arXiv:2304.06275v1 [cs.CV] 13 Apr 2023

基于元相似度校正的噪声对应学习

Haochen Han[1], Kaiyao Miao[2],Qinghua Zheng[1],Minnan Luo[1*1] 西安交通大学计算机科学与技术学院，中国[2] 西安交通大学网络空间安全学院，中国{hhc1997, miaoky814}@stu.xjtu.edu.cn，{qhzheng, minnluo}@xjtu.edu.cn

摘要

尽管多模态学习在跨模态检索任务中取得了成功，但这一显著进展依赖于多媒体数据之间的正确对应关系。然而，收集此类理想数据既昂贵又耗时。在实践中，大多数广泛使用的数据集都从互联网中收集而来，不可避免地包含不匹配对。在此类噪声对应数据集上进行训练会导致性能下降，因为跨模态检索方法可能会错误地强制不匹配数据变得相似。为解决这一问题，我们提出了一种元相似度校正网络（*MSCN*），以提供可靠的相似度分数。我们将二分类任务视为元过程，促使 *MSCN* 从正负元数据中学习判别能力。为进一步减轻噪声的影响，我们设计了一种有效的数据净化策略，利用元数据作为先验知识来移除噪声样本。我们在合成噪声和真实世界噪声两种情况下进行了广泛实验，包括 *Flickr30K*、*MS-COCO* 和 *ConceptualCaptions*，以证明我们方法的优势。我们的代码已公开发布。[1]

1. 引言

近年来，随着多媒体数据的快速增长，跨模态检索备受关注。给定特定模态的查询样本，跨模态检索旨在检索不同模态中的相关样本。现有的跨模态检索工作 [1, 6, 11] 通常会学习一个可比较的共同空间，以连接不同模态，并在许多应用中取得了显著进展，包括视频-音频检索 [1, 33], 视觉问答 [11, 26], 和图文匹配 [44]。

尽管前景广阔，跨模态检索中的一个核心假设是多个.....之间存在正确的对应关系

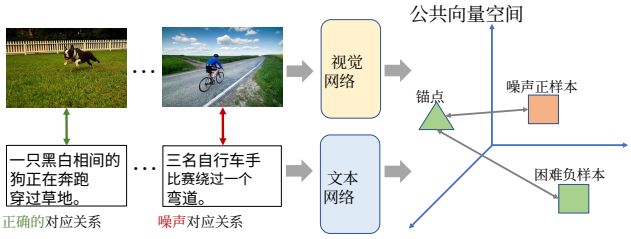

图 1. 图文检索中噪声对应关系的示意图。标准三元组损失用于确保共同空间中的正样本对比负样本更接近。噪声对应关系是指不匹配对被错误地视为正样本对，因此会导致模型性能下降。

模态。然而，收集此类理想数据既昂贵又耗时。在实践中，大多数广泛使用的数据集都从互联网中获取，并且不可避免地包含噪声对应关系 [31]。如图1所示，跨模态检索方法在利用噪声对应关系进行学习时，会错误地强制不匹配的数据变得相似，这可能会显著影响检索性能。迄今为止，针对这一问题的研究工作很少。Huang [15]首次研究了这一问题，并提出 NCR 方法，以稳健地从噪声图文对中进行训练。受噪声标签研究 [20],先前成功经验的启发，NCR 将数据划分为干净数据分区和噪声数据分区，并利用自适应模型校正对应关系。然而，NCR 基于深度神经网络的记忆效应 [3],，导致其在高噪声比例下性能不佳。

为应对这一挑战，我们提出**元相似度校正网络（MSCN）**，旨在为主网络中的噪声特征提供可靠的相似度分数。我们将二分类任务视为元过程：给定一个多模态样本，MSCN 将学习判断各模态是否相互对应，其预测结果可自然地视为相似度分数。具体而言，我们使用少量干净数据对构建正负元数据，并将二者视为元知识，以促使 MSCN 学习判别能力。同时，主网络

*Minnan Luo 为通讯作者。[1]

https://github.com/hhc1997/MSCN

利用 MSCN 校正后的相似度分数进行训练，并采用自适应边界实现稳健学习。这种相互依赖关系使主网络和 MSCN 能够相互促进，共同抵御噪声。然而，由于三元组损失的特性，即使采用理想的相似度分数，噪声样本对仍会产生正损失。为此，我们进一步提出一种元知识引导的数据净化策略，以移除可能存在错误对应关系的样本。我们进行了大量实验，证明了该方法在合成噪声和真实世界噪声下的优势。

本工作的主要贡献总结如下：

- 我们率先探索了针对噪声对应关系问题的元学习，并提出一种元校正网络，以在噪声干扰下提供可靠的相似度分数。
- 我们提出了一种新颖的元过程，首先将正样本和负样本都视为元知识，以促使 MSCN 从中学习判别能力。
- 我们设计了一种有效的数据净化策略，将元数据用作先验知识，以去除噪声样本。

2. 相关工作

2.1. 跨模态检索

跨模态检索旨在将一种模态的数据作为查询，以检索其他模态中的相关数据。根据对标注信息的利用方式，跨模态检索方法大致可分为两类：1）无监督方法 [8, 22, 25, 36]。它仅利用共现信息来学习多模态数据之间的共同表示。2）有监督方法 [14,34,38,45]。采用额外的标签信息来增强共同表示的判别能力。然而，缺乏标注限制了其在现实世界中的实用性。

我们提出的方法属于无监督方法这一类别。例如，Wang [36]使用双分支神经网络来学习多模态数据的联合嵌入。受困难负样本挖掘的启发，VSE++[8] 使用困难负样本来提升检索性能。为了捕捉模态之间的细粒度交互，SCAN [18] 引入堆叠交叉注意力，使注意力能够同时利用图像和句子的上下文。最近，受图模型强大学习能力的启发，GSMN [25] 和 SGRAF [7] 为多模态数据构建图结构，以促进细粒度对应关系的学习。尽管现有工作已经取得了显著成果，但它们通常依赖于多模态数据之间的正确对应关系，并且无法解决噪声问题。因此，探索如何利用噪声对应关系进行跨模态检索学习具有重要意义，但以往研究很少涉及这一问题。

