arXiv:2602.01059v1 [cs.CV] 1 Feb 2026

DRFormer：用于行人重识别的双重正则化双向变换器

Ying Shu [1] **Pujian Zhan** [1] **Huiqi Yang** [1] **Hehe Fan** [2] **Youfang Lin** [1] **Kai Lv** [1]

摘要

细粒度判别细节和全局语义特征都有助于解决行人重识别中的遮挡、姿态变化等挑战。视觉基础模型（例如，DINO）擅长挖掘局部纹理，而视觉语言模型（例如，CLIP）能够捕捉显著的全局语义差异。现有方法主要依赖单一范式，忽视了将二者结合的潜在优势。本文分析了这两种架构的互补作用，并提出一种框架，通过双重正则化双向**Transformer（DRFormer）**协同发挥它们的优势。双重正则化机制确保了特征提取的多样性，并使两种模型的贡献达到更好的平衡。在五个基准数据集上进行的大量实验表明，我们的方法能够有效协调局部和全局表征，在性能上与最先进方法相比具有竞争力。

1. 引言

行人重识别（ReID）旨在根据查询图像从图库中检索目标行人。由于模糊、遮挡、光照变化以及巨大的姿态变化，ReID任务极具挑战性。我们期望模型能够捕捉整体人体结构，**并**关注相似个体之间的细微差异，例如配饰和服装纹理。

许多现有的行人再识别方法专注于具有判别性的细粒度细节（Zhu 等人，2022b；Hu 等人，2025）**或**全局语义特征（Li 等人，2023b；Yang 等人，2024；Wei 等人，2025）。一方面，DINO 等视觉基础模型（VFMs）（Caron 等人，2021年；Oquab 等人，2024）在细粒度特征挖掘方面展现出了强大的能力。例如，有研究采用灾难性遗忘分数来衡量 DINO 的预训练数据与微调数据之间的差距，随后设计了一个行人重识别专用模块（Luo 等人，2021年）。另一方面，CLIP 等视觉语言模型（VLMs）（Radford 等人，2021年）擅长提取高层语义特征。CLIP-ReID（Li 等人，2023b）通过充分利用 CLIP 的跨模态描述能力，将 CLIP 应用于行人再识别。然而，在行人再识别领域，大多数现有方法主要依赖这两种视角中的一种，忽视了将二者结合所带来的潜在益处。目前的文献对同时利用 DINO 的局部细粒度判别能力与 CLIP 的全局语义理解能力，探索仍然不足。

在本文中，我们首先探讨一个问题：**DINO与CLIP在行人再识别中分别扮演何种具体角色？** 1）如图1（a）和（c）所示，我们观察到一个明显区别：DINO展现出关注细粒度局部特征的强大能力，注意力峰值沿对角线出现；而CLIP仅将注意力集中在少数固定词元上，缺乏局部分辨能力。2）图1（b）和（d）展示了微调后的DINO与CLIP在行人身份上的特征分布。我们发现，DINO特征高度重叠且分散，而CLIP特征呈现聚类结构。其原因在于，全局语义上下文有助于行人再识别中的类间分离。上述比较揭示了一个极具吸引力的机会：DINO提供必要的细粒度纹理，而CLIP确保关键的语义可分性。因此，提升再识别性能的关键在于协调这些截然不同却互补的优势。

受这一洞见启发，我们着手解决一个关键问题：**如何有效结合 DINO 和 CLIP 的能力？** 为此，我们围绕两项关键原则设计了框架：对于每个单独的模型，我们力求增强特征多样性，认为更多样化的信息能够为有效融合提供更丰富的基础。对于两个模型的融合，我们优先考虑有效协作，以充分整合它们互补的优势。

