IIITD-20K：用于文本-图像行人重识别的密集描述

A V Subramanyam
IIITD，印度
subramanyam@iiitd.ac.in

Niranjan Sundararajan*
IIITD，印度
niranjan20090@iiitd.ac.in

Vibhu Dubey*IIITD，
印度
vibhu20150@iiitd.ac.in

Brejesh Lall IITD，
印度
brejesh@ee.iitd.ac.in

arXiv:2305.04497v1 [cs.CV] 8 May 2023

摘要

文本到图像（T2I）行人重识别近年来受到了广泛关注。目前有三个可用于评估 T2I 行人重识别方法的基准数据集：CUHK-PEDES、RSTPReid 和 ICFG-PEDES。RSTPReid 和 ICFG-PEDES 均包含来自 MSMT17 的身份，但由于唯一身份数量有限，数据多样性不足。另一方面，CUHK-PEDES 包含 13,003 个身份，但其平均文本描述相对较短。此外，这些数据集是在受限环境中采集的，摄像机数量有限。为了进一步丰富身份多样性并提供密集的图像描述，我们提出了一个名为 IIITD-20K 的全新数据集。IIITD-20K 包含在自然场景中采集的 20,000 个唯一身份，为文本到图像行人重识别提供了丰富的数据。每条图像描述至少包含 26 个词，因此每张图像都有密集的描述文本。我们还使用在本数据集上训练的 Stable Diffusion 和 BLIP 模型合成生成图像并生成细粒度描述文本。我们采用最先进的文本到图像行人重识别模型和视觉语言预训练模型进行了详尽实验，并对该数据集进行了全面分析。实验还表明，合成生成数据能够显著提升同数据集设置和跨数据集设置下的性能。本数据集可通过 https://bit.ly/3pkA3Rj 获取。

CCS 概念

• **信息系统** → **图像搜索**；• **计算方法** →**匹配**。

关键词

文本到图像 ReID，基准，合成数据，IIITD-20K 数据集

1 引言

文本到图像行人重识别（T2I ReID）旨在将对某人的查询文本描述与一组图库图像进行匹配，以检索目标身份 [19]。与查询和图库均为图像的单模态图像行人重识别相比，这项跨模态任务极具挑战性。文本描述通常较为笼统、粗略，而图像则包含丰富、细致的视觉信息。因此，给定的文本描述可能对应多张视觉细节各异的图像，导致文本特征的类间方差较大。因此，文本到图像行人重识别面临较大的模态间隙，难度远高于基于图像的行人重识别。此外，由于文本描述通常含糊且带有主观性，因此很难将其与视觉特征准确匹配。因此，为文本到图像行人重识别开发有效的方法行人重识别是计算机视觉和自然语言处理领域一项重要且具有挑战性的研究方向。

CUHK-PEDES [19] 是首次尝试构建文本到图像行人重识别数据集。该数据集由 5 个不同的数据集构建而成：CUHK03[21], 、Market-1501 [37], 、SSM [34], 、VIPER [10] 和 CUHK01 [20]。每张图像配有两条图像描述，平均词长为 23.5。部分描述文本还包含动作和背景细节 [5]。CUHK-PEDES 中的图像是在相似条件下获取的。为克服这些局限，RSTPReid [39] 和 ICFG-PEDES [5] 使用了来自 MSMT17 的图像；与 CUHK-PEDES 的图像相比，这些图像是在更复杂的条件下拍摄的。RST- PReid 中每张图像有两条描述文本，最短词长为 23。此外，针对 CUHK-PEDES 描述文本平均词长较短且包含无关细节的问题，Ding 等人提出了 ICFG-PEDES。该数据集的身份数量较少，但平均词长约为 35。此外，描述以身份为中心，并避免提及动作或背景细节。

本文提出了一个用于文本到图像 ReID 的新数据集。我们的数据集平均词长为 36，是 CUHK-PEDES 的 1.56× 倍。与 RSTPReid、ICFG-PEDES 和 CUHK-PEDES 相比，我们的身份数量分别为 4.87× 倍和 1.53× 倍。我们的数据集中的图像采集自公开可用的来源，因此不受环境限制，这与现有数据集不同。因此，该数据集规模庞大且具有多样性，可加速文本到图像 ReID 领域的研究。

2 相关工作

近年来，跨模态文本目标识别受到了广泛关注。这主要得益于视觉和 NLP 社区开发的高效深度学习模型[2]。这些方法大致可分为基于全局特征嵌入的方法和基于注意力的 Re-ID 方法。