2.2. 噪声对应学习

作为一个新近提出的问题，噪声对应关系指的是不匹配对却被错误地视为正样本。Huang [15]首次研究了这一问题，并提出 NCR 方法，以利用噪声对应对鲁棒地训练图文匹配模型。受此前噪声标签研究成功的启发 [20], , NCR 根据深度神经网络的记忆效应，将数据划分为干净数据和噪声部分，然后利用自适应模型校正对应关系。Qin [29]提出了一种基于不确定性的方法，以实现高效学习。此外，一些研究[40, 41]探讨了部分视图未对齐问题，该问题可视为广义的噪声对应关系学习。近期，一些研究 [13,39] 探索了更具挑战性的场景，即噪声标签和噪声对应关系可能同时出现。相比之下，我们从元学习视角应对跨模态检索中的这一挑战。

2.3. 元学习

元学习的目标是在比传统学习更高的层次上进行学习，例如学习更新规则 [28], 寻找易于微调的参数 [9], 或适应新任务 [19]。近期，研究人员利用元学习寻找能够抵御噪声标签的模型参数 [21, 32, 46]。例如，Li [21] 在传统训练前优化元目标，使模型不会对噪声过拟合。MW-Net [32] 采用元过程自动为训练样本分配权重。MLC [46] 提出了一种以元学习方式训练的标签校正网络，用于为含噪训练数据生成可靠标签。

与本文最相关的现有方法是 MLC [46], , 它首先将噪声标签问题表述为元标签校正问题。然而，我们认为，所提方法与 MLC 有两方面不同。首先，我们关注多模态数据中的噪声对应问题，而非单模态分类场景。其次，我们利用正负元数据来引导模型学习判别能力，这一点在以往的元学习方法中尚未得到探索。

3. 所提方法

3.1. 问题表述

在不失一般性的前提下，我们以图文检索任务为代理，研究跨模态检索中的噪声对应问题。给定一个训练集 $\mathcal{D}_{train} = \{(I_i,\ T_i,\ y_i)\}_{i=1}^{N}$，其中（$I_i$，$T_i$）是第 i个图像-

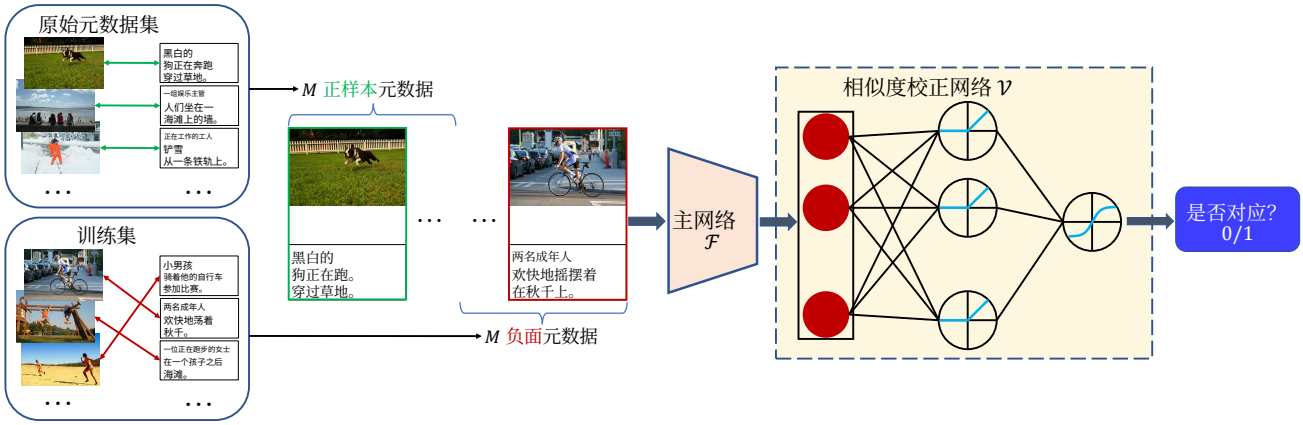

图2. 所提出元过程的概览。正样本对取自包含对应图像和文本的原始元数据，而负样本对则从训练集中的不同样本对中提取，以构造不匹配对。随后，训练 MSCN 以判断这些模态是否彼此对应。

文本对，$y_i \in \{1, 0\}$ 表示该对为正例（匹配），即（I_i, T_i），或负例（不匹配），即（I_i，$T_{j\neq i}$），N是整个训练数据的数量。双模态数据中的噪声对应是指负例却被错误地视为 $y_i = 1$。为解决这一问题，我们提出一种基于元学习的方法，以实现稳健训练。接下来将详细介绍。

3.2 基于元学习的相似度校正

训练目标。首先，我们介绍标准投影模块，使跨模态特征具有可比性。令 $f(I; W_f)$ 和 $g(T; W_g)$ 分别表示将视觉模态和文本模态映射到联合嵌入空间的模态专属网络。然后，我们通过函数计算双模态特征之间的相似度特征

$$S\left(f\left(I\right), g\left(T\right); W_s\right) = \frac{W_s|f(I) - g(T)|^2}{\|W_s|f(I) - g(T)|^2\|_2}, \quad (1)$$

其中，W_s 是用于获取低维相似度表示的可学习参数矩阵。为方便记号，我们将 $\{W_f, W_g, W_s\}$ 汇集为 W，作为主网络 $\mathcal{F} = (f, g, S)$的参数，并将相似度特征记为$\mathcal{F}_W(I, T)$。存在噪声对应关系时，若直接使用相似度表示，模型不可避免地会对噪声过拟合。为解决这一问题，我们提出元相似度校正网络（MSCN），将获取相似度分数的过程视为一个元过程，以噪声相似度表示为输入，并输出校正后的相似度分数。具体而言，我们的 MSCN 表述为一个多层感知机网络 $\mathcal{V}_\Theta(\cdot)$，其参数为 Θ。输出 $s_i = \mathcal{V}_\Theta(\mathcal{F}_W(I_i, T_i))$经 Sigmoid函数激活，以表示位于$[0, 1]$中的相似度分数。

对于含噪训练对，主网络在 MSCN 生成的校正后相似度分数的引导下学习抗噪特征；而 MSCN 则基于主网络提取的特征预测可靠的相似度分数。这种相互依赖使主网络和元网络能够相互强化以抵御噪声，并通过双层优化进行优化。形式上，最优参数 W^* 通过最小化训练数据上的累积损失来计算：

$$W^*(\Theta) = \arg\min_W \mathbb{E}_{(I_i,T_i)\in\mathcal{D}_{train}}\, l^{train}(I_i, T_i), \quad (2)$$