本文提出了双重正则化双向Transformer （DRFormer）。我们采用双向交

[1]网络科学与智能系统研究院，北京交通大学，北京，中国 [2]浙江大学，杭州，浙江，中国。通讯作者：Kai Lv< lvkai@bjtu.edu.cn>。

预印本。*2026*年*2*月*3*日。

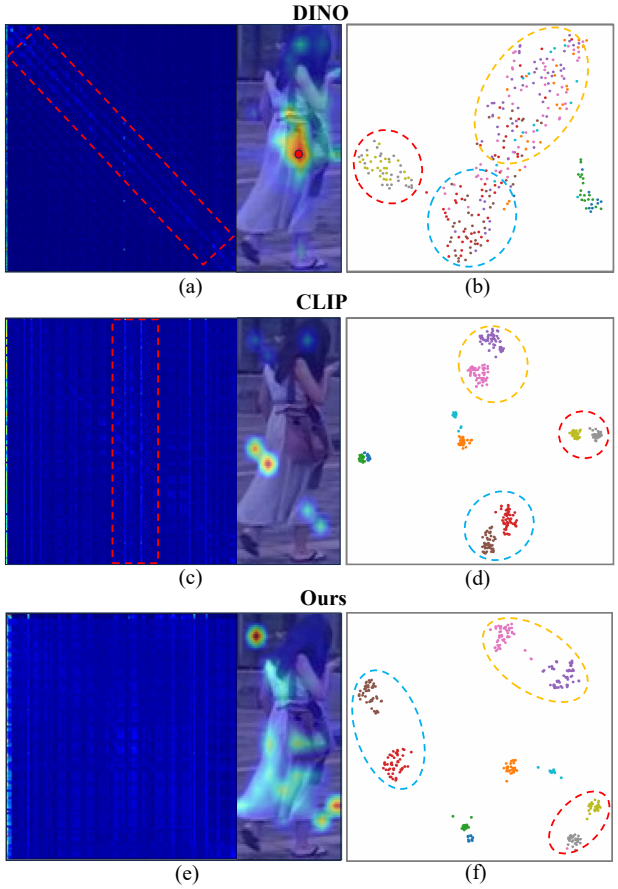


图*1*。DINO、CLIP 与我们的方法在注意力模式和特征分布上的比较。左列展示了所有词元及一个选定词元的注意力权重：图 (a) 显示红点处的查询图像词元，而图 (c) 和 (e) 展示了 [CLS] 词元的注意力权重。右列展示了微调后的DINO、微调后的CLIP 和我们的方法在部分样本上的特征分布。

叉注意力机制，使 DINO 和 CLIP 的特征交替作为查询，以促进深度交互。此外，我们引入了两个辅助正则化器。模型内令牌多样性正则化器旨在促进可学习令牌学习多样化特征。我们最大化这些令牌之间的余弦距离，以驱动模型关注互补区域。模型间偏差正则化器旨在消除模型间学习抑制，并提升两个模型的贡献比例。在理论分析的指导下，我们将该正则化器表述为最小化每个独立模型的预测误差。我们的贡献可概括如下：

- 针对什么问题，我们分析并验证 DINO 和 CLIP 的互补作用：DINO 能捕捉细粒度分类所需的局部判别模式，而 CLIP 则施加全局语义约束，以增强身份可分性。
- 针对如何解决这一问题，我们提出了一种双向 Transformer 模块，并引入了两个正则化器。模型内正则化器鼓励充分利用多样化的表征，而模型间正则项则平衡 DINO 和 CLIP 的贡献。
- 在五个行人重识别数据集上，与最先进的行人重识别方法相比，我们的模型取得了具有竞争力的性能，证明了所提方法的有效性。

2. 相关工作

2.1. 行人重识别

行人重识别旨在跨不重叠的摄像机视角匹配行人，即使存在姿态变化和遮挡，也能识别出同一身份。该领域的研究已从基于 CNN 的架构发展到基于 Transformer 的方法。基于 CNN 的方法（Sun 等人，2018；He 等人，2023）主要通过度量学习来学习判别性特征表征。如今，基于 Transformer 的方法（He 等人，2021；Zhu 等人，2022a）被提出，并利用自注意力建模长程依赖关系。开创性工作 TransReID（He 等人，2021）引入拼图块和侧信息嵌入，以增强特征鲁棒性。本文提出一种基于 Transformer 的框架，以实现鲁棒且具有判别性的行人重识别。

