基于文本的人物检索的自适应不确定性学习

Li，何晨，徐星，沈福民，杨洋，沈恒涛
中国电子科技大学计算机科学与工程学院、未来媒体研究中心
lishenshen727@gmail.com，陈he1219@outlook.com, xing.xu@uestc.edu.cn, fumin.shen@gmail.com, yang.yang@uestc.edu.cn, shenhengtao@hotmail.com

摘要

基于文本的人物检索旨在基于文本描述从图库中检索特定的行人图像。主要的挑战是如何在类内差异显著、类间差异极小的情况下克服固有的异质模态差距。现有的方法通常采用视觉语言预训练或注意机制，从噪声输入中学习适当的跨模态对齐。尽管取得了值得称道的进展，但目前的方法不可避免地存在两个缺陷:i)匹配歧义，这主要源于不可靠的匹配对;2)片面的跨模态对齐，源于缺乏对一对多对应的探索，即粗粒度语义对齐。这些关键问题会显著降低检索性能。为此，我们提出了一种新的框架，称为基于不确定性的自适应不确定性学习(AUL)，用于基于文本的个体检索。具体来说，我们的AUL框架由三个关键部分组成:1)不确定感知匹配过滤，利用主观逻辑有效地减轻不可靠匹配对的干扰，选择高置信度的跨模态匹配进行训练;2)基于不确定性的对齐细化，它不仅通过构建不确定性表示来模拟粗粒度对齐，而且还执行渐进式学习，以适当地合并粗粒度和细粒度对齐;3)跨模态掩模建模(Cross-modal mask Modeling)，旨在探索视觉和语言之间更全面的关系。大量的实验表明，我们的AUL方法在监督、弱监督和域泛化设置的三个基准数据集上始终如一地实现了最先进的性能。我们的代码可在https://github.com/CM-MSG/Code-AUL上获得。

I介绍

基于文本的人物检索任务(Li等人，2017;Ding et al. 2021)旨在通过提供的文本描述查询从候选集合中定位特定的行人图像。与传统的基于图像的人检索(Specker, Cormier, and Beyerer 2023)或基于视频的人检索(Hou et al. 2021)相比，基于文本的人检索的查询提供了一个易于访问和直观的

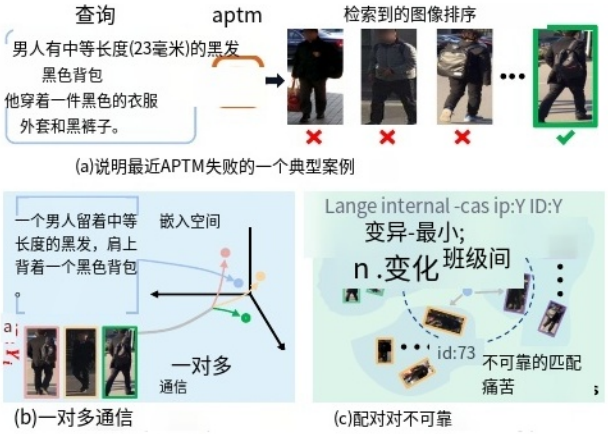

图1。存在问题的说明性例子:(a)最新方法APTM的代表性失效案例。(b)一对多对应的存在是显而易见的。(c)不可靠的匹配对源于较大的类内差异和最小的类间差异。

描述目标人物属性的手段，使其成为一个流行和活跃的研究领域。然而，由于类内差异很大，类间差异很小，基于文本的人物检索任务在克服固有的异质神经模态差距方面面临越来越大的挑战(Jiang等人，2022;Li等人，2023a)，这大大阻碍了整体检索性能。

为了解决上述问题，以前的方法主要是通过利用补充信息(Zhu et al. 2021)或结合多种注意机制(Suo et al. 2022)来关注跨模态对齐。此外，最近的方法(Jiang和Ye 2023)受到视觉语言预训练能力的影响，以增强表征学习，这可以有效地表征视觉和语言之间的关联。

如图1(a)所示，虽然上述方法取得了进展，但它们仍然存在匹配模糊和片面跨模态对齐的问题，从而导致性能下降和泛化受限。这两个问题可以归结为以下几个方面:1)缺乏一对多对应关系，这源于只考虑一对一的限制

"Corresponding Author

一个匹配。如图1(b)所示，语言和视觉之间确实存在一对多对应关系。具体来说，视觉数据可以完全捕获所有对象，但缺乏相应文本中的上下文，并且语言不能基于人类注释的标题完全描述场景的每个细节。这种固有的性质导致了探索视觉和语言之间一对多对应关系的必要性。2)不可靠的匹配对，主要来自于类内差异较大和类间差异较小所引入的固有数据噪声。如图1(c)所示，仅仅基于相似性来选择跨模态匹配可能是不合适的。这是因为某些负样本可能会因为与目标图像的相似性而被错误地识别为ground truth。这些问题从不同的角度共同降低了跨模态匹配的准确性。