2.1 全局特征嵌入

一些早期基于深度学习的方法侧重于提取全局特征 [7, 8] ，并采用了多种损失函数，例如 N-pair [29] 和 bi-rank [23]。Zhang 等人 [35]提出了用于投影匹配（CMPM）和分类（CMPC）的跨模态损失函数，以学习具有判别性的文本-图像表示。然而，基于全局特征嵌入的策略忽视了文本-图像对中细粒度视觉语义相似性的重要性。因此，学习到的表示会受到背景或动作等无关细节的影响。为解决这一问题，研究者提出了基于注意力的方法，接下来我们将对此进行讨论。

*贡献相同

表1：数据集

数据集	平均词数长度	唯一IDs	唯一词数	最少词数/图像描述	最多词数/图像描述		图像描述/图像	摄像机
CUHK-PEDES	23.5	13,003	9,408	12	96	40,206	2	≪ 身份
RSTPReid	25.75	4,101	4,628	11	68	20,505	2	
ICFG-PEDES	34.86	4,102	4,411	9	79	54,522	1	
IIITD-20K	35.9	20,000	5,281	26	97	20,000	2	≈ ID

2.2 基于注意力的模型

Li 等人 [19] 提出了一种 GNA-RNN 方法，该方法结合视觉子网络和语言子网络，有效构建词图关系，并对语言和视觉信息进行编码。单元级注意力和词级门控分别根据输入词和不同词语的重要性，对视觉单元进行加权。此外，汇总所有单元激活后，可生成句子与人物图像之间的最终亲和度。基于注意力的机制也应用于其他研究中 [1, 3, 5, 13, 18, 22, 38, 39]。

凭借 Transformer 的卓越性能，BERT 等自然语言处理模型 [4] 已被应用于 HGAN [36], NAFS [9], LapsCore [33],LGUR [28]。

2.3 视觉语言模型

受 BERT 类模型成功的启发，视觉语言领域受到广泛关注。视觉语言模型旨在学习图像及其对应文本描述的联合表征，并将其应用于图像描述、检索、视觉问答等下游任务。这些模型大致可分为双流模型和单流模型。ViLBERT[24] 和 LXMERT [30], 等双流模型使用两个模块分别处理文本和图像。相比之下，VisualBERT[17], 、VL-BERT 和 UNITER 等单流模型在同一模块中对两种模态进行编码。CLIP[25], 、ALIGN [14]和 BLIP [16] 等视觉语言预训练模型在基于文本的检索任务中取得了显著进步。IRRA [15] 在基于文本的行人重识别任务中也取得了显著进步。

2.4 基线

我们选择了三个不同的基线，即 IRRA [15], 、LGUR [28] 和 SSAN [5], , 用于分析所提出的数据集。

IRRA [15] 采用跨模态方法，学习将视觉和文本模态映射到共同潜在空间，从而有效匹配多模态数据。该方法采用一种新颖的相似度分布匹配（SDM）损失，通过最小化图像-文本相似度与根据标签得到的分布之间的 KL 散度来进行优化。

LGUR [28]由两个模块组成：基于字典的粒度对齐（DGA）和基于原型的粒度统一（PGU）。DGA通过使用基于 Transformer 的多模态共享字典重构视觉特征和文本特征，弥合图像与文本模态之间的差距。由于 DGA 基于重构任务，在缺乏明确指导的情况下，它无法学习到稳健的判别性特征。为解决这一问题，PGU



图 1：数据集中的示例图像。

通过使用一组共享且可学习的原型，将两种模态投影到共享特征空间中，从而将它们结合起来。与使用跨模态注意力操作的方法相比，共享原型还有一项额外优势：它能显著降低 LGUR 的计算成本。

Ding 等人提出了一种名为语义自对齐网络（SSAN） [5] 的新模型。SSAN 利用语言描述中的上下文线索，通过推断词语与部位之间的对应关系，提取部位级视觉和文本特征。然而，这种映射忽略了身体部位之间的相关性，以及文本描述中指定的图像区域之间的空间关系。因此，SSAN 提出了一种多视图非局部网络（MV-NLN），用于捕捉身体部位之间的关系。此外，为了克服描述中的类内差异，作者提出了复合排序（CR）损失，其中包含强监督和弱监督两个部分，并将其用作一种新颖的数据增强策略。

3 所提出的数据集

我们的数据集旨在收集大量在非受限环境下拍摄的唯一身份，并由多名标注人员为其添加图像描述。这些图像抓取自



图2：所提出的数据集中的高频词。

从网络上的公开来源收集，并紧密裁剪以包含完整的人体。为抓取图像，我们使用了以下关键词，

（1）城市名称：柏林、伦敦、纽约、班加罗尔、金奈、巴黎、加尔各答、悉尼、布里斯班、东京、加利福尼亚、德里、孟买；（2）上传年份：2012—2022年

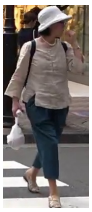
我们的数据集包含 22,727 张图像，其中 20,000 张带有图像描述，每张图像对应两条描述文本。共有 14,72,005 个词和 53,358 个句子。共有 30,874 条描述文本的词数在 20 至 40 个之间，另有 8,889 条描述文本的词数在 40 至 60 个之间。我们剔除了质量较差、存在噪声或遮挡的图像。图1展示了一些示例图像。图2中的词云展示了高频词。