其中，$l^{train}(\cdot)$ 是三元组排序损失：

$$\begin{aligned} l^{train}(I_i, T_i) = &[\hat{\gamma} - \mathcal{V}_\Theta\left(\mathcal{F}_W(I_i, T_i)\right) + \mathcal{V}_\Theta\left(\mathcal{F}_W(I_i, T_i^-)\right)]_+ \\ &+ [\hat{\gamma} - \mathcal{V}_\Theta\left(\mathcal{F}_W(I_i, T_i)\right) + \mathcal{V}_\Theta\left(\mathcal{F}_W(I_i^-, T_i)\right)]_+, \end{aligned} \quad (3)$$

其中 $[x]_+ = max(x, 0)$、I_i^- 和 T_i^- 是与给定样本对 (I_i, T_i) 相对应的最难负样本，与 VSE 类似++[8]。为实现稳健的跨模态检索，我们设计了一种自适应边界 $\hat{}\gamma$，以替代 VSE 中的固定值++。理想情况下，该边界应对噪声样本对赋予较小的值，对干净数据样本对赋予较大的值。为此，我们根据相似度分数调整边界：

$$\hat{\gamma} = \frac{1}{1 + (\frac{s_i}{1-s_i})^{-\tau}}\gamma, \quad (4)$$

其中 γ 是原始边际值，$\tau > 0$ 是参数。这种自适应边界使三元组损失更加关注相似度分数较高、可能是干净数据样本对的样本。

元训练过程。为确保 MSCN 生成的相似度分数的可靠性，我们使用少量

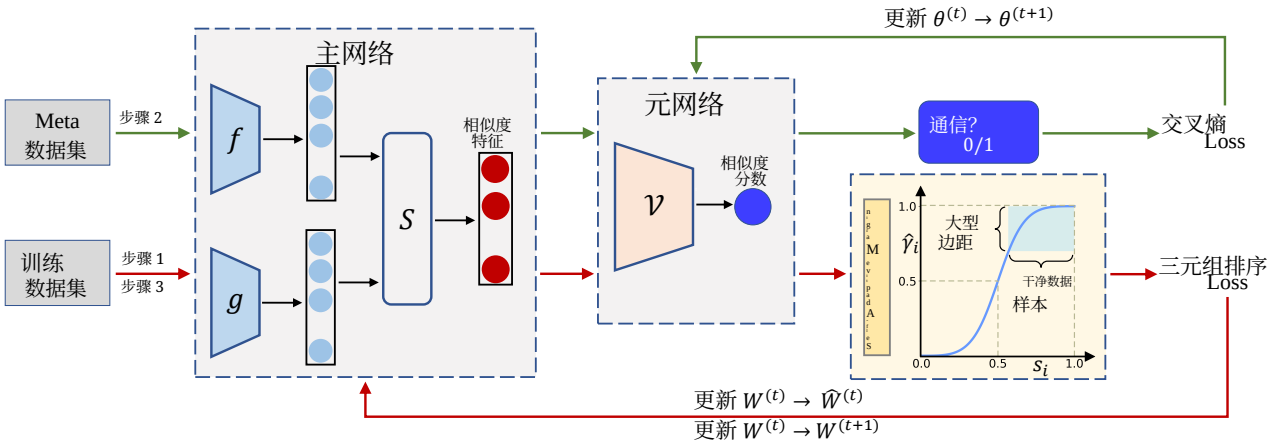

图 3. 所提出的双层优化流程图。在每次迭代中，我们按照式（7）至（9）交替更新主网络和元网络（MSCN）。

干净元数据集 $\mathcal{D}_{meta} = \{(I_i, T_i)\}_{i=1}^{M}$ 来指导训练，其中 M 远小于 N。尽管元学习在近期文献中已得到广泛研究 [10, 32, 46], ，但在我们的目标中构建元过程仍具有挑战性。一方面，主网络提取的相似度特征无法直接用于生成损失。另一方面，元过程应使 MSCN 的输出能够表示相似度分数。为此，我们将一个二分类任务视为元过程：给定一个图文对，训练 MSCN 来判断这两种模态是否相互对应。具体而言，正样本对取自包含对应图像和文本的元数据，即 $(I_i, T_i, y_i = 1)$；负样本对则从训练集中的不同图文对中抽取，以构造不匹配对，即 $(I_i, T_{j\neq i}, y_i = 0)$。二元标签由数据本身生成，可视为指导 MSCN 训练的理想相似度分数。MSCN 的预测衡量一对样本属于干净数据的概率，因此可以自然地等同于相似度分数。具体来说，在每次训练迭代中，我们从训练集中随机构造 M 个不匹配对，以扩充元数据集，得到 $\mathcal{D}'_{meta} = \{(I_i, T_i, y_i)\}_{i=1}^{2M}$。随后，通过最小化在元数据上计算的以下元损失，可以学习到最优参数 Θ^*：

$$\Theta^* = \arg\min_{\Theta} \mathbb{E}_{(I_i,T_i)\in\mathcal{D}'_{meta}}\, l^{meta}(I_i, T_i, y_i), \tag{5}$$

其中，$l^{meta}(\cdot)$ 是交叉熵损失 [35, 37]:

$$l^{meta}(I_i, T_i, y_i) = -y_i \cdot log\mathcal{V}_{\Theta}\left(\mathcal{F}_{W^*(\Theta)}(I_i, T_i)\right). \tag{6}$$

如图 2所示，尽管简单，我们提出的元过程不仅能够促使 MSCN 学习可靠的相似度分数，还具有以下优势：1）以往工作仅使用真实标注数据指导元网络；然而，我们利用正数据和负数据作为元知识，促使 MSCN 从中学习判别能力。2）我们克服了元学习方法的一个局限，即缺乏元数据。我们从训练集中的不同样本对中提取数据，以构建大量负元数据。此外，这更符合噪声对应关系的规律，因为噪声数据的不规则性会呈现出多种模式。

双层优化。 受近期工作的启发 [9, 16, 30, 32], 我们采用双层优化策略，以确保更新主网络和 MSCN 的效率。如图 3所示，该问题实际上通过两个优化嵌套循环求解。具体而言，在训练的每次迭代 t 中，我们采样一个训练对小批量$\{(I_i, T_i)\}_{i=1}^{n}$，其中 n 表示小批量的大小。我们首先沿三元组损失的方向执行一次梯度下降步骤，以更新主网络参数，具体可表示如下：

$$\hat{W}^{(t)}(\Theta) = W^{(t)} - \alpha \sum_{i=1}^{n} \nabla_W l^{train}(I_i, T_i)\big|_{W^{(t)}}, \tag{7}$$