基于上述观察，我们提出了一种新的框架，称为自适应不确定性学习(AUL)，用于从不确定性角度进行基于文本的人物检索。具体而言，如图2所示，它由三个关键组成部分组成:1)不确定性感知匹配过滤(UMF)，它最初采用主观逻辑理论(Jesang 2016)对不确定性建模，以测量匹配模糊程度，称为匹配不确定性。随后，它利用这种不确定性自适应地为每个训练对分配权重，旨在防止匹配模糊性的影响，并在模型学习过程中选择高置信度的跨模态匹配。2)基于不确定性的对齐优化(UAR)，它不仅通过构建不确定性表示来探索一对多对应关系，而且还通过渐进式学习来适当地整合粗粒度和细粒度对齐。请注意，一对多对应的概念可以比作粗粒度对齐。这个模块巧妙地解决了一对多对应的不足，并通过一种易学难学的方法，引导模型逐渐获得更全面的对齐。3)跨模态掩模建模(Cross-modal mask Modeling, CMM)，设计具有跨模态交互的掩模信号建模，有效挖掘图像和文本之间的细粒度关系。我们在三个广泛使用的基于文本的人物检索基准(即CUHK-PEDES, ICG-PEDES和RSTPReid)上评估了我们的方法。实验结果表明，我们的AUL方法在监督、弱监督和领域泛化设置方面明显优于最近最先进的方法。

我们的主要贡献可以概括如下:

- 在仔细考虑匹配不确定性的基础上，设计了一种不确定性感知匹配过滤策略，利用主观逻辑自适应选择高置信度的跨模态匹配，减轻不可靠匹配对对训练的干扰。我们提出了一个基于不确定性的对齐优化模块，该模块不仅通过构造不确定性表示来模拟粗粒度对齐，而且可以逐步组织多粒度对齐。
- 我们部署了一个跨模态掩模建模模块，通过全面的跨模态交互来重建图像和文本模态信号，进一步探索两种模态之间的对应关系。

相关工作

基于文本的人物检索。基于文本的人物检索的目标是根据提供的文本准确地识别目标行人。(Li et al. 2017)首先介绍了该任务，并发布了开创性数据集CUHK-PEDES。大多数后续方法(Niu et al. 2020;牛，黄，王2020;郑等人2020;Jing et al. 2020;Ding 等 人 。2021;Suo 等 。2022;Farooq等人，2022;Wang等人。2020;Aggarwal, Babu, and Chakraborty 2020)在很大程度上依赖于注意力模块或补充信息来实现有效的跨模态对齐。例如，(Wu et al. 2021)采用颜色推理来获得信息语义。(Farooq等人，2022;Li等人(202b)设计了一个统一的多层网络，从图像和文本模式中动态提取全局和局部级别的语义。此外，最近的方法(Li等人，2023c;Jiang and Ye 2023)逐渐使用视觉语言预训练模型或基于外部知识的预训练来增强对齐能力。(Yan et al. 2022a)有效地利用CLIP (Radford et al. 2021)的优势进行基于文本的人物检索，而(Jiang and Ye 2023)通过对其构建的数据集进行预训练来提高检索性能。然而，他们忽略了由不可靠的匹配对引起的匹配不确定性的干扰，这构成了我们工作的动机。

基于不确定性的学习。为了解决量化预测置信度的挑战，基于不确定性的学习已经成为一种很有前途的方法。(Kendall and Gal 2017)将不确定性分为两种不同的类型:认知不确定性和任意不确定性。为了解决认知的不确定性，一些研究将贝叶斯网络(Gal and Ghahramani 2016)和主观逻辑(Jesang 2016)与Dempster-Shafer证据理论(Yager and Liu 2008)结合起来，旨在了解权重的分布，而不是直接获得具体权重。在任意不确定性方面，先前的研究已经在各个领域进行了探索，例如图像检索(Warburg et al. 2021)和分割(Zheng and Yang 2021)。与此不同的是，我们提出了一种基于自适应不确定性的学习框架，通过构建基于不确定性的表示来克服匹配不确定性的干扰，挖掘一对多对应关系。

屏蔽信号建模。(He et al. 2022;Vaswani等人，2017;Jiang等人(2023)提出屏蔽信号建模是各种视觉和语言任务中常用的组件，包括掩模语言建模(MLM)和掩模图像建模(MIM)。例如，(Vaswani等人。2017)在广泛的自然语言处理任务中验证了传销的泛化性。(He et al. 2022)预测蒙面像素以细化视觉表征。由于其卓越的性能，MLM和MIM在我们的任务中也发挥了重要作用。(Jiang和Ye 2023)介绍了一种根据未掩码文本预测掩码文本标记的方法