图3展示了与图像对应的描述文本。对于第一张图像，我们可以看到，标注捕捉到了所有属性——白色帽子、护目镜、项链、手表以及衣服的颜色。第一条描述文本没有提到右手中的包，而第二条标注则补充了这一点。类似地，其他描述文本也详细说明了人物的外观。

我们在表1中比较了这四个数据集的唯一词语和标识符数量、图像数量、每条描述文本的最少和最多词数，以及每张图像的描述文本数量。为此，我们遵循 [19] 中的分词方式，并使用相同的方法处理所有四个数据集。我们发现，我们的数据集在唯一标识符和词语数量方面表现更好，在每条图像描述的平均、最少和最多词数方面也更胜一筹。需要指出的是，ICFG-PEDES [5] 和 RSTPReid [39] 中给出的唯一词语数量和每条图像描述的平均词数与表1中的数字不同。这可能是由于分词方式不同所致。

3.1 合成数据集

以合成数据集的形式进行数据增强有助于提升性能。潜在扩散模型（LDM） [26] 和BLIP [16] 分别在高清图像合成和



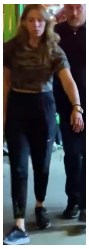
一名留着黑色短发的女孩，穿着奶油色圆领长袖上衣和深蓝色正装裤，脚穿一双奶油色带饰拖鞋，脖子上戴着白色项链，左手腕上戴着手表，头上戴着护目镜和白色帽子。

一位女士戴着白色帽子、太阳镜和白色项链，穿着长袖圆领象牙色上衣，搭配深蓝色裤子和象牙色凉鞋。她左手戴着手表，右手提着一个白色塑料袋，肩上背着一个背包。



一名男子留着黑色短直发，穿着白色正装翻领衬衫，外搭浅棕色长袖西装外套，系着深黑色皮带，穿着浅奶油色正装裤和深黑色皮鞋，左手提着一个深黑色包。

一名秃顶且留着短黑发的男子，穿着白色翻领衬衫，外搭桃色长袖外套，搭配奶油色裤子，腰间系着黑色皮带，脚穿一双黑鞋，左手提着一个黑色手提包。



一位披着金发的女士，穿着绿色迷彩短袖露脐上衣，搭配黑色九分慢跑裤和黑色运动鞋。她左手腕上也戴着一个手环。

一名留着浅棕色头发、穿着半袖深灰色露脐上衣的女性，脚踝长度深蓝色裤子，搭配黑白鞋。她还穿着一件她左手腕上的手环。

图3：所提出的数据集中的示例图像和描述文本。

描述生成方面取得了长足进展。为了充分利用这些强大的模型，我们首先使用自己的数据集对LDM进行微调。潜在扩散模型是一种条件UNet模型 [27]，通过训练来预测给定输入图像中的噪声。文本条件基于交叉注意力模型 [32]。在我们的研究中，我们使用图文对作为输入。由于微调LDM会导致灾难性遗忘，我们采用了 [12]中提出的LoRA训练技术。LoRA不仅能克服灾难性遗忘，而且需要训练的参数更少，还可以使用较高的学习率。推理时，我们将训练图像描述作为提示词，并使用DDPM采样从微调后的LDM模型中采样生成图像。

另一方面，BLIP 是一种强大的图像描述生成模型，使用大型语料库进行训练，并采用核采样 [11] 生成多样化的描述文本。然而，由于它使用的训练数据集仅包含突出要点作为图像描述，因此生成的描述文本缺乏图文行人重识别所需的细粒度信息。因此，为了生成更详细、能够捕捉图像不同属性的描述文本，我们使用自己的数据集对 BLIP 模型进行微调。

随后，将 LDM 生成的图像输入微调后的 BLIP 模型，以获取相应的图像描述。LDM 生成的图像以及 BLIP 生成的描述文本现可组成图像-描述文本对。我们进一步使用这些合成数据来训练 ReID 模型。在实验中，我们发现，使用这些合成数据增强训练能带来显著提升。

4 评估方案

这里我们讨论评估方案。按照香港中文大学 CUHK-PEDES 的做法，我们将数据集划分为训练集、验证集和测试集。训练集包含 15K 个图文对，验证集和测试集则各包含 2.5K 个图文对。训练集、验证集和测试集之间没有重叠。该数据集共包含 22,727 张图像，其余 2,727 张图像的描述文本未被采纳。我们不将这些图像纳入评估，但可以将它们作为干扰项。