2.2. 大规模预训练模型

视觉基础模型（VFMs）（He 等人，2022；Kirillov 等人，2023；Oquab 等人，2024）和视觉语言模型（VLMs）（Radford 等人，2021年；Jia 等人，2021年；Zhai 等人，2023）在大规模数据集上预训练，并展现出强大的图像理解能力。VFMs 通常作为视觉骨干网络，并被广泛应用于许多下游任务（Li 等人，2023a；Wang 等，2025）。DINO（Caron 等人，2021年；Oquab 等人，2024）通过直接预测教师网络的输出简化了自监督训练，并擅长捕捉细粒度细节。掩码自编码器（He 等人，2022）显著改进了面向视觉 Transformer 的掩码图像建模（MIM）（Dosovitskiy 等人，2021年）。与此同时，VLMs 在多样化任务中展现出强大的泛化能力。CLIP（Radford 等人，2021年）通过对比学习在共享嵌入空间中对齐图像与文本，从而有效捕捉全局语义上下文。SigLIP（Zhai 等人，2023）通过采用基于 sigmoid 的目标函数进一步增强了这一能力。

许多 ReID 方法基于现有的视觉基础模型或视觉语言模型构建。基于 DINO，PersonViT（Hu 等人，2025）结合了

MIM 与判别式对比学习，以提取全局和局部特征。CLIP-ReID（Li 等人，2023b）采用提示学习，为每个身份生成文本标记，并充分利用 CLIP 的先验知识。PromptSG（Yang 等人，2024）进一步通过将图像特征转换为伪标记来提升泛化能力。不同于这些仅关注单一方面的工作，我们的方法旨在结合视觉基础模型和视觉语言模型的优势，以取得更好的性能。

2.3. 特征融合方法

特征融合整合不同层级的信息，以获得更强的表征。最简单的方法是直接拼接，但这需要仔细对齐特征图维度。随着 Transformer 的引入，许多基于注意力的特征融合方法（Xu 等人，2018；Zhou 等人，2022）相继提出。双重注意力网络（DANet）旨在自适应地将局部特征与其全局依赖关系整合起来（Fu 等人，2019）。GCNet（Cao 等人，2019）利用全局上下文信息增强局部特征，同时显著降低计算复杂度。对于预训练模型的特征融合，由于其固有的特征层级不匹配，难度更大。在单目深度估计（MDE）等其他领域，一些研究（Zhang 等人，2025）利用文本引导 DINO 与 CLIP 的对齐。DeCLIP（Wang 等人，2025）通过蒸馏 DINO 的注意力图，增强 CLIP 稠密特征的判别性和空间一致性，但这需要使用大规模数据集进行训练。本文通过结合双重正则化机制的双向 Transformer，实现了 DINO 特征与 CLIP特征的高效融合。

3. 提出的方法

3.1. 关键观察与动机

在回答如何整合 DINO 和 CLIP 的能力这一问题之前，我们首先开展实验，以确定特征融合过程中的关键挑战。如图2所示，这些挑战体现在两个方面：

DINO 和 CLIP 内部的特征利用不足。为了充分提取 DINO 和 CLIP 中的有用信息，我们认为仅依赖分类损失是不够的。如图2 （a）所示，我们观察到，不同可学习令牌的注意力区域几乎完全相同。这意味着同一模型中的不同可学习令牌倾向于提取相对同质的特征