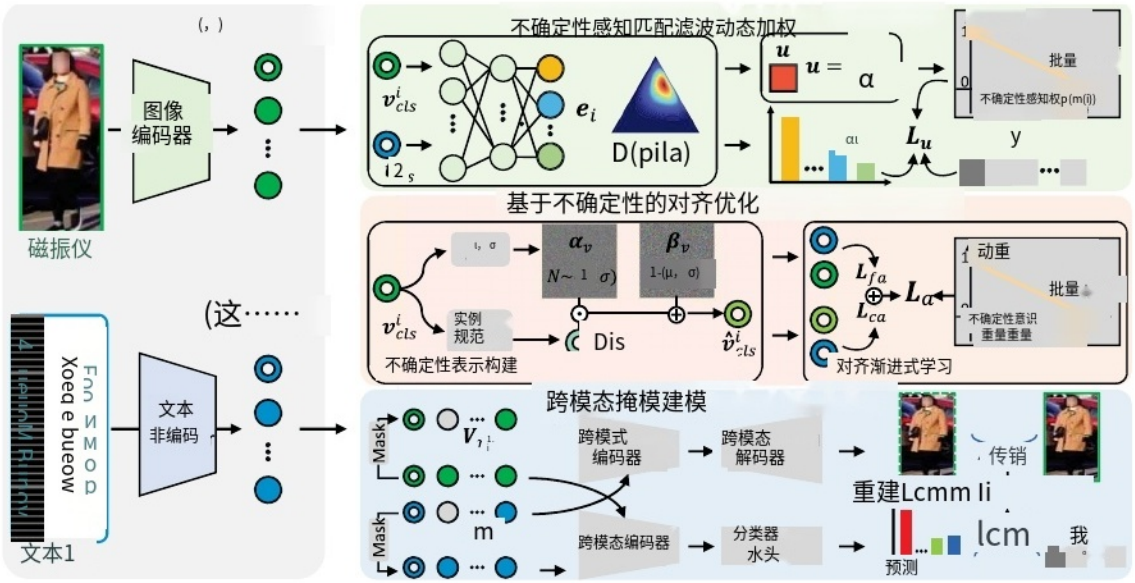

图2。我们提出的AUL方法的总体框架。它由三个关键部分组成:1)不确定性感知匹配滤波(UMF);2)基于不确定性的对齐优化(UAR);3)跨模态掩模建模。

以及视觉标记。在本文中，我们通过同时考虑情态语义来预测掩码文本和视觉标记。

我们的AUL方法

初步

基于文本的人物检索任务的目标是在提供的文本查询的指导下，从候选图库中识别和检索最相似的人物图像。为了获得正确的行人，我们提出的框架侧重于通过学习文本描述和相应人物图像之间存在的相似性来促进准确的对齐。形式上，我们将(I, T)定义为训练数据集中的图像-文本对。每对由一个人物图像I和它对应的文本描述Ti组成。我们首先将图像I输入到图像编码器中，产生一系列视觉特征(vs,v, v)，其中vils作为全局视觉特征，{v。，v表示视觉贴片特征。此外，我们利用文本编码器获得一系列文本表示(tels,t, t)，其中tfs和ft，…，t)表示全局文本特征和令牌特征。

不确定性感知匹配过滤

主观逻辑的背景。主观逻辑(SL)提供了Dempster-Shafer (Yager and Liu 2008)理论的不确定性原理在识别框架内的形式化表示，建模为Dirich-let分布。因此，它提供了在严格建立的理论框架内采用SL理论原理量化不确定性的手段。具体来说，我们首先获得证据向量e;预测为第i个单例。然后我们对每个单子的不确定性u和信念质量p = (p1)进行建模，其可表述如下:

$$p_k = \frac{e_k}{S}, \quad u = \frac{N}{S}, \tag{1}$$

其中S=(ek + 1)可以认为是狄利克雷分布的强度，相信概率pk对应于相应的Dirich-let分布的参数a =(ek + 131)。注意，不确定性u与总证据呈反比关系。最后，以a为特征的狄利克雷分布可定义为:

$$D(\mathbf{p}|\boldsymbol{\alpha}) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{N} p_j^{\alpha_j - 1} & \text{for } \mathbf{p} \in \mathcal{S}_N, \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

其中B(a)表示n维函数，Sy为n维单位单纯形。

不确定性感知学习。为了有效地消除不可靠匹配对带来的不确定性的影响，对匹配不确定性建模的必要性是显而易见的。虽然主观逻辑(SL)理论在不确定性建模方面取得了显著进展，但将SL理论直接应用于基于文本的人物检索还不合适。为了将SL扩展到这个特定的任务，第一步是表示第i个文本和第j个图像之间的跨模态匹配证据eij = earpf (Sim(，))的预测，其中Sim()和f表示余弦相似度的计算和ReLU函数。证据e;的i-th文本的总匹配可以表示为ei= fei=1

根据前一节提到的主观逻辑，我们得到a;并对匹配不确定性u建模如下:

$$\boldsymbol{\alpha}_i = \mathbf{e}_i + 1, \quad \mathbf{u} = \frac{N}{\mathbf{S}}, \tag{3}$$

其中S =Σ (a)可视为狄利克雷分布的强度。基于获得的匹配不确定性，我们执行不确定性感知学习(uncertainty - aware Learning)，自适应过滤不可靠的匹配对，选择高置信度的跨模态匹配。具体而言，我们设计了具有不确定性感知的动态权函数(m(i))的交叉熵损失Cu，以便在优化过程中对匹配不确定性较低的跨模态匹配分配较大的权重，对匹配不确定性较高的跨模态匹配分配较小的权重，从而减少不可靠匹配对带来的负面影响。损失函数Cu可以表示为:

$$\mathcal{L}_u = \lambda \sum_{i=1}^{N} \varphi(m(i)) \mathbf{Y}_i \Big(\log(S_i) - \log(\boldsymbol{\alpha}_i) \Big), \quad (4)$$

其中A为超参数，Y, i为第i个样本的单热表, um() m) (o. () mmmics通过对匹配不确定性u按降序排序得到的第i个跨模态匹配的序数。

基于不确定性的对齐优化

由于缺乏视觉和语言之间的一对多对应关系，现有的方法主要集中在探索片面的跨模态对齐，即一对一对应，导致检索性能退化。为了解决这个限制，我们提出了一个Ücertainty-based对齐优化(UAR)模块，它模拟粗粒度的对齐，并采用渐进式学习以一种容易而困难的方式协作地优化粗粒度和细粒度的对齐。

不确定性表示构建(Uncertainty Representation Construction)。给定N个图像-文本对的全局表示(vls, tls)，我们需要首先明确地构建具有不确定性的视觉表示，这是通过附加原始特征分布的高斯噪声来实现的。高斯噪声的均值和标准差均由原始特征vils导出。然后，我们通过将生成的高斯噪声添加到白化特征vls中来构建具有不确定性ls的视觉表示，其可以计算如下:

$$\hat{\mathbf{v}}^{cls} = \boldsymbol{\alpha}_v \cdot \bar{\mathbf{v}}^{cls} + \boldsymbol{\beta}_v, \quad (5)$$