5 实现细节

对于 BLIP，我们使用 ViT-B 作为图像编码器 [6]，使用 BERTbase [6] 作为文本编码器。这些模型使用 [16]提供的检查点进行初始化。我们采用 Adam 优化器，学习率设为 1e-6，权重衰减设为 5e-2，并且只训练一个轮次。其他设置与 [16]中给出的微调过程保持相似。

LDM [26] 使用 Hugging Face 的“stable diffusion v1-4”*模型和基于 CLIP 的文本编码器 [25]。我们以 12 的批次大小、1e-4 的学习率训练模型 23K 步。

IRRA 使用 CLIP 模型 [25]。LGUR 使用 DeiT-Small [31] 作为图像骨干网络，并使用 BERT [4] 作为文本编码器。SSAN 使用 ResNet-50 作为图像编码器，使用双向长短期记忆网络作为文本编码器。我们采用各自论文和发布代码中提供的相同设置来训练这些模型。

6 实验

CUHK-PEDES 包含 40,206 张图像和 80,412 条文本描述，涵盖 13,003 个身份。我们使用包含 11,003 个身份的训练集，以及验证集和测试集，两者各包含 1,000 个身份。

ICFG-PEDES 的训练集包含 3,102 个身份，测试集包含 1,000 个身份。

RSTPReid 包含 4,101 个身份。训练集包含 3,701 个身份，验证集和测试集则各包含 200 个身份。

评估指标我们采用广泛使用的 Rank-k 和平均精度均值（mAP）作为评估标准。Rank-k 用于计算在检索结果排名前 k 的图像中至少找到一张相关图像的概率。此外，mAP 会计算每个查询的平均精度，再对所有查询的平均精度取均值。

6.1 合成图像和描述文本

合成数据按照第3.1节所述的方法生成。我们为每个人物 ID 生成两张图像，并为每张图像生成一条描述文本。总计，我们额外生成了 3 万张图像和描述文本。我们分配相同的标签
与原始数据集相同。在图4中，我们展示了由 LDM 生成的图像，以及由微调后的 BLIP 模型生成的相应描述文本。可以看到，最上方图像对应的描述部分正确。“绿色夹克”和“蓝色牛仔裤”等属性与图像相符，而图像中缺少“白色鞋子”和“腰带”这部分内容。同样，第二张图像中，“盘发”这部分与图像不符。尽管这些描述文本存在噪声，但它们有助于模型训练。

*https://huggingface.co/CompVis/stable-diffusion-v1-4

此外，我们使用数据集中的原始图像，通过 BLIP 模型生成了 15K 条描述文本。在训练期间，我们将这些描述文本与数据集中的原始图像结合使用，以分析这些描述文本的质量。



一个黑色短发的男孩，穿着绿色长袖夹克和蓝色牛仔裤，搭配白色鞋子，腰间系着黑橙色腰带



一位扎着黑色盘发的女孩，穿着白色长袖卫衣和黑色裤子，脚穿棕色鞋履，肩上还披着一件黑色夹克



一名留着黑色短发的男子，穿着灰色圆领短袖 T 恤和黑色短裤，同时穿着黑色乐福鞋



一位留着黑色短发的人，穿着浅蓝色短袖翻领 T 恤和浅棕色裤子，腰间系着深棕色腰带，脚穿深棕色正装鞋，右手戴着棕色配饰

图4：合成图像和描述文本。前两行：来自 LDM [26] 的生成图像，以及来自 BLIP [16]的我们数据集训练集描述文本。后两行：为 CUHK-PEDES 生成的样本。

6.2 消融研究
所提出的数据集

在此，我们分析了表2中从合成数据获得的结果。我们使用最先进的基线方法 IRRA [15], LGUR [28] 和 SSAN [5]对我们的数据集进行了分析。首先，我们使用 LGUR 和 IRRA 在相同的数据集设置下计算结果。在表2中，“Ours”表示使用原始数据集进行训练。B1 和 B2 表示批次，其中 B1 表示由 LDM-BLIP 生成的一组包含 15K 个图文对的数据，B2 表示另一组包含 15K 个图文对的数据。我们报告了使用原始数据集和合成数据集进行训练的各种组合的结果。

LGUR 在使用原始数据集训练时，R-1 为 65.96%；IRRA 的 R-1 为 77.08%，对应表2中的“Ours”变体。如果我们只使用原始图像和 BLIP 描述文本，则观察到的 R-1 为 17.94%。这表明，尽管这些描述文本存在噪声，但可能有助于增强原始图像描述。