其中 α 表示学习率。更新后的 $\hat{W}^{(t)}$ 可视为最优 $W^*(\Theta)$ 在 t 次迭代时的近似值，其中包含了 Θ的梯度信息。然后，在固定主网络参数的情况下更新 MSCN 的参数。在式（5）中的目标损失和一批元数据 $\{(I_i, T_i, y_i)\}_{i=1}^{m}$的指导下，其中 m 为小批量大小。MSCN 的更新参数计算如下：

$$\Theta^{(t+1)} = \Theta^{(t)} - \beta \frac{1}{m} \sum_{i=1}^{m} \nabla_{\Theta} l^{meta}(I_i, T_i, y_i)\big|_{\Theta^{(t)}}, \tag{8}$$

其中， β 是元过程的学习率。最后，在固定 MSCN 参数的情况下更新主网络的参数。借助新更新的 $\Theta^{(t+1)}$ 生成的校正后相似度分数，我们可以通过梯度下降更新主网络参数 $W^{(t+1)}$：

$$W^{(t+1)} = W^{(t)} - \alpha \sum_{i=1}^{n} \nabla_W l^{train}(I_i, T_i)\big|_{W^{(t)}}. \quad (9)$$

上述过程以交替方式进行，以优化主网络和 MSCN，直到超过最大迭代次数。

3.3. 元知识引导的数据净化

如前所述，我们的MSCN 可以提供校正后的相似度分数以避免噪声。然而，由于三元组损失的性质，即使采用理想的相似度分数（例如，$\mathcal{V}(F(I_i, T_i)) \approx 0 < \mathcal{V}\left(F(I_i^-, T_i)\right) + \hat{}\gamma)$ ，模型仍会对噪声进行拟合。为解决这一问题，我们采用样本选择策略去除噪声数据，仅使用干净样本作为训练数据。以往的工作 [2, 42]受到深度神经网络记忆效应的启发，认为干净样本在训练初期具有较低的损失。不幸的是，三元组排序损失更加复杂，即，同时受到正样本和负样本的影响，从而导致次优的选择结果。基于图 4所示的观察结果，干净样本对和噪声样本对往往可以根据相似度分数加以区分。高斯混合模型（GMM）是一种广泛用于建模混合分布的技术 [15, 20]。然而，我们发现相似度分布向1 呈现高度偏斜，导致 GMM的近似效果较差。因此，我们拟合了一个双分量 Beta 混合模型（BMM） [27] 来建模每对样本的相似度分数 s_i，以更好地区分干净样本和噪声样本：

$$p(s_i) = \sum_{k=1}^{K} \lambda_k \phi(s_i|\alpha_k, \beta_k), \quad (10)$$

其中 $K = 2$ 和 λ_k 是混合系数，而$\phi(s|\alpha_k, \beta_k)$表示参数为 $\alpha_k, \beta_k > 0$的概率密度函数。

我们使用期望最大化 [5] 算法来优化 BMM，这面临着一个关键挑战：模型参数的初始化。现有方法通常采用 K-均值算法 [23] 进行初始化；然而，这种方法耗时（即，时间复杂度为$O(TKN)$，其中 T是迭代次数），且性能不佳。因此，我们提出一种新颖的初始化方法，将元数据作为先验知识。具体而言，我们使用正负元数据拟合两个 Beta 分量，以对干净数据对和噪声数据对的分布进行建模，
分别。给定对应于 $\mathcal{S}_p = \{s_i\}_{i=1}^{M}$ 的相似度分数

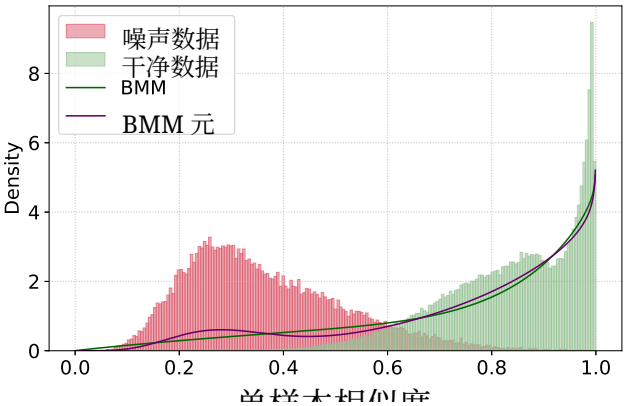

图4. 在 Flickr30K 中，经过 5 轮预热后，20% 对应关系含噪时的逐样本相似度分数分布、估计得到的 BMM，以及结合元先验知识的 BMM。

对应于正元数据的相似度分数后，可通过以下方式初始化干净数据分布建模组件的参数:

$$\alpha_k = \frac{(1 - E(\mathcal{S}_p))\, E(\mathcal{S}_p)^2}{V(\mathcal{S}_p)} - E(\mathcal{S}_p),$$
$$\beta_k = \frac{\alpha_k (1 - E(\mathcal{S}_p))}{E(\mathcal{S}_p)}, \quad (11)$$

其中 $E(\cdot)$ 和 $V(\cdot)$ 分别是均值函数和方差函数。噪声分布分量模型的初始化与之类似，但使用负元数据的相似度分数 $\mathcal{S}_n = \{s_i\}_{i=1}^{M}$。如图 4所示，我们的元知识引导初始化方法近似效果更好，同时时间复杂度更低（即，$O(KM)$）。

最后，我们计算后验概率 $p_i = p(k|s_i) = p(k)p(s_i|\alpha_k, \beta_k)/p(s_i)$，将其作为第 i 个样本为干净数据的概率。净化后的训练集定义如下:

$$\mathcal{D}'_{train} = \{(I_i, T_i) \in \mathcal{D}_{train} | p_i > 0.5\}. \quad (12)$$

在每次迭代中，我们都会在训练元学习目标之前生成净化后的数据集。此外，按照 [12, 15, 20],，我们维护两个网络 $\{\mathcal{F}_W^1, \mathcal{V}_\Theta^1\}$ 和$\{\mathcal{F}_W^2, \mathcal{V}_\Theta^2\}$，以避免误差累积：一个网络生成净化后的训练集，用于训练另一个网络。完整算法见补充材料。

4. 实验

4.1. 数据集和评估协议

数据集。我们在三个标准图文检索数据集上评估了我们的方法。具体而言，Flickr30K[43] 从 Flickr 网站收集了 31,783 张图像，每张图像对应 5 条字幕。按照 [15], 中的划分，我们使用 1,000 个图文对作为验证集，1,000 个图文