其中ay和3是引入噪声的不确定性向量，ay ~ N(1，)，β，~ N(u, o)，而icle是白化特征idla -

对齐渐进式学习。基于获得的具有不确定性和文本表征信息的视觉表示，我们采用了InfoNCE损失信息(Lee, Kim, and Han 2021;Yang et al. 2023)执行粗粒度对齐，并进一步探索一对多对应关系。粗粒度对齐的损失可以定义如下:

$$\mathcal{L}_{ca} = \frac{\mathcal{L}_{info}\left(\hat{\mathbf{v}}_i^{cls}, \mathbf{t}_i^{cls}\right)}{2\sigma^2} + \frac{1}{2}\log\sigma^2, \quad (6)$$

在细粒度比对方面，即一对一对应，我们设计了一个成对损失函数Cja来减轻密集采样机制的不利影响(Zhou et al. 2023)。仅使用一个负样本腿的成对损失函数可以写成:

$$\mathcal{L}_{fa} = -\log \frac{\psi\left(\mathbf{v}_i^{cls}, \mathbf{t}_i^{cls}\right)}{\psi\left(\mathbf{v}_i^{cls}, \mathbf{t}_{neg}^{cls}\right) + \psi\left(\mathbf{v}_i^{cls}, \mathbf{t}_i^{cls}\right)}, \quad (7)$$

直观地说，进行细粒度比对明显比粗粒度比对更具挑战性。因此，我们的策略包括在一开始就为粗粒度对齐分配更高的权重，为细粒度对齐分配更低的权重，在训练过程中逐渐逆转这种分配。我们提出对齐渐进式学习(APL)，将动态权重纳入损失函数，允许以“易-难”的方式逐步关注多粒度对齐，同时优化以下目标La:

$$\mathcal{L}_a = \sum_{i=1}^{N} \varphi(m(i))(\gamma \mathcal{L}_{ca} + (1-\gamma)\mathcal{L}_{fa}), \quad (8)$$

其中y = exp(- 70”ocepion)， d = tti wigingh。

跨模态掩模建模

为了增强图像和文本之间的交互作用，我们设计了跨模态掩模建模(CMM)，使用掩模输入重构一个模态的固有信号，该信号以图像和文本模态的非掩模输入为条件。该模型可进一步分为两个部分:跨模态掩模年龄建模(CMIM)和跨模态掩模语言建模(CMLM)。

以CMIM为例，遵循MAE (He et al. 2022)，我们得到掩膜图像的表示Vm = (v)， nu表示未掩膜的to-kens个数。然后，我们利用包含多个头部交叉注意层和3层变压器块的跨模态编码器，根据掩码图像的表示Vm和原始文本表示E，对所有原始符号进行预测;= {t} -1。最后，通过图像跨模态解码器fa将预测映射回RGB图像空间，其结构与编码器相同，后面是线性层。CMIM的总过程表示为:

$$\mathcal{L}_{cmim} = \frac{1}{\Omega(\mathbf{I}_i)} \|\mathbf{I}_i - f_d(f_e(\mathbf{V}_{m_i}, \mathbf{E}_i))\|_1, \quad (9)$$

其中2()为像素数，损失函数Cemim基于li损失。

与CMIM类似，给定表示Em;我们利用交叉熵损失函数H来度量预测与被屏蔽文本标记Em之间的距离，即进行跨模态掩模语言建模。因此，CMM的目标可以计算为:

$$\mathcal{L}_{cmm} = \mathcal{L}_{cmim} + \mathcal{H}(\mathbf{y}_{m_i}, f_{td}(f_{te}(\mathbf{V}_i, \mathbf{E}_{m_i})), \quad (10)$$

其中，Ym，为第i个掩码的单键标签，fte与CMIM的跨模态编码器相同，fid为分类器头。通过最小化Ccmm，模型被迫通过跨模态交互来执行原始信号的重建。这一过程有效地促进了对图像和文本模态之间存在的更深层次关系的探索。

最后，训练的总体损失Cotat表示为:

$$\mathcal{L}_{total} = \mathcal{L}_u + \mathcal{L}_a + \mathcal{L}_{cmim}. \quad (11)$$

实验

实验设置

数据集。我们在三个基准数据集上评估我们的模型，包括:1)CUHK-PEDES (Li et al. 2017)共包含40,206张图像，捕获13,003个不同的身份，并伴随着68,120个文本描述。2) ICFG-PEDES (Ding et al. 2021)是一个大规模的人数据集，包含54,522张图像的大量集合，其中训练集包含34,674对图像-文本，而测试集包含19,848对图像-文本。3) RSTPReid (Zhu等人，2021)共包含20505张图像，跨越4101个不同的身份，全部由15台不同的相机拍摄。此外，每张图像都丰富了两种文本描述。

评价。遵循之前的方法(Jiang和Ye, 2023;Yang等人，2023)，我们采用Rank@K (R@K)作为所有数据集的标准评估，其中表示从top-k候选图像中检索到至少一个奇异对应目标图像的百分比。

实现细节。我们使用PyTorch实现我们的模型。为了公平起见，我们遵循基于vlp的方法(Yang等人，2023)，并使用相同的视觉和文本编码器，即Swin Transformer (Liu等人，2021)和Bert (Vaswani等人，2017)。整个训练过程可以分为两个阶段。我们首先在(Yang et al. 2023)提出的数据集上训练整个模型。然后我们将图像大小调整为384 x 128，并将每个文本标记序列的长度设置为56。通过第一阶段的参数初始化，我们使用Adam优化器(Kingma and Ba 2015)用PyTorch训练了35个epoch的AUL模型，其学习率初始化为5e-5，并在线性学习率衰减后衰减为5e-6。批大小设置为128个。最后，所有实验的A和yo分别设置为0.8和1.0。