BLIP。现在，我们讨论使用原始数据集（包括图像和描述文本）以及由 BLIP 生成的描述文本时的结果。由于 BLIP 描述文本存在噪声，我们为 BLIP 描述文本计算出的损失赋予较低的权重。我们用 w 表示该权重。我们发现，“Ours+BLIP (w=0.1)”这一变体在 LGUR 上表现更好。此处，$w = 0.1$ 表示使用合成描述文本计算出的损失权重为 0.1，10× 低于原始图像描述对应损失的权重。同样，$w = 1$ 表示原始数据集和合成数据集计算出的损失具有相同的权重。不过，与 $w = 0.1$ 变体相比，这种设置的结果较差。我们还尝试了学习率退火，但未能取得良好效果。

表2：使用合成数据时在相同数据集上的性能。B1+B2 表示训练期间使用了两个批次

模型	设置	R-1	mAP
LGUR	Ours	65.96	71.06
	BLIP	17.94	23.48
	我们的方法+BLIP (w=0.1)	66.70	72.02
	我们的方法+BLIP (w=1)	64.66	70.27
	我们的方法+BLIP（退火）	64.82	70.18
	B1 BLIP+LDM	15.46	20.16
	B1+B2 BLIP+LDM	21.32	26.80
	我们的+B1+B2 BLIP+LDM (w=0.1)	67.10	72.18
	我们+B1+B2 BLIP+LDM (w=1)	**67.12**	**72.46**
IRRA	Ours	77.08	83.10
	我们的方法+BLIP（w=0.1）	78.2	83.98
	我们的方法+BLIP（w=1）	76.26	82.34
	B1 BLIP+LDM	35.44	45.14
	B1+B2 BLIP+LDM	37.06	46.99
	我们的方法+B1+B2 BLIP+LDM（w=0.1）	77.54	83.48
	我们+B1+B2 BLIP+LDM（w=1）	**78.94**	**84.56**

BLIP+LDM。另一方面，当我们仅使用合成数据（B1 BLIP + LDM）时，也就是使用来自 LDM 的图像和来自 BLIP 的描述文本时，LGUR 的 R-1 为 15.46%。当我们添加更多合成数据（B1+B2 BLIP+LDM）后，R-1 提升至 21.32%。当我们将合成图像和描述文本与原始数据集结合使用时，“我们+B1+B2 BLIP+LDM（w=1）”变体在所有设置中表现最佳。LGUR 的 R-1 为 67.12%，相比不使用合成数据时提升了 1.16%。平均精度均值也呈现出类似的提升。在 IRRA 中，$w = 1$ 变体的得分为 78.94%，相比非合成数据情形提升了 1.86%。

跨数据集合成数据。在表 3 的跨数据集设置中，我们发现同时使用合成图像和描述文本时，性能最佳。对于 CUHK-PEDES，LGUR 的最佳性能为 36.68%，提升了 2.86%。我们还观察到 ICFG-PEDES 的性能提升了 1.66%。然而，RSTPReid 的提升幅度较小。IRRA 在 CUHK-PEDES 上提升了 2.97%，在 ICFG-PEDES 上则取得了显著提升。

表 3：使用合成数据的跨数据集性能。报告 Rank-1 准确率。

模型	设置	CUHK	ICFG	RSTP
LGUR	Ours	33.82	28.91	31.15
	我们的+BLIP（w=0.1）	34.97	29.14	33.15
	我们的方法+BLIP（w=1）	35.02	29.02	32.30
	我们+BLIP（退火）	34.19	27.72	30.40
	我们+B1+B2 BLIP+LDM（w=0.1）	36.39	29.85	**33.60**
	我们+B1+B2 BLIP+LDM（w=1）	**36.68**	**30.57**	31.20
IRRA	Ours	43.74	35.61	**38.00**
	我们+B1+B2 BLIP+LDM（w=1）	**46.71**	**36.64**	37.00

CUHK-PEDES

我们还为 CUHK-PEDES 合成生成了 11,003 个训练图文对。在此设置下，我们观察到性能提升了 0.44%。由于 LDM 和 BLIP 都是在我们的数据集上训练的，为 CUHK-PEDES 生成的图文对噪声更大，因此性能提升不如在我们的数据集上显著。不过，如果使用 CUHK-PEDES 训练 LDM 和 BLIP，我们认为这些结果将会有显著提升。

表4：使用合成数据在香港中文大学 CUHK-PEDES 数据集上的性能。报告 Rank-1 准确率。

设置	R-1
CUHK	65.35
香港中文大学+ B1+B2 BLIP+LDM (w=0.1)	65.58
香港中文大学+ B1+B2 BLIP+LDM (w=1)	**65.77**

6.3 在相同数据集设置下的性能

在表5中，我们展示了所有原始数据集上的结果。这里，模型在同一数据集上进行训练和测试。在我们的数据集上，IRRA 的 R-1 准确率达到 77.08%，而 LGUR 的 R-1 准确率为 65.96%。SSAN 的准确率为 57.56%。IRRA 的得分最高。