对作为测试集，其余部分作为训练集。MS-COCO[24] 包含 123,287 张图像，每张图像对应 5 条字幕。按照 [15] 的划分方式，我们将 MS-COCO 拆分为 113,287 个图文对作为训练集、5,000 个图文对作为验证集，以及 5,000 个图文对作为测试集。Conceptual Captions [31]是一个大规模数据集，包含 330 万张图像，每张图像对应一条字幕。我们采用了遵循 [15], 的相同子集，即CC152K。具体而言，CC152K 包含训练集中的 150,000 个图文对作为训练数据、验证集中的 1,000 个图文对作为验证数据，以及验证集中的 1,000 个图文对作为测试数据。此外，我们还使用了一个额外的元数据集，其中包含验证集中的 3,000 个图文对。

评估协议。 我们使用 K 值召回率（R@K）指标评估检索性能。简而言之，R@K 是在与查询最接近的 K 个项目中检索到的相关项目所占的比例。在实验中，我们分别以图像和文本作为查询，并报告 R@1、R@5 和 R@10 结果，以进行全面评估。遵循 [15], , 我们在推理阶段对两个网络的相似度分数取平均。

4.2. 实现细节

作为一种通用方法，我们的方法可以轻松应用于几乎所有跨模态检索方法，以提升鲁棒性。为进行公平比较，我们采用了与 NCR [15], 即相同的网络骨干，即投影模块和相似度函数。对于所有数据集，训练过程在经过 5 个 epoch 的预热后进行 50 个 epoch。我们使用 ADAM [17]作为主网络和元网络的优化器，批量大小设为 64。为拟合 BMM，我们将 EM过程的停止阈值设为 10^{-2}，最大迭代次数设为 10。需要注意的是，为了稳定拟合，我们将相似度分数截断到 $[10^{-4}, 1-10^{-4}]$ 范围内，而不是 [0,1] 。此外，我们将边界参数设为 0。2，并将 $\tau = 2$ 用于计算自适应边界。对于所有数据集，真实元数据的数量约占训练数据的 2%，主网络和元网络的学习率分别初始化为 2×10^{-4} 和 1.7×10^{-5}。对于包含真实噪声的 CC152K，由于缺少真实数据，我们使用验证数据作为元数据集。对于所有实验，我们在 30 个 epoch 后将学习率衰减 0。1。

4.3. 与最先进方法的比较

在本节中，我们将所提方法与 7 种最先进方法进行对比实验，包括 SCAN [18], VSRN[22], IMRAM [4], SGR、SAF、SGRAF [7], 和 NCR [15]。由于 Flick30K 和 MS-COCO 是标注良好的数据集，我们通过随机打乱训练图像和描述来生成合成噪声对应关系，并针对特定噪声比例开展实验，设置三种不同的

	图像到文本			
方法	R@1	R@5	R@10	SUM
SCAN（ECCV'18）	30.5	55.3	65.3	151.1
VSRN（ICCV'19）	32.6	61.3	70.5	164.4
IMRAM（CVPR'20）	33.1	57.6	68.1	158.8
SAF (AAAI'21)	31.7	59.3	68.2	159.2
SGR (AAAI'21)	11.3	29.7	39.6	80.6
SGR* (AAAI'21)	35.0	63.4	73.3	171.7
NCR(NIPS'21)	39.5	64.5	73.5	177.5
MSCN*（我们的方法）	39.7	65.4	75.3	180.4
MSCN（我们的方法）	**40.1**	**65.7**	**76.6**	**182.4**
	文本到图像			
SCAN（ECCV'18）	26.9	53.0	64.7	144.6
VSRN(ICCV'19)	32.5	59.4	70.4	162.3
IMRAM(CVPR'20)	29.0	56.8	67.4	153.2
SAF（AAAI'21）	31.9	59.0	67.9	158.8
SGR（AAAI'21）	13.1	30.1	41.6	84.8
SGR*（AAAI'21）	34.9	63.0	72.8	170.7
NCR(NIPS'21)	40.3	64.6	73.2	178.1
MSCN*（我们的方法）	39.8	66.1	75.0	180.9
MSCN（我们的方法）	**40.6**	**67.4**	**76.3**	**184.3**

表 1. 在含真实噪声的 CC152K 上的图文检索性能，最佳结果以**粗体**突出显示。

噪声比例，即., 分别为 20%、50% 和 70%。对于所有方法，我们选择在验证集上表现最佳的检查点，并报告其在测试集上的性能。遵循 NCR 的做法，我们还报告了两种基于 SGR 的强基线方法，即 SGR-C 和 SGR*。简而言之，SGR-C 仅使用干净数据进行训练，而 SGR* 则采用预训练过程，并在训练时不使用困难负样本，以提高鲁棒性。为了在真实世界的噪声数据上评估我们的方法，我们在 CC152K 上开展实验。该数据集采集自互联网，包含约 3% ∼ 20 % 的噪声对应关系 [31]。

CC152K 上的实验。 表 1 展示了在具有真实噪声对应关系的 CC152K 上的结果。在 CC152K 上，我们的 MSCN在所有指标方面都达到了最先进的性能。具体而言，MSCN 的总分在文本检索和图像检索方面分别比最佳基线方法高 4.9% 和 6.2 %。此外，我们还展示了一个更严格的比较结果（记为 MSCN*），该结果仅使用 1,000 条元数据（约占全部数据的 0.67% ）训练我们的方法。可以看出，我们的 MSCN* 也取得了具有竞争力的性能。具体而言，除图像检索的 R@1 外，其余所有指标均优于基线方法。

在 Flickr30K 和 MS-COCO 上的实验。表2展示了在一系列合成噪声率下，Flickr30K 和 MS-COCO 上的结果。按照 [15], , 我们报告