特征。此外，两个词元都未能关注背包区域，而该区域可以为区分相似个体提供重要线索。这些观察表明，需要施加额外约束，以促进每个模型内部多样且互补的特征提取。

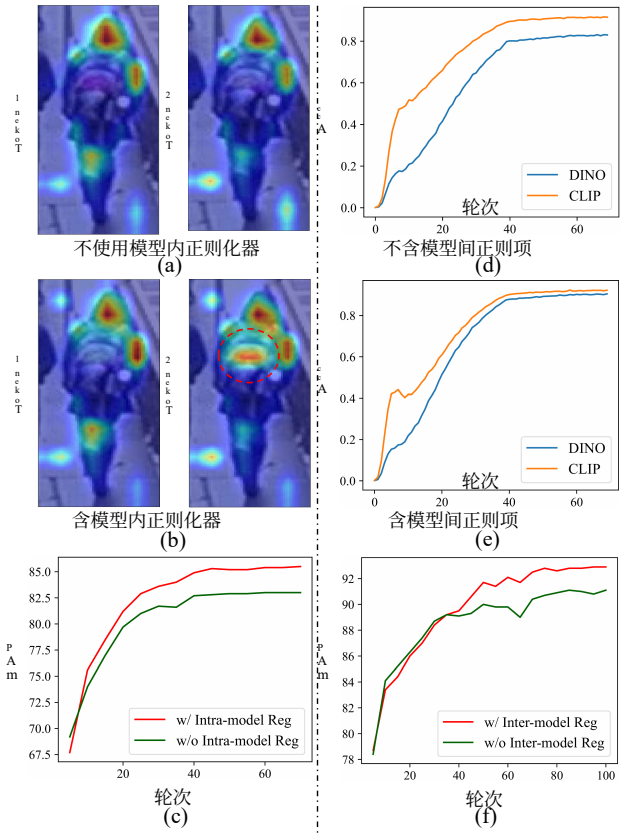


图2. 左、右两列分别展示了模型内正则化器和模型间正则化器的动机及效果。左侧：图(a)和图(b)展示了DukeMTMC数据集上，添加模型内正则化器前后两个可学习令牌的注意力权重。图(c)显示了性能提升（2.5% ↑）。右侧：图(d)和图(e)展示了Market-1501数据集上，添加模型间正则化器前后DINO和CLIP的准确率曲线。图(f)显示了性能提升（1.8% ↑）。

为鼓励词元提取多样化的特征，我们引入了**模型内令牌多样性正则化器**。如图 2(b) 所示，标记 2 的注意力图突出了背包区域，体现了所学习特征的多样性。图 2(c) 显示，该正则化器在 DukeMTMC 上带来了 2.5% 的性能提升，验证了其有效性。

DINO 和 CLIP 的不平衡学习。由于预训练范式和数据集不同，DINO 和 CLIP 提取的特征存在显著差异。尽管 DINO 和 CLIP 的配置对称，我们推测它们的学习过程本质上是不平衡的。为验证这一点，我们在统一框架内联合训练了两个模型，但在 Market-1501 上分别评估了各分支的性能。如图所示