表5：所有数据集上的性能比较。报告 Rank-1 准确率

模型	CUHK-PEDES	ICFG-PEDES	RSTPReid	**Ours**
IRRA	**73.68**	**63.43**	**57.00**	**77.08**
LGUR	65.35	59.53	48.65	65.96
SSAN	60.71	54.05	38.80	57.56

6.4 跨数据集性能

文本到图像

表6展示了跨数据集测试的结果。在此设置中，我们在一个数据集上训练，并在其余三个数据集的测试集上进行测试。评估方案与各数据集所采用的方案相同。

表6：跨数据集文本到图像：在一个数据集上训练，并在其余测试集上测试。

训练	CUHK-PEDES						ICFG-PEDES					
Test	ICFG-PEDES		RSTPReid		**Ours**		CUHK-PEDES		RSTPReid		**Ours**	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
IRRA	43.13	23.37	53.30	40.05	64.56	72.84	33.48	31.56	45.30	36.82	43.82	53.79
LGUR	34.52	15.85	41.30	27.98	54.42	60.49	26.54	21.63	46.25	33.09	28.82	34.81
SSAN	24.47	10.59	13.75	8.935	37.12	43.83	15.50	13.37	35.50	28.30	16.36	21.38

训练	RSTPReid						**Ours**					
Test	CUHK-PEDES		ICFG-PEDES		**Ours**		CUHK-PEDES		ICFG-PEDES		RSTPReid	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
IRRA	32.49	30.30	31.23	19.36	26.32	35.79	43.74	39.72	35.62	17.51	38.00	28.57
LGUR	15.50	13.06	26.48	15.71	10.66	14.89	33.82	27.91	28.91	12.12	31.15	21.17
SSAN	10.04	8.39	20.19	12.15	4.92	7.83	21.86	17.96	18.65	7.26	9.90	6.12

表 7：跨数据集图像到文本

训练	CUHK-PEDES						ICFG-PEDES					
Test	ICFG-PEDES		RSTPReid		**Ours**		CUHK-PEDES		RSTPReid		**Ours**	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
IRRA	42.38	20.51	54.70	35.80	73.52	71.79	48.96	27.91	64.80	40.33	51.60	51.60
LGUR	30.71	14.55	44.9	27.65	59.84	56.7	37.54	20.22	60.60	36.01	35.56	33.62
SSAN	22.45	10.10	16.60	11.24	44.96	40.44	22.45	11.42	52.50	30.56	21.12	19.53

训练	RSTPReid						**Ours**					
Test	CUHK-PEDES		ICFG-PEDES		**Ours**		CUHK-PEDES		ICFG-PEDES		RSTPReid	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
IRRA	41.77	25.38	36.94	18.85	33.92	34.91	58.17	35.39	35.62	16.14	43.40	26.13
LGUR	19.91	10.96	33.61	16.11	13.12	13.08	45.41	25.06	26.30	11.88	35.60	21.72
SSAN	13.01	6.66	24.82	11.02	7.04	6.67	28.85	15.31	16.67	6.915	11.60	8.18

在 *CUHK-PEDES* 上训练。 我们首先在 CUHK-PEDES 上训练，并在 ICFG-PEDES、RSTPReid 和我们自己的数据集上测试。IRRA 在所有数据集和所有设置下均取得最佳结果，在 ICFG-PEDES 上达到 43.13%，在 RSTPReid 上达到 53.30%，在我们自己的数据集上达到 64.56%。对于 LGUR，我们在 ICFG-PEDES 上取得了 34.52% 的 R-1，在 RSTPReid 上取得了 41.30%，在我们自己的数据集上取得了 54.42%。

在所提出的数据集上训练。 当我们在自己的数据集上训练 IRRA，并在 CUHK-PEDES、ICFG-PEDES 和 RSTPReid 上测试时，分别取得了 43.75%、35.62% 和 38.00% 的 R-1 分数。其后是 LGUR 和 SSAN。

我们还推断，CUHK-PEDES 具有如此强的跨数据集性能，原因在于其词汇丰富，包含超过 9K 个唯一词语。虽然无法直接比较，但我们注意到，相较于在 ICFG-PEDES 或 RSTPReid 上训练并在 CUHK-PEDES 上测试，在我们的数据集上训练并在 CUHK-PEDES 上测试能取得不错的结果。例如，在所提出的数据集上训练的 IRRA 在 CUHK-PEDES 上取得了 43.75% 的 R-1。另一方面，在 ICFG-PEDES 上训练的 IRRA 在 CUHK-PEDES 上取得了 33.48%。

此外，我们在以 RSTPReid 和 Ours 作为训练集的情况下，在 ICFG-PEDES 上测试 IRRA，得到的 R-1 分别为 31.23% 和 35.62%。尽管 ICFG-PEDES 和 RSTPReid 都源自同一个数据集（MSMT17），在所提出的数据集上训练仍能取得更好的性能。

