权 利 要 求 书

1、一种基于可扩展表征学习的行人重识别方法，其特征在于，包括以下步骤：

步骤 S1：获取训练样本，每一个训练样本包括：训练图像及与训练图像对应的训练文本描述；

步骤 S2：利用图像-文本编码器分别提取所述训练图像及训练文本描述的特征，得到训练图像特征和训练文本特征；

步骤 S3：利用非文本对齐的视觉编码器提取训练图像特征，优先考虑图像本身，在语义空间中有效地分离相似图像。然后通过向量余弦相似度获取与主干模型相对应的维度的图像以及文本特征，该特征即为示范样本特征；

步骤 S4：利用 do 算子的几何特征进行因果表征学习，它采用结构因果模型，将数据集偏差视为混杂因素，并随后消除它们的混杂效应；

步骤 S5：训练模型至收敛。然后将所述行人图像特征和文本特征输入训练后的行人重识别模型，得到行人重识别结果。

2、根据权利要求 1 所述的可扩展表征学习的行人重识别方法，其特征在于，所述步骤 S2 包括：

通过 CLIP 模型的图像特征提取器和文本特征提取器分别从训练图像提取训练图像特征，对所述的训练文本描述提取训练文本特征；

3、根据权利要求 1 所述的可扩展表征学习的行人重识别方法，其特征在于，所述步骤 S3 包括：

利用非文本对齐的视觉编码器 EVA 提取训练图像特征。利用 CLIP 的文本特征提取器提取示范样本特征；

具体地，所述步骤 S2 和步骤 S3 中，所述图像和文本特征提取器分别为 CLIP 模型的图像特征提取器和 CLIP 的文本特征提取器，以及 EVA 图像特征提取器。由于来自 CLIP 的与文本对齐的视觉图像编码器依赖于文本且难以区分相似图像，因此利用另一种同样具有 ViT 架构的非文本对齐的视觉编码器 EVA 作为语义空间的补充。

4、根据权利要求 1 所述的可扩展表征学习的行人重识别方法，其特征在于，所述步骤 S4 包括：

需要考虑不同空间的 ViT 视觉编码特征的适当融合策略，这里通过因果注意力机制来融合这两种表征，增强视觉特征的鲁棒性。由于跨模态行人重识别任务主要利用文本检索图库中行人图像，所以来自 CLIP 模型的特征仍然需要处于主导地位，因此应用因果注意力机制进行特征交互采用一个单向过程，如下所示：

$$\mathcal{M}[i,j]=\begin{cases}0, & 0<i<\mathbb{M},\mathbb{M}\le j<(\mathbb{M}+\mathbb{S})\\ 1, & \text{others}\end{cases}$$

其中 M 是由具有文本监督的 CLIP 模型的图像编码器编码的视觉表征，S 是由非文本对齐的视觉编码器编码的视觉表征；此外，i 和 j 是行和列索引，$M[i,j]$ 表示第 i 个图像在 Transformer 层中可以关注到第 j 个图像的表征；最后，双重视觉表征 $f_i^{'m}$ 的表达式为：

$$f_i^{'m}=[f_i^m,\tilde{f}_i^{\tilde{m}}]$$

同样，文本表征学习执行因果注意力机制如下：

$$\mathcal{W}[i,j]=\begin{cases}0, & 0<i<\mathbb{D},\mathbb{D}\le j<(\mathbb{D}+\mathbb{T})\\ 1, & \text{others}\end{cases}$$

其中 $\mathbb{T}$ 和 $\mathbb{D}$ 是由 CLIP 模型的文本编码器编码的文本表征；此外，i 和 j 是行和列索引，$W[i,j]$ 表示第 i 个文本在 Transformer 层中可以关注到第 j 个文本的表征。