		Flickr30K						MS-COCO					
		图像到文本			文本到图像			图像到文本			文本到图像		
噪声	方法	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
20%	SCAN	59.1	83.4	90.4	36.6	67.0	77.5	66.2	91.0	96.4	45.0	80.2	89.3
	VSRN	58.1	82.6	89.3	40.7	68.7	78.2	25.1	59.0	74.8	17.6	49.0	64.1
	IMRAM	63.0	86.0	91.3	41.4	71.2	80.5	68.6	92.8	97.6	55.7	85.0	91.0
	SAF	51.0	79.3	88.0	38.3	66.5	76.2	67.3	92.5	96.6	53.4	84.5	92.4
	SGR*	62.8	86.2	92.2	44.4	72.3	80.4	67.8	91.7	96.2	52.9	83.5	90.1
	SGR-C	72.8	90.8	95.4	56.4	82.1	88.6	75.4	95.2	97.9	60.1	88.5	94.8
	NCR	75.0	93.9	97.5	58.3	83.0	89.0	77.7	95.5	98.2	62.5	89.3	95.3
	MSCN	**77.4**	**94.9**	**97.6**	**59.6**	**83.2**	**89.2**	**78.1**	**97.2**	**98.8**	**64.3**	**90.4**	**95.8**
50%	SCAN	27.7	57.6	68.8	16.2	39.3	49.8	40.8	73.5	84.9	5.4	15.1	21.0
	VSRN	14.3	37.6	50.0	12.1	30.0	39.4	23.5	54.7	69.3	16.0	47.8	65.9
	IMRAM	9.1	26.6	38.2	2.7	8.4	12.7	21.3	60.2	75.9	22.3	52.8	64.3
	SAF	30.3	63.6	75.4	27.9	53.7	65.1	30.4	67.8	82.3	33.5	69.0	82.8
	SGR*	36.9	68.1	80.2	29.3	56.2	67.0	60.6	87.4	93.6	46.0	74.2	79.0
	SGR-C	69.8	90.3	94.8	50.1	77.5	85.2	71.7	94.1	97.7	57.0	86.6	93.7
	NCR	72.9	93.0	**96.3**	54.3	79.8	86.5	74.6	94.6	97.8	59.1	87.8	94.5
	MSCN	**74.4**	**93.2**	96.0	**55.3**	**80.4**	**86.8**	**77.5**	**96.2**	**98.7**	**60.7**	**89.1**	**94.9**
70%	SCAN	5.6	19.3	27.4	2.2	8.0	12.8	18.1	43.1	57.4	0.3	1.3	2.3
	VSRN	0.8	2.5	4.1	0.5	1.5	2.7	5.1	15.7	24.6	2.5	8.8	13.3
	IMRAM	1.3	3.1	3.9	0.3	1.2	2.8	7.1	20.0	33.4	5.3	15.2	22.0
	SAF	0.5	2.2	3.0	0.2	0.8	1.7	0.1	1.7	4.0	0.6	1.9	3.0
	SGR*	17.9	42.1	51.9	14.6	31.0	40.8	35.7	71.2	85.4	31.6	65.8	79.0
	SGR-C	65.0	89.3	**94.7**	48.1	**74.5**	**81.1**	69.8	93.6	97.5	56.5	86.0	93.4
	NCR	16.1	38.5	52.8	11.0	29.5	41.4	35.4	69.5	83.4	31.5	66.4	81.1
	MSCN	**69.0**	**89.3**	93.0	**49.2**	73.1	79.0	**74.4**	**94.9**	**97.7**	**58.8**	**87.2**	**93.7**

表 2. 在 Flickr30K 和 MS-COCO 1K 上，合成噪声率为 20%、50% 和 70% 时的图文检索性能，最佳结果以**粗体**突出显示。

结果为对 5 个折中的 1K 测试图像取平均值。根据表 2，我们可以观察到，噪声对应关系会显著影响跨模态检索方法的性能。随着噪声率增加，检索准确率会迅速下降。在不同噪声率下，我们的 MSCN 在几乎所有指标上都优于基线方法。在低噪声情况下，即，20% 和 50% 时，与 NCR 相比，我们的 MSCN 在四种评估中将 R@1（图像和文本之和）分别提高了 3.7%、2.2%、2.5% 和 4.5%。在高噪声情况下，即，70% 时，所有基线方法都很容易对噪声对应关系过拟合，导致性能不佳。与 NCR 相比，我们的 MSCN 在这四种评估中将 R@1 分别提高了 52.9%、38.2%、39.4% 和 27.3%。即使与仅使用干净数据训练的 SGR-C 相比，我们的 MSCN 也取得了有竞争力的性能。

4.4. 消融研究

为评估所提出组件（即、自适应边界和数据净化）的性能，我们在带有 20% 噪声对应关系的 Flickr30K 上进行了消融研究。请注意，对于不带 $\mathcal{D}'_{train}$ 的 MSCN，我们仅使用单个模型进行实验。如表 3 所示，我们可以看到，不带 $\hat{}\gamma$ 或 $\mathcal{D}'_{train}$ 的性能都不如完整的 MSCN，表明所有组件对于取得有利结果都很重要。我们将在补充材料中进一步解释数据选择策略。

方法			图像到文本		
MSCN	不含$\hat{\gamma}$	不含 $\mathcal{D}'_{train}$	R@1	R@5	R@10
✓			**77.4**	**94.9**	**97.6**
✓	✓		75.3	94.5	97.2
✓		✓	75.8	93.8	96.2
✓	✓	✓	74.1	91.5	94.7
			文本到图像		
✓			**59.6**	**83.2**	**89.2**
✓	✓		58.3	83.1	88.9
✓		✓	55.8	74.5	77.4
✓	✓	✓	53.4	71.2	72.2

T表 3. Flickr30K 上 20% 噪声条件下的消融研究　率。

4.5. 渐进式比较

图。6 绘制了随着训练进行，MSCN 和 NCR 在测试集上的平均召回率。我们展示了一个使用含 50% 噪声对应关系的 Flickr30K 得到的代表性结果。从结果可以看出，我们的 MSCN 在整个训练过程中表现更加稳定。尽管 NCR



图 5. 我们的 MSCN 检测到的真实世界噪声示例。

无需额外的元数据集来指导训练，而是依赖干净数据验证集来选择最佳模型参数。这种稳定性也表明，我们的方法更好地减轻了噪声对应关系的干扰。

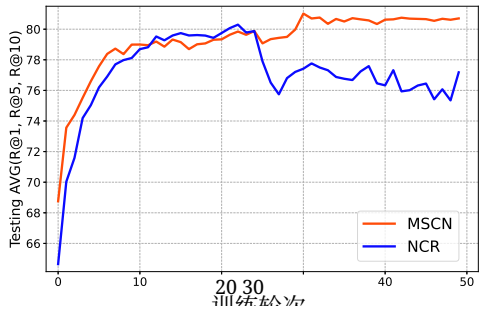

图 6. 随着训练的进行，MSCN 和 NCR 在测试集上的渐进式性能比较。

4.6. 检测到的噪声样本

图 5 展示了 MSCN 在 Conceptual Captions 中检测到的一些真实噪声图文对（相似度分数较低）。我们还展示了 MSCN 预测的相似度分数。对于大多数噪声图文对，图像和文本包含完全无关的语义信息。此外，MSCN 甚至能够发现噪声图文对中的细微错误。最后一对中，图片显示一名男子捂着耳朵，却被错误地描述为正在使用手机。尽管这张图片具有误导性，我们的方法仍成功检测出了这一细微差异。