图像到文本

我们在表 7 中展示了 I2T 结果。我们观察到的情况与 T2I 类似。在所提出的数据集上训练的 IRRA 在 CUHK-PEDES、ICFG-PEDES 和 RSTPReid 上的得分分别为 58.16%、35.62% 和 43.40%。LGUR 在 CUHK-PEDES 上的 R-1 为 45.41%，在 ICFG-PEDES 上为 26.30%，在 RSTPReid 上为 35.60%。我们使用 ICFG-PEDES 和所提出的数据集作为训练集，在 CUHK-PEDES 上进行测试，发现对于所有模型而言，在我们的数据集上训练在 R-1 和平均精度均值方面都表现显著更好。因此，在所提出的数据集上训练能够取得丰富的跨数据集结果。

7 结论

本文介绍了一个新的文本-图像行人重识别数据集，该数据集在身份数量、图像采集环境和文本描述方面都十分丰富。我们还采用了最先进的基线模型，并进行了全面实验，以分析所提出的数据集。此外，我们训练了 LDM 和 BLIP 等生成模型来创建合成文本-图像对，并使用原始数据和合成数据，在同一数据集设置和跨数据集设置下进行了大量实验。结果表明，

除了原始数据之外，使用合成数据还能显著提升模型的性能。总而言之，所提出的数据集及相关实验为文本-图像行人重识别领域的未来研究提供了宝贵的见解和基准。

参考文献

[1] Surbhi Aggarwal、Venkatesh Babu Radhakrishnan 和 Anirban Chakraborty。2020。通过属性辅助匹配进行基于文本的行人搜索。载于 WACV。2617–2625。[2] Min Cao、Shiping Li、Juntao Li、Liqiang Nie 和 Min Zhang。2022。图像文本检索：近期研究与发展的综述。IJCAI（2022）。[3] Dapeng Chen、Hongsheng Li、Xihui Liu、Yantao Shen、Jing Shao、Zejian Yuan 和 Xiaogang Wang。2018。通过全局和局部图像-语言关联改进行人重识别的深度视觉表征。载于 ECCV。54–70。[4] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018。BERT：用于语言理解的深度双向变换器预训练。arXiv预印本 arXiv:1810.04805（2018）。[5] Zefeng Ding、Changxing Ding、Zhiyin Shao 和 Dacheng Tao。2021。用于文本到图像的部件感知行人重识别的语义自对齐网络。arXiv预印本 arXiv:2107.12666（2021）。[6] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xi- aohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。2020。一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer。arXiv预印本 arXiv:2010.11929（2020）。[7] Fartash Faghri、David J Fleet、Jamie Ryan Kiros 和 Sanja Fidler。2017。VSE++：利用困难负样本改进视觉-语义嵌入。arXiv预印本 arXiv:1707.05612（2017）。[8] Andrea Frome、Greg S Corrado、Jon Shlens、Samy Bengio、Jeff Dean、Marc'Aurelio Ranzato 和 Tomas Mikolov。2013。DeViSE：一种深度视觉-语义嵌入模型。NeurIPS 26（2013）。[9] Chenyang Gao、Guanyu Cai、Xinyang Jiang、Feng Zheng、Jun Zhang、Yifei Gong、Pai Peng、Xiaowei Guo 和 Xing Sun。2021。面向基于文本的行人搜索，在全尺度表征上进行上下文非局部对齐。arXiv预印本 arXiv:2101.03036（2021）。[10] Douglas Gray、Shane Brennan 和 Hai Tao。2007。评估用于识别、重新捕获和跟踪的外观模型。载于 PETS，第 3 卷。1–7。[11] Ari Holtzman、Jan Buys、Li Du、Maxwell Forbes 和 Yejin Choi。2020。神经文本退化的奇特现象。ICLR（2020）。[12] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 和 Weizhu Chen。2021。LoRA：大语言模型的低秩适配。arXiv预印本 arXiv:2106.09685（2021）。[13] Yan Huang、Wei Wang 和 Liang Wang。2017。利用选择性多模态长短期记忆网络进行实例感知的图像与句子匹配。载于 CVPR。2310–2318。[14] Chao Jia、Yinfei Yang、Ye Xia、Yi-Ting Chen、Zarana Parekh、Hieu Pham、Quoc Le、Yun-Hsuan Sung、Zhen Li 和 Tom Duerig。2021。通过噪声文本监督扩展视觉与视觉-语言表征学习。载于 ICML。PMLR，4904–4916。[15] Ding Jiang 和 Mang Ye。2023。面向文本到图像行人检索的跨模态隐式关系推理与对齐。CVPR（2023）。[16] Junnan Li、Dongxu Li、Caiming Xiong 和 Steven Hoi。2022。BLIP：用于统一视觉-语言理解与生成的语言-图像预训练引导。载于 ICML。PMLR，12888–12900。