总体对比结果

我们将我们提出的方法CMAP与最新的最先进的方法进行了比较，包括:(1)传统的预训练方法，通过注意机制和额外的信息线索来提高跨模态匹配的准确性，如LGUR (Shao等人，2022)，LBUL (Wang等人，2022b);几种方法DSSL (Zhu等人，2021)、AXM Net (Farooq等人，2022)、ISANet (Yan等人，2022b)、CAIBC (Wang等人，2022a)、RKT (Wu等人，2023)和SRCF (Suo等人，2022)提出了一些简单的策略来实现不同的语义，以便进行适当的对齐。(2)利用超大图像文本语料库先验知识的视觉语言预训练方法，包括CFine (Yan等);

	方法	r@	R@5	R@10
非VLP	DSSL (mm '21)	59.98	80.41	87.56
	SSAN (arXiv'21)	61.37	80.15	86.73
	AXM-Net (AAAT'22)	61.90	79.40	85.75
	Caibc(22年)	64.43	82.87	88.37
	Lbul (mm '22)	64.04	82.66	87.22
	Lgur (mm '22)	64.21	81.94	87.93
	C2a2 (mm '22)	64.82	83.54	89.77
	ISANet (arXiv'22)	63.92	82.15	87.69
	SRCF (ECCV'22)	64.04	82.99	88.81
	RKT (tmm '23)	61.48	80.74	87.28
VLP	阿萨姆(23岁	65.66	84.53	90.21
	TVT (ecvw 22)	65.59	83.11	89.21
	CFine(2014年12月)	69.57	85.93	91.15
	TP-TPS [j]	70.16	86.10	90.98
	Irra (cvpr '23)	73.38	89.93	93.71
	[j] [c] [b]	76.51	90.29	94.25
	APTM (MM'231)	76.17	89.47	93.57
	AUL(我们的	77.23	90.43	94.41

表1。CUHK-PEDES比较。

	方法	r@	R@5	R@10
非VLP	DSSL (mm '21)	39.05	62.60	73.95
	SSAN (arXiv'21)	43.50	67.80	77.15
	Lbul (mm '22)	45.55	68.20	77.85
	C2a2 (mm '22	51.55	76.75	85.15
VLP	(ecvw '22)	46.70	70.00	78.80
	CFine(2014年12月)	50.55	72.50	81.60
	TP-TPS (arXiv'23)	50.65	72.45	81.20
	Irra (cvpr '23)	60.20	81.30	88.20
	RaSa (ICAI'23)	66.90	86.50	91.35
	Aptm (mm '23)	66.45	85.60	90.60
	AUL(我们的)	71.65	87.5	92.05

表2。RSTPReid与近期方法的比较。

2022a)、TP-TPS (Wang 等，2023)、IRRA (Jiang 和 Ye, 2023)、RaSa (Bai等，2023)和APTM (Yang等，2023)。监督设置的比较。通过对表1、2、3三个数据集的对比，我们可以发现:(1)我们的AUL模型在RSTPReid上达到了惊人的71.65% R@1，比RaSa和APTM分别高出4.75%和5.20%。这些结果表明，我们的AUL方法在缓解严重匹配模糊的不利影响和挖掘视觉和语言之间更细粒度的对应关系方面表现出色。(2)此外，与现有方法相比，我们在所有三个数据集上都取得了显着改进，这表明我们的AUL模型通过模拟粗粒度对齐有效地改进了跨模态对齐，并逐步协同多粒度对齐。

弱监督与域泛化设置的比较。此外，我们还在弱监督和域泛化设置下评估了我们的AUL模型。从表5和表6中，我们可以观察到:(1)我们的AUL模型比当前最先进的APTM方法表现出了实质性的性能提升，特别是在R@1方面。这种改进可以归因于这样一个事实，即在只有成对关系的情况下

	方法MIA	r@	R@5	R@10
	(TIPP20)	46.49	67.14	75.18
P-I 2O M	SSAN (arXiv'21)	54.23	72.63	79.53
	Lgur (mm '22)	57.42	74.97	81.45
	ISANet (arXiv'22)	57.73	75.42	81.72
	SRCF (ECCV'22)	57.18	75.01	81.49
	阿萨姆(23岁)	57.09	76.33	82.84
	(ecvw '22)	56.04	3.60	80.22
	CFine(2014年12月)	60.83	76.55	82.42
A T A M	TP-TPS (arXiv'23)	60.64	75.97	81.76
	Irra (cvpr '23)	63.46	80.25	85.82
	RaSa (ICAI'23)	65.28	80.40	85.12
	Aptm (mm '23)	68.22	82.87	87.50
	AUL(我们的)	69.16	83.32	88.37

表3。ICFG-PEDES的比较。

没有。	组件war umf	Cfa Lca		三坐标测量机	RSTPReid R@1 R@5 R@10
	-	-	-	-	68.15 85.10 89.20
	-	-	-	✔	69.25 85.65 90.20
	-	✔	✔	-	69.15 86.10 90.35
	✔	-	-	-	69.55 85.40 89.50
	-	✔	-		70.85 85.95 90.53
	-	*	✔		70.95 87.10 91.25
	✔	-	-	✔	71.35 86.85 90.90
7	✔	✔	✔	✔	71.65 87.55 92.05

