1675，法国尼斯，October - 。ACM，–。

[44] Yaqian Zhou、Yu Liu、Heyu Zhou、Zhiyong Cheng、Xuanya Li 和 An-An Liu。2022年。学习用于基于二维图像的三维模型检索的可迁移且具有判别性的表示。*IEEE*《电路与系统视频技术汇刊》32，10（2022年），7147–7159。