[17] Liunian Harold Li、Mark Yatskar、Da Yin、Cho-Jui Hsieh 和 Kai-Wei Chang。2019。VisualBERT：一种简单且高性能的视觉与语言基线模型。arXiv预印本 arXiv:1908.03557（2019）。[18] Shuang Li、Tong Xiao、Hongsheng Li、Wei Yang 和 Xiaogang Wang。2017。通过潜在协同注意力进行身份感知的文本-视觉匹配。载于 ICCV。1890–1899。[19] Shuang Li、Tong Xiao、Hongsheng Li、Bolei Zhou、Dayu Yue 和 Xiaogang Wang。2017。利用自然语言描述进行行人搜索。载于 CVPR。1970–1979。[20] Wei Li、Rui Zhao 和 Xiaogang Wang。2013。通过迁移度量学习实现人体重识别。载于 ACCV。31–44。[21] Wei Li、Rui Zhao、Tong Xiao 和 Xiaogang Wang。2014。DeepReID：用于行人重识别的深度滤波器配对神经网络。载于 CVPR。152–159。[22] Jiawei Liu、Zheng-Jun Zha、Richang Hong、Meng Wang 和 Yongdong Zhang。2019。用于基于文本的行人搜索的深度对抗图注意力卷积网络。载于 ACM MM。665–673。[23] Yu Liu、Yanming Guo、Erwin M Bakker 和 Michael S Lew。2017。学习用于多模态匹配的循环残差融合网络。载于 ICCV。4107–4116。[24] Jiasen Lu、Dhruv Batra、Devi Parikh 和 Stefan Lee。2019。ViLBERT：为视觉与语言任务预训练任务无关的视觉语言表征。NeurIPS 32（2019）。[25] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021。通过自然语言监督学习可迁移视觉模型。载于 *ICML*。PMLR，8748–8763。[26] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2022。利用潜在扩散模型进行高分辨率图像合成。载于 *CVPR*。10684–10695。[27] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。2015。U-Net：用于生物医学图像分割的卷积网络。载于 *MICCAI*。234–241。[28] Zhiyin Shao、Xinyu Zhang、Meng Fang、Zhifeng Lin、Jian Wang 和 Changxing Ding。2022。学习用于文本到图像行人重识别的粒度统一表示。载于 *ACM MM*。5566–5574。[29] Kihyuk Sohn。2016。采用多类 n 对损失目标改进深度度量学习。*NeurIPS* 29（2016）。[30] Hao Tan 和 Mohit Bansal。2019。LXMERT：从 Transformer 学习跨模态编码器表示。*arXiv*预印本 *arXiv:1908.07490*（2019）。[31] Hugo Touvron、Matthieu Cord、Matthijs Douze、Francisco Massa、Alexandre Sablayrolles 和 Hervé Jégou。2021。通过注意力训练数据高效的图像 Transformer 并进行蒸馏。载于 *ICML*。PMLR，10347–10357。[32] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017。注意力就是你所需要的一切。*NeurIPS* 30（2017）。[33] Yushuang Wu、Zizheng Yan、Xiaoguang Han、Guanbin Li、Changqing Zou 和 Shuguang Cui。2021。LapsCore：通过颜色推理进行语言引导的行人搜索。载于 *ICCV*。1624–1633。[34] Tong Xiao、Shuang Li、Bochao Wang、Liang Lin 和 Xiaogang Wang。2016。端到端行人搜索深度学习。*arXiv*预印本 *arXiv:1604.01850* 2，2（2016），4。[35] Ying Zhang 和 Huchuan Lu。2018。用于图像文本匹配的深度跨模态投影学习。载于 *ECCV*。686–701。[36] Kecheng Zheng、Wu Liu、Jiawei Liu、Zheng-Jun Zha 和 Tao Mei。2020。用于基于文本的行人搜索的分层 Gumbel 注意力网络。载于 *ACM MM*。3441–3449。[37] Liang Zheng、Liyue Shen、Lu Tian、Shengjin Wang、Jingdong Wang 和 Qi Tian。2015。可扩展的行人重识别：一个基准。载于 *ICCV*。1116–1124。[38] Zhedong Zheng、Liang Zheng、Michael Garrett、Yi Yang、Mingliang Xu 和 Yi-Dong Shen。2020。带实例损失的双路径卷积图像文本嵌入。*ACM*多媒体计算、通信与应用汇刊 16，2（2020），1–23。[39] Aichun Zhu、Zijie Wang、Yifeng Li、Xili Wan、Jing Jin、Tian Wang、Fangqiang Hu 和 Gang Hua。2021。DSSL：面向基于文本的行人检索的深度环境与行人分离学习。载于 *ACM MM*。209–217。