表4。RSTPReid上模型部件的烧蚀研究。

方法	R@1 R@S	R@10
Cmpm + mmt (iccv21) 50.5170.23		78.98
CMPM+SpCL (ICCV'21) 51.13 71.54		80.03
CMMT (iccv '21) 57.10 78.14 caibc (mm '22)		85.23
58.64 79.02		85.93
Irra (cvpr '23)	70.94 88.39	93.06
Aptm (mm '23)	74.57 88.95	93.18
AUL(我们的)	75.86 90.11	94.02

表5。CUHK-PEDES与最新技术比较(弱监督)。

	方法	@ 1	R@5	R@10
	SSAN (arXiv'21)	2924	4900	58.3
	Lgur (mm '22)	34.25	52.58	60.85
	C2a2 (mm '22)	27.61	47.48	57.03
?	阿萨姆(23岁)	30.22	50.51	59.59
	Irra (cvpr '23)	41.89	61.56	69.04
	Aptm (mm '23)	46.20	65.13	72.59
	AUL(我们的)	49.29	67.46	74.42
	SSAN (arXiv'21)	21.07	38.94	48.54
	Lgur (mm '22)	25.44	44.48	54.39
?	C2a2 (mm '22)	16.48	34.03	43.88
个	阿萨姆(23岁)	17.99	35.30	44.75
	Irra (cvpr '23)	31.04	52.18	63.53
	Aptm (mm '23)	48.67	68.75	77.06
	AUL(我们的)	56.79	76.14	83.14

表6.与最先进技术(领域泛化)的比较。C代表CUHK-PEDES, r代表ICFG-PEDES。

可用和身份信息缺失，显著的类内变化的影响变得更加明显，导致检索性能受损。因此，我们提出了UMF来缓解匹配模糊性的影响。(2)此外，我们的AUL模型在I→C方面比最近的APTM方法在R@1和R@5上分别提高了8.12%和7.39%。这些结果表明，我们的AUL模型可以有效地量化跨模态匹配模糊性中固有的不确定性，并过滤掉高置信度的对齐。

进一步分析

消融研究。如表4所示，我们列出了以下结论:(1)与No.0和No.3的比较表明，我们提出的UMF显著提高了检索性能。这再次表明，引入SL理论对跨模态匹配模糊度的不确定性建模对于过滤高置信度对齐是有效的，这使得我们的模型专门用于可靠的重新检索结果。(2) No.5的模型性能优于No.1的结果，特别是在R@5和R@10方面。结果表明，利用基于高斯噪声的不确定性表示，UAR可以有效地探索一对多对应关系。此外，UAR采用的渐进式学习方法可以适当地协作粗粒度和细粒度对齐。(3)通过对比No.6和No.3，我们推测添加CMM对re-的影响更大检索性能。一个可能的原因是，执行具有进一步跨模态交互的MLM和MIM在视觉和语言之间的细粒度和相关关系挖掘方面产生了额外的优势。

CMLM和CMIM的选择分析。我们进一步探讨了CMLM和CMIM的重要性。如图4所示，我们可以观察到:(1)消融模型w/ Cemlm的性能优于基线。我们认为，优越的性能是由于图像和文本之间的充分交互，这更有助于弥合视觉和语言之间显著的模态差距。(2)此外，单独应用CMLM的损失Lemlm并不像CMIM和CMLM的组合(即w/ CMM)那样有效。这表明，同时使用屏蔽文本和视觉标记作为挖掘综合跨模态关系的锚点是必不可少的。

UAR中对齐渐进式学习分析。

在这里，我们研究了我们提出的Align-的进步性渐进式学习(APL)，旨在全面探索一对一和一对多对应关系。通过观察图3，我们可以发现:(1)引入动态权值y优于Avg的细化模型，我们推测其原因是利用渐进式学习在学习综合多粒度对齐方面发挥了重要作用。(2)本文提出的APL算法在出口处有效地为粗粒度排列分配更高的权重，并逐渐向分配方向转移

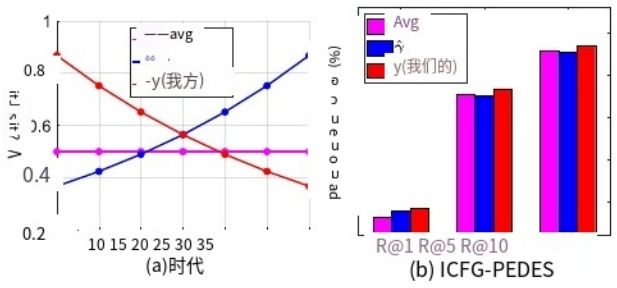

图3。关于UAR对齐渐进式学习的分析。y和f分别表示我们提出的渐进式学习方式和相反的渐进式学习方式，Avg为平均权重分配。

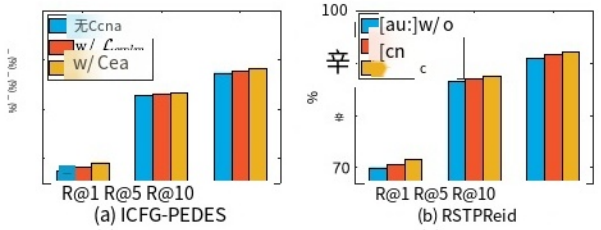

图4。跨模态掩模建模中各分量对ICFG-PEDES和RSTPReid的影响。

对细粒度对齐的更高权重。(3)此外，我们进一步探索了以易难方式学习多粒度对齐的有效性。特别地，我们比较了利用y和的性能。显然，前者更适合于正确的对齐合并和检索精度。这一发现支持了我们的理论，即引导模型以一种简单而困难的方式逐步学习适当的多粒度对齐比其他方法更合理。