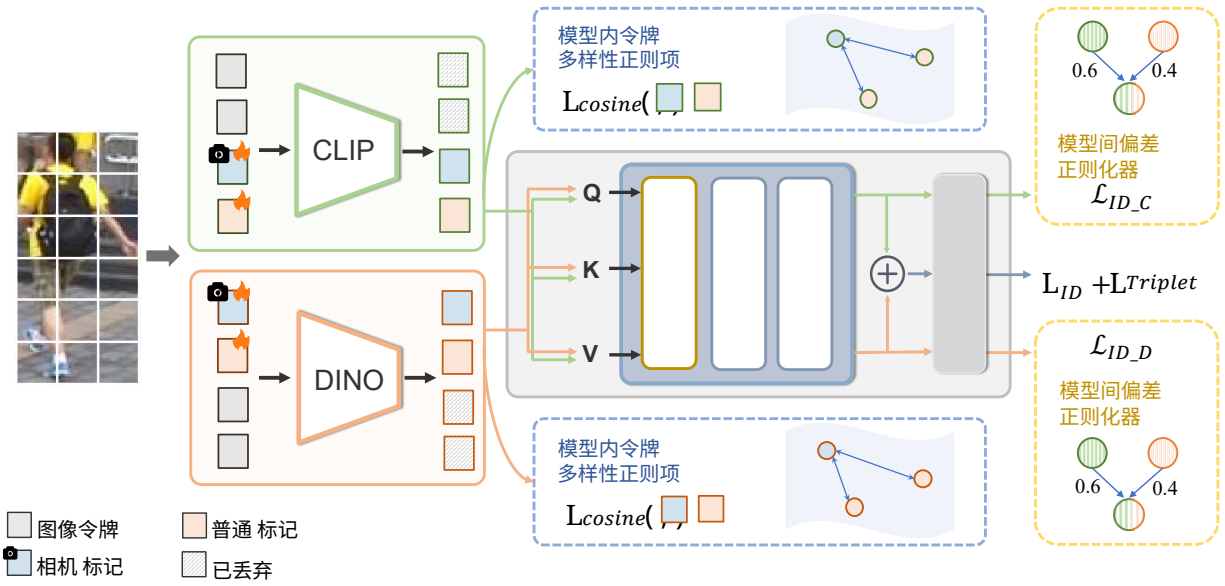


图*3.* 我们方法的框架。DRFormer主要由CLIP图像编码器、DINO编码器和双向融合变换器组成。输入图像被分割为图像块，并与可学习令牌一起分别输入CLIP和DINO。两个模型中可学习令牌的输出有两个用途：执行模型内正则化，以及作为交叉注意力的查询。融合模块的两个输出被拼接后送入线性分类器。此外，每个输出也会单独传入分类器，以进行模型间正则化。

所示2(d)，在训练初期，CLIP 的准确率提升速度显著快于 DINO。随着训练过程逐渐收敛，两者之间形成了稳定且平行的差距。实验结果表明，DINO 和 CLIP 编码器的学习难度各不相同。

为协调它们各自的贡献，我们引入了**模型间偏差正则化器**。图2 （e）和（f）展示了引入模型间正则项后的实验结果。该正则化器有效消除了DINO与CLIP之间的准确率差距，并在Market-1501上大幅提升了性能。

3.2. 双向融合变换器

我们的方法旨在有效利用 DINO 和 CLIP 的互补优势。DINO 擅长细粒度细节提取，而 CLIP 能够捕捉全局语义上下文。我们同时使用 DINO 编码器 $\mathcal{E}_{\text{DINO}}$ 和 CLIP 图像编码器 $\mathcal{E}_{\text{CLIP}}$ 从同一幅图像中提取互补特征。具体而言，我们将图像划分为 M 个互不重叠的图像块，并将其嵌入为图像块词元序列，分别记为 DINO 的 $\boldsymbol{I}^D_{1\sim M} \in \mathbb{R}^{M\times d_D}$ 和 CLIP 的 $\boldsymbol{I}^C_{1\sim M} \in \mathbb{R}^{M\times d_C}$，其中 d_D 和 d_C分别是两个编码器的词元嵌入维度。我们为每个模型引入 N个可学习令牌（即，$\boldsymbol{T}^D_{1\sim N} \in \mathbb{R}^{N\times d_D}$和 $\boldsymbol{T}^C_{1\sim N} \in \mathbb{R}^{N\times d_C}$）。这些令牌与图像词元拼接后，分别输入两个模型，如公式 1 和 2所示。

$$[\boldsymbol{f}^D_T; \boldsymbol{f}^D_I] = \mathcal{E}_{\text{DINO}}([\boldsymbol{T}^D_1, ..., \boldsymbol{T}^D_N; \boldsymbol{I}^D_1, ..., \boldsymbol{I}^D_M]) \quad (1)$$

$$[\boldsymbol{f}^C_T; \boldsymbol{f}^C_I] = \mathcal{E}_{\text{CLIP}}([\boldsymbol{T}^C_1, ..., \boldsymbol{T}^C_N; \boldsymbol{I}^C_1, ..., \boldsymbol{I}^C_M]) \quad (2)$$

可学习令牌用作查询，从两个模型中提取有用信息以进行特征融合。为降低计算成本，我们舍弃图像词元的输出。只有与可学习令牌对应的输出特征，即即 $\boldsymbol{f}^D_T$ 和 $\boldsymbol{f}^C_T$，会被送入融合模块。