4.7. 相似度分数可视化

图 7绘制了 MSCN 对干净数据和噪声训练样本预测的相似度分数分布。可以看出，几乎所有较大的相似度分数都属于干净图文对，而噪声图文对的分数低于干净样本，这表明训练后的 MSCN 能够成功提供经过校正的相似度分数。

5. 结论

在本文中，我们探索了一种元学习方法，以解决带噪声对应关系学习的问题——

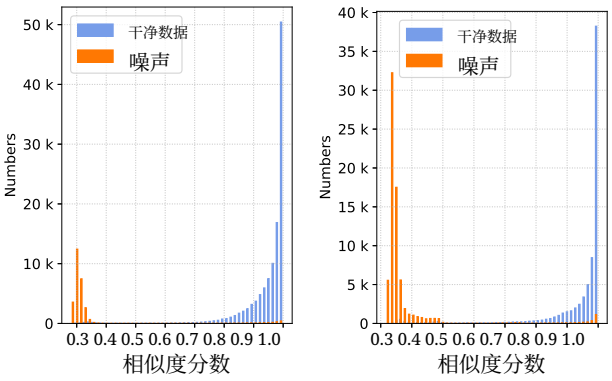

图 7. Flickr30K 训练数据中逐样本相似度分布。(a) 20% 噪声对应关系。(b) 50% 噪声对应关系。

具体而言，我们提出了一种元校正网络（MSCN），用于提供可靠的相似度分数。我们的 MSCN 通过一种新颖的元过程进行训练，该过程将正数据和负数据都视为元知识，以促进 MSCN 从中学习判别能力。为进一步减轻噪声干扰，我们设计了一种数据净化策略，利用元数据作为先验知识，高效地净化噪声数据。我们在三个广泛使用的数据集上开展了全面实验。结果验证了我们的方法在合成噪声和真实噪声下的有效性。

致谢

本研究得到国家重点研发计划（2020AAA0108800）、国家自然科学基金（62192781、62272374、62202367、62250009、62137002、61937001）、国家自然科学基金创新研究群体项目（61721002）、教育部创新研究团队项目（IRT 17R86）、中国工程科技知识中心项目以及中国工程院“面向‘一带一路’MOOC中国培训的线上线下混合教育服务系统”项目的资助。我们谨向王宽诚教育基金会的支持表示感谢。

参考文献

[1] Relja Arandjelovic 和 Andrew Zisserman。能够发声的物体。载于《欧洲计算机视觉会议》（ECCV）论文集，第 435–451 页，2018 年。1[2] Eric Arazo、Diego Ortego、Paul Albert、Noel O'Connor 和 Kevin McGuinness。无监督标签噪声建模与损失校正。载于《国际机器学习大会》论文集，第 312–321 页。PMLR，2019 年。5[3] Devansh Arpit、Stanislaw K Jastrzebski、Nicolas Ballas、David Krueger、Emmanuel Bengio、Maxinder S Kanwal、Tegan Maharaj、Asja Fischer、Aaron C Courville、Yoshua Bengio 等。深入探究深度网络中的记忆现象。载于 ICML，2017 年。1[4] Hui Chen、Guiguang Ding、Xudong Liu、Zijia Lin、Ji Liu 和 Jungong Han。IMRAM：用于跨模态图文检索的递归注意力记忆迭代匹配。载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 12655–12663 页，2020 年。6[5] Arthur P Dempster、Nan M Laird 和 Donald B Rubin。通过 EM 算法从不完整数据中进行最大似然估计。《英国皇家统计学会会刊：B 辑（方法论）》，39(1)：1–22，1977 年。5[6] Cheng Deng、Zhaojia Chen、Xianglong Liu、Xinbo Gao 和 Dacheng Tao。用于跨模态检索的基于三元组的深度哈希网络。《IEEE 图像处理汇刊》，27(8)：3893–3903，2018 年。1[7] Haiwen Diao、Ying Zhang、Lin Ma 和 Huchuan Lu。图文匹配的相似性推理与筛选。载于 AAAI 人工智能会议论文集，第 35 卷，第 1218–1226 页，2021 年。2, 6[8] Fartash Faghri、David J Fleet、Jamie Ryan Kiros 和 Sanja Fidler。VSE++：利用困难负样本改进视觉语义嵌入。arXiv 预印本 arXiv:1707.05612，2017 年。2, 3[9] Chelsea Finn、Pieter Abbeel 和 Sergey Levine。用于深度网络快速适应的模型无关元学习。载于《国际机器学习大会》论文集，第 1126–1135 页。PMLR，2017 年。2, 4[10] Luca Franceschi、Paolo Frasconi、Saverio Salzo、Riccardo Grazzi 和 Massimiliano Pontil。用于超参数优化与元学习的双层规划。载于《国际机器学习大会》论文集，第 1568–1577 页。PMLR，2018 年。4[11] Dalu Guo、Chang Xu 和 Dacheng Tao。用于视觉对话的图像—问题—答案协同网络。载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 10434–10443 页，2019 年。1[12] Bo Han、Quanming Yao、Xingrui Yu、Gang Niu、Miao Xu、Weihua Hu、Ivor Tsang 和 Masashi Sugiyama。协同教学：利用极高噪声标签进行鲁棒的深度神经网络训练。《神经信息处理系统进展》，31，2018 年。5[13] Haochen Han、Qinghua Zheng、Minnan Luo、Kaiyao Miao、Feng Tian 和 Yan Chen。面向音视动作识别的噪声容忍学习。arXiv 预印本 arXiv:2205 07611，2022 2 。 .[14] Peng Hu、Xi Peng、Hongyuan Zhu、Liangli Zhen 和 Jie Lin。利用噪声标签学习跨模态检索。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第5403–5413页，2021年。2[15] Zhenyu Huang、Guocheng Niu、Xiao Liu、Wenbiao Ding、Xinyan Xiao、Hua Wu和Xi Peng。利用噪声对应关系进行跨模态匹配学习。《神经信息处理系统进展》，34:29406–29419，2021年。1, 2, 5, 6[16] Haoxuanye Ji、Le Wang、Sanping Zhou、Wei Tang、Nanning Zheng和Gang Hua。用于无监督行人重识别的元成对关系蒸馏。载于《IEEE/CVF国际计算机视觉会议论文集》，第3661–3670页，2021年。4[17] Diederik P Kingma和Jimmy Ba。Adam：一种随机优化方法。载于国际学习表征会议（海报），2015年。6[18] Kuang-Huei Lee、Xi Chen、Gang Hua、Houdong Hu和Xiaodong He。用于图文匹配的堆叠交叉注意力。载于《欧洲计算机视觉会议论文集》（ECCV），第201–216页，2018年。2, 6[19] Da Li、Yongxin Yang、Yi-Zhe Song和Timothy Hospedales。学习泛化：用于领域泛化的元学习。载于《AAAI人工智能会议论文集》，第32卷，2018年。2[20] Junnan Li、Richard Socher和Steven CH Hoi。DivideMix：将带噪声标签的学习作为半监督学习。载于《国际学习表征会议》，2019年。1, 2, 5[21] Junnan Li、Yongkang Wong、Qi Zhao和Mohan S Kankanhalli。学习如何从带噪声标签的数据中学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第5051–5059页，2019年。2[22] Kunpeng Li、Yulun Zhang、Kai Li、Yuanyuan Li和Yun Fu。用于图文匹配的视觉语义推理。载于《IEEE/CVF国际计算机视觉会议论文集》，第4654–4662页，2019年。2, 6[23] Aristidis Likas、Nikos Vlassis和Jakob J Verbeek。全局k均值聚类算法。《模式识别》，36(2):451–461，2003年。5[24] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar和C Lawrence Zitnick。Microsoft COCO：上下文中的常见物体。载于《欧洲计算机视觉会议论文集》，第740–755页。施普林格，2014年。6[25] Chunxiao Liu、Zhendong Mao、Tianzhu Zhang、Hongtao Xie、Bin Wang和Yongdong Zhang。用于图文匹配的图结构网络。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第10921–10930页，2020年。2[26] Lin Ma、Zhengdong Lu和Hang Li。使用卷积神经网络学习回答图像相关问题。载于第三十届AAAI人工智能会议，2016年。1[27] Zhanyu Ma和Arne Leijon。利用变分推断对贝塔混合模型进行贝叶斯估计。《IEEE模式分析与机器智能汇刊》，33(11):2160–2173，2011年。5[28] Nikhil Mishra、Mostafa Rohaninejad、Xi Chen和Pieter Abbeel。一种简单的神经注意力元学习器。载于《国际学习表征会议》，2018年。2