匹配不确定性感知动态权重分配分析。为了进一步验证匹配模糊性的存在以及我们提出的UMF的重要性，我们深入研究了各种不确定性感知权重分配与整体性能之间的关系。从图5中得出的观察结果概述如下:(1)分布分析清楚地揭示了不可靠匹配对的存在，其特征是明显的匹配不确定性。这种不确定性来自于显著的类内变化和有限的类间变化，阻碍了检索性能的提高。(2)为了强调匹配不确定性感知动态权重分配的有效性，我们比较了不同权重分配的性能。将高不确定性的跨模态匹配设置为1(-)会产生最差的性能，这反映了我们的动机的合理性，我们的模型遭受了严重的匹配歧义。

定性分析。如图6所示，我们进行了定性分析，将我们的AUL方法的前6个检索结果与最近的APTM方法(Yang et al. 2023)进行了比较。根据可视化结果，

图5。ICFG-PEDES匹配不确定性感知权值分配的有效性。当不确定度大于0.5时，0(-)和1(-)分别表示样本的权重设置为0和1。

图6。APTM和我们提出的AÜL在ICFG-PEDES和CUHK-PEDES上的定性结果。

我们的AUL在检索精度上优于APTM方法。具体来说，我们的方法AUL可以同时满足细粒度和粗粒度的检索需求(如“长袖”或“高个子”)，因为我们提出的UAR逐步全面地获取多粒度语义。此外，AUL还建立了匹配不确定性模型，量化了类内变化大而类间变化小所带来的模糊性，从而减轻了不可靠匹配对的干扰，从而提高了性能。

结论

本文提出了一种新的基于不确定性的自适应学习(AUL)方法，用于基于文本的人物检索。我们提出了不确定性感知匹配过滤(UMF)来量化和防止不可靠匹配对造成的歧义影响。此外，我们设计了基于不确定性的对齐优化(UAR)和跨模态掩模建模(CMM)来增强对齐学习并关注适当的跨模态关系。在三个基准上进行的大量实验证明了我们提出的AUL方法的优越性。在未来，我们将探索其他策略来提高检索性能。

致谢

国家自然科学基金项目(No. 62222203,0和3)、科纳斯通科学基金项目(XPLORER PRIZE)、深圳市科技创新委员会基金项目(No. 62222203,0和3)的部分资助。jcyj20210324132203007)。

参考文献

Aggarwal, S; Babu, R. V; and Chakraborty, A. 2020. Text-based Person Search via Attribute-aided Matching. In IEEE Winter Conference on Applications of Computer Vision, 2606-2614.

Bai, Y; Cao, M.; Gao, D.; Cao, Z.; Chen, C.; Fan, Z.; Nie, L.; and Zhang, M. 2023. RaSa: Relation and Sensitiv-ity Aware Representation Learning for Tex-based Person Search. CoRR, abs/2305.13653.

Ding, Z. :; Ding, C. ; Shao, Z. ; and Tao, D. 2021. Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification. CoRR, abs/2107.12666.

Farooq, A.; Awais, M.; Kittler, J.;, and Khalid, S. S. 2022. AXM-Net: Implicit Cross-Modal Feature Alignment for Person Re-identification. In AAAI, 4477-4485.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the International Conference on Machine Learning, volume 48, 1050-1059.

He, K.; Chen, X.; Xie, S:; Li, Y; Dollár, P; and Girshick, R. B. 2022. Masked Autoencoders Are Scalable Vision Learners. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15979-15988.

Hou, R.; Chang, H.; Ma, B.; Huang, R.; and Shan, S. 2021. BiCnet-TKS: Learning Efficient Spatial-Temporal Repre sentation for Video Person Re-Identification. In IEEE Conference on Computer Vision and Pattern Recognition, 2014-2023.

Jiang, D.; and Ye, M. 2023. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. CoRR, abs/2303.12501.

Jiang, X.; Xu, X.; Zhang, J.; Shen, F,; Cao, Z.; and Shen, H. T. 2022. Semi-supervised video paragraph grounding with contrastive encoder. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2466-2475.

Jiang, X.; Zhou, Z.; Xu, X.; Yang, Y.; Wang, G.; and Shen, H. T. 2023. Faster Video Moment Retrieval with Point-Level Supervision. arXiv preprint arXiv:2305.14017.

Jing, Y.; Si, C. Wang, J; Wang, W; Wang, L.; and Tan, T. 2020. Pose-Guided Multi-Granularity Attention Network for Text-Based Person Search. In AAA, 11189-11196. Jesang, A. 2016. Subjctive Logic - A Formalism for Rea-soning Under Uncertainty. Springer.

Kendall, A.; and Gal, Y. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In Advances in Neural Information Processing Systems, 5574-5584.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In ICLR.

Lee, S.; Kim, D.; and Han, B. 2021. CoSMo: Content-Style Modulation for Image Retrieval With Text Feedback. In CVPR, 802-812.

Li, S:; Xiao, T, Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017. Person Search with Natural Language Description. In IEEE Conference on Computer Vision and Pattern Recognition, 5187-5196.

Li, S.: Xu, X.; Jiang, X.; Shen, F.; Liu, X.; and Shen, H. T. 2023a. Multi-Grained Attention Network with Mutual Exclusion for Composed Query-Based Image Retrieval. IEEE Transactions on Circuits and Systems for Video Technology. Li, S.; Xu, X.; Shen, F; and Yang, Y. 2023b.Mult-granularity Separation Network for Text-Based Person Re-trieval with Bidirectional Refinement Regularization. In Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, 307-315.