在这里，我们使用双向 Transformer 融合 DINO 和 CLIP 的特征。融合模块由一个交叉注意力层和两个自注意力层组成。通过执行双向交叉注意力，利用 CLIP 的语义上下文从 DINO 中查询细粒度视觉细节，同时利用从 DINO 中捕获的细节查询 CLIP 中编码的全局语义。具体而言，$\boldsymbol{f}^D_T$和 $\boldsymbol{f}^C_T$交替用作查询，而另一个则用作键和值，如公式3 和 4所示。随后，我们可以得到两个输出特征，即 $\mathbf{H}_{D\to C}$ 和 $\mathbf{H}_{C\to D}$。 $\mathbf{H}_{D\to C}$ 和 $\mathbf{H}_{C\to D}$是通过 DINO 与 CLIP 之间的双向交互得到的两个增强特征。

$$\mathbf{H}_{D\to C} = \text{Transformer}\big(\mathbf{Q} = \boldsymbol{f}_T^D, \mathbf{K} = \boldsymbol{f}_T^C, \mathbf{V} = \boldsymbol{f}_T^C\big) \tag{3}$$

$$\mathbf{H}_{C\to D} = \text{Transformer}\big(\mathbf{Q} = \boldsymbol{f}_T^C, \mathbf{K} = \boldsymbol{f}_T^D, \mathbf{V} = \boldsymbol{f}_T^D\big) \tag{4}$$

将两个查询得到的输出拼接起来，并用作简单线性分类器 f_{linear} 的输入。具体可表述如下:

$$\hat{y} = f_{linear}([\mathbf{H}_{D\to C}; \mathbf{H}_{C\to D}]) \tag{5}$$

最终，我们采用标准的行人重识别损失，即，三元组损失 $\mathcal{L}_{\text{Tri}}$ 和身份分类损失 $\mathcal{L}_{\text{ID}}$ (He 等人，2023)，以优化我们的网络。这两种损失的计算方式如下:

$$\mathcal{L}_{ID} = \sum_{j=1}^{K} -q_j \log(p_j) \tag{6}$$

$$\mathcal{L}_{Tri} = \max\left(d_p - d_n + \alpha, 0\right) \tag{7}$$

其中， q_j 表示目标分布中的值， p_j表示类别 j的身份预测逻辑值， d_p 和 d_n 分别表示正负样本对的特征距离，α 表示 $\mathcal{L}_{Tri}$的间隔。

3.3. 模型内令牌多样性正则化器

模型内正则化器旨在使不同的可学习词元捕获多样化的表征。首先，我们使用侧信息嵌入（SIE）（He 等人，2021年）将摄像头信息融入第一个可学习词元。我们分别将摄像头信息嵌入 $\boldsymbol{E}_{cam}$添加到 DINO 和 CLIP 的第一个可学习词元中，如公式8和9所示。

$$\boldsymbol{T}_1^D = \boldsymbol{T}_1^D + \boldsymbol{E}_{cam} \tag{8}$$

$$\boldsymbol{T}_1^C = \boldsymbol{T}_1^C + \boldsymbol{E}_{cam} \tag{9}$$

接着，我们最大化余弦距离，即最小化第一个标记与普通可学习标记之间的余弦相似度，如公式10 和11所示。这会促使这些标记捕获多样且互补的信息。因此，第一个标记能够在摄像头信息的引导下聚合更全面的特征，而其余标记则通过提示来学习视角不变表征。

$$\mathcal{L}_{cos}^D = \frac{1}{N-1}\sum_{i=2}^{N} cos(\boldsymbol{f}_1^D, \boldsymbol{f}_i^D) \tag{10}$$

$$\mathcal{L}_{cos}^C = \frac{1}{N-1}\sum_{i=2}^{N} cos(\boldsymbol{f}_1^C, \boldsymbol{f}_i^C) \tag{11}$$