[29] Yang Qin、Dezhong Peng、Xi Peng、Xu Wang和Peng Hu。利用噪声对应关系进行深度证据学习，以实现跨模态检索。载于第30届ACM多媒体国际会议论文集，第4948 4956页，2022年 2 –。[30] Mengye Ren、Wenyuan Zeng、Bin Yang和Raquel Urtasun。学习为样本重新赋权，以实现鲁棒深度学习。载于《国际机器学习大会》，第4334–4343页。PMLR，2018年。4[31] Piyush Sharma、Nan Ding、Sebastian Goodman和Radu Soricut。Conceptual Captions：用于自动图像描述的清理后、添加上位词的图像替代文本数据集。载于《计算语言学协会第56届年会论文集》（第1卷：长篇论文），第2556–2565页，2018年。1, 6[32] Jun Shu、Qi Xie、Lixuan Yi、Qian Zhao、Sanping Zhou、Zongben Xu和Deyu Meng。Meta-weight-net：学习用于样本加权的显式映射。《神经信息处理系统进展》，32，2019年。2, 4[33] Didac Sur´ís、Amanda Duarte、Amaia Salvador、Jordi Torres和Xavier Gir´o-i Nieto。用于视频和音频检索的跨模态嵌入。载于《欧洲计算机视觉会议》（ECCV）研讨会论文集，第0 0页，2018年 1 –。[34] Jun Tang、Ke Wang和Ling Shao。用于跨模态检索的监督式矩阵分解哈希。《IEEE图像处理汇刊》，25(7):3157–3166，2016年。2[35] Bokun Wang、Yang Yang、Xing Xu、Alan Hanjalic和Heng Tao Shen。对抗式跨模态检索。载于第25届ACM多媒体国际会议论文集，第 154 162, 2017 4 –。[36] Liwei Wang、Yin Li和Svetlana Lazebnik。学习保持结构的深度图文嵌入。载于《IEEE计算机视觉与模式识别会议论文集》，第5005–5013页，2016年。2[37] Yunchao Wei、Yao Zhao、Canyi Lu、Shikui Wei、Luoqi Liu、Zhenfeng Zhu和Shuicheng Yan。利用CNN视觉特征进行跨模态检索：一种新的基线方法。《IEEE控制论汇刊》，47(2):449–460，2016年。4[38] Xing Xu、Li He、Huimin Lu、Lianli Gao和Yanli Ji。用于跨模态检索的深度对抗度量学习。《万维网》，22(2):657–672，2019年。2[39] Mouxing Yang、Zhenyu Huang、Peng Hu、Taihao Li、Jiancheng Lv和Xi Peng。利用双重噪声标签学习进行可见光-红外行人重识别。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第14308–14317页，2022年。2[40] Mouxing Yang、Yunfan Li、Peng Hu、Jinfeng Bai、Jiancheng Lv和Xi Peng。利用不完整信息进行鲁棒多视图聚类。《IEEE模式分析与机器智能汇刊》，45(1):1055–1069，2022年。2[41] Mouxing Yang、Yunfan Li、Zhenyu Huang、Zitao Liu、Peng Hu和Xi Peng。利用鲁棒噪声对比损失进行部分视图对齐的表示学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第1134–1143页，2021年。2[42] Quanming Yao、Hansi Yang、Bo Han、Gang Niu和James Tin-Yau Kwok。探索如何利用记忆效应噪声标签学习。在 国际机器学习大会，第 10789–10798 页。PMLR，2020。 5[43] Peter Young、Alice Lai、Micah Hodosh 和 JuliaHockenmaier。从图像描述到视觉指称：用于事件描述语义推理的新相似度度量。计算语言学协会汇刊，2：67–78，2014。 5[44] Qi Zhang、ZhenLei、Zhaoxiang Zhang和 Stan Z Li。用于图文检索的上下文感知注意力网络。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 3536–3545 页，2020 年。 1[45] Liangli Zhen、PengHu、XuWang 和 Dezhong Peng。深度监督跨模态检索。载于 *IEEE/CVF*计算机视觉与模式识别会议论文集，第 10394–10403 页，2019 年。 2[46] Guoqing Zheng、AhmedHassanAwadallah 和 Susan Dumais。噪声标签学习的元标签校正。载于*AAAI*人工智能会议论文集，第 35 卷，第 11053–11061 页，2021 年。 2，4