Li, S.; Xu, X.; Yang, Y; Shen, F; Mo, Y.; Li, Y; and Shen, H. T. 2023c. DCEL: Deep Cross-modal Evidential Learning for Text-Based Person Retrieval. In Proceedings of the 31st ACM International Conference on Multimedia, 6292-6300. Liu, Z.; Lin, Y; Cao, Y; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In EEE/CVF International Conference on Computer Vision, 9992-10002. Niu, K.; Huang, Y; Ouyang, W.; and Wang, L. 2020. Improving Description-Based Person Re-Identification by Multi-Granularity Image-Text Alignments. IEEE Trans. Im-age Process, 29: 5542-55556.

Niu, K. ; Huang, Y. ; and Wang, L. 2020. Textual Dependency Embedding for Person Search by Language. In Chen, C. W.; Cucchiara, R.; Hua, X.; Qi, G.; Ricci, E.; Zhang, Z.; and Zimmermann, R., eds., ACM International Conference on Multimedia, 4032-4040.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G; Askell, A.; Mishkin, P; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transfer-able Visual Models From Natural Language Supervision. In ICML, volume 139, 8748-8763.

Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.: Wang, J; and Ding, C. 2022. Learning Granularity-Unified Representations for Text-to-Image Person Re-identification. In The ACM International Conference on Multimeda, 5566-5574.

Specker, A.; Cormier, M.; and Beyerer, J.023. UPAR: Unified Pedestrian Attribute Recognition and Person Retrieval. In WACV, 981-990.

Suo, W.; Sun, M.; Niu, K.; Gao, Y; Wang, P; Zhang, Y; and Wu, Q. 2022. A Simple and Robust Correlation Filtering Method for Text-Based Person Search. In ECCV, 726-742. Vaswani, A.; Shazeer, N; Parmar, N; Uszkoreit, J; Jones, L.; Gomez, A. N; Kaiser, L.; and Polosukhin, I. 2017. At-tention is All you Need. In Advances in Neural Information Processing Systems, 5998-6008.

Wang, G; Yu, F; Li, J; Jia, Q and Ding, S. 2023. Exploit-ing the Textual Potential from Vision-Language Pre-training for Text-based Person Search. CoRR, abs/2303.04497.

Wang, Z; Fang, Z; Wang, J.; and Yang, Y. 2020. VITAA: Visual-Textual Atributes Alignment in Person Search by Natural Language. In Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII, volume 12357, 402-420. Wang, Z.; Zhu, A.; Xue, J; Wan, X.; Liu, C; Wang, T.; and Li, Y. 2022a. CAIBC: Capturing All-round Information Beyond Color for Text-based Person Retrieval. In ACM International Conference on Multimedia, 5314-5322. Wang, Z.; Zhu, A.; Xue, J.; Wan, X.; Liu, C.; Wang, T; and Li, Y. 2022b. Look Before You Leap: Improving Text-basedPerson Retrieval by Learning A Consisent Cross-modal Common Manifold. In The ACM International Conference on Multimedia, 1984-1992.

Warburg, F.; Jorgensen, M; Civera, J.; and Hauberg, S. 2021. Bayesian Triplet Loss: Uncertainty Quantification in Image Retrieval. In IEEE/CVF Internationai Conference on Computer Vision, 12138-12148.

Wu, Y.; Yan, Z.; Han, X.; Li, G:; Zou, C.; and Cui, S. 2021. LapsCore: Language-guided Person Search via Color Rea-soning. In IEEE/CVF International Conference on Com-puter Vision, 1604-1613.

Wu, Z.; Ma, B.; Chang, H.; and Shan, S. 2023. Refined Knowledge Transfer for Language-Based Person Search. IEEE Transactions on Multimedia, 1-15. Yager, R. R.; and Liu, L. 2008. Classic Works of the Dempster-Shafer Theory of Belief Functions, volume 219. Springer.

Yan, S.; Dong, N.; Zhang, L.; and Tang, J. 2022a. CLIP-Driven Fine-grained Text-Image Person Re-identification. CoRR, abs/2210.10276.

Yan, S; Tang, H.; Zhang, L.; and Tang, J. 2022b. Image Specific Information Suppression and Implicit Local Alignment for Text-based Person Search. CoRR, abs/2208. 14365. Yang, S.; Zhou, Y.; Wang, Y.; Wu, Y.; Zhu, L.; and Zheng, Z. 2023. Towards Unified Text-based Person Retrieval: A Large-scale Multi-Attribute and Language Search Bench-mark. CoRR, abs/2306.02898.

Zheng, Z; and Yang, Y. 2021. Rectifying Pseudo Label Learning via Uncertainty Estimation for Domain Adaptive Semantic Segmentation. Int. J. Comput. Vis., 1294): 1106-1120.

Zheng, Z; Zheng, L.; Garrett, M,; Yang, Y; Xu, M.; and Shen, Y. 2020. Dual-path Convolutional Image-Text Embeddings with Instance Loss. ACM Trans. Multim. Comput. ComTun. Appl., 16: 51:1-51:23.

Zhou, X.; Zhong, Y; Cheng, Z.; Liang, F; and Ma, L. 2023. Adaptive Sparse Pairwise Loss for Object Re-Identification. CoRR, abs/2303.18247.

Zhu, A.; Wang, Z; Li, Y; Wan, X.; Jin, J; Wang, T; Hu, F; and Hua, G. 2021. DSSL: Deep Surroundings-person Separation Learning for Text-based Person Retrieval. In Pro ceedings of the 29th ACM International Conference on Mul-timedia, 209-217.