$$\mathcal{L}_{intra} = \mathcal{L}_{cos}^D + \mathcal{L}_{cos}^C \tag{12}$$

3.4. 模型间偏差正则化器

模型间正则项旨在找到 DINO 与 CLIP 之间的最优贡献比例。我们刚刚将融合模块的两个输出即，$\mathbf{H}_{D\to C}$ 和$\mathbf{H}_{C\to D}$，拼接起来并输入分类器，如公式5所示。线性分类器可表示为:

$$\boldsymbol{z}_f = [\mathbf{H}_{D\to C}; \mathbf{H}_{C\to D}] \tag{13}$$

$$f_{linear} = \boldsymbol{W}\boldsymbol{z}_f + \boldsymbol{b} \tag{14}$$

公式14 可进一步展开，得到以下公式:

$$f_{linear} = \boldsymbol{W}_D \cdot \mathbf{H}_{D\to C} + \boldsymbol{b}_D + \boldsymbol{W}_C \cdot \mathbf{H}_{C\to D} + \boldsymbol{b}_C \tag{15}$$

$$\boldsymbol{s}^D = \boldsymbol{W}_D \cdot \mathbf{H}_{D\to C} + \boldsymbol{b}_D \tag{16}$$

$$\boldsymbol{s}^C = \boldsymbol{W}_C \cdot \mathbf{H}_{C\to D} + \boldsymbol{b}_C \tag{17}$$

其中， $\boldsymbol{s}^D$ 表示 DINO 的逻辑值输出， $\boldsymbol{s}^C$表示 CLIP 的逻辑值输出。因此，最终输出是 $\boldsymbol{s}^D$ 与 $\boldsymbol{s}^C$ 之和。因此，优化梯度由 $\boldsymbol{s}^D$ 和 $\boldsymbol{s}^C$决定。

假设两个模型输出的贡献比例分别为w_0 和 w_1，且 $w_0 + w_1 = 1$。如公式18所示，最优解应使泛化误差 $\mathcal{L}$ 最小化。

$$\min_{w_0,w_1} \; g = \mathbb{E}\mathcal{L}\big(f(s), y\big) \tag{18}$$

$$f(s) = w_0 s^D + w_1 s^C \tag{19}$$

通过对泛化误差进行偏差-方差分解（Shahhosseini 等人，2019），解可改写如下:

$$g = (Bias(f(s), y))^2 + Var(f(s)) + Var(\epsilon) \tag{20}$$

$$w_0 = \frac{Bias(s^C, y)}{Bias(s^C, y) - Bias(s^D, y)} \tag{21}$$

$$w_1 = \frac{-Bias(s^D, y)}{Bias(s^C, y) - Bias(s^D, y)} \tag{22}$$

其中， $Bias(.)$ 表示预测偏差， $Var(.)$ 表示 $f(s)$ 的方差。 $Bias(s^D, y) = \mathbb{E}[s^D - y]$ 和 $Bias(s^C, y) = \mathbb{E}[s^C - y]$ 分别衡量 DINO 和 CLIP 的预测误差。然而，在公式 21 和 22 中， w_0 或 w_1 必须小于 0，这与 $w_0 > 0$ 和 $w_1 > 0$ 相矛盾。因此，在约束条件 $w_0 + w_1 = 1$下，我们无法求得该方程的数值解。理论推导的更多细节见附录 A。

在公式20中，由于我们不能仅依靠重新加权来最小化误差项 $(Bias(f(s)\ y))^2$,，最有效的替代方案是最小化每个独立模型的内在预测偏差，即， $Bias(s^D, y)$ 和 $Bias(s^C, y)$。由此得到以下命题：

命题 3.1。 给定加权融合 $f(s) = w_0 s^D + w_1 s^C$，其中 w_0、 $w_1 > 0$ 和 $w_0 + w_1 = 1$，最小化泛化误差需要最小化各分支的预测偏差。

因此，保持 DINO 和 CLIP 各自互补的判别优势，即分别发挥局部细节建模和全局语义理解的作用，对于实现有效融合至关重要。我们为每个模型引入一个正则化项以实现这一目标。具体而言，我们分别复制 $\mathbf{H}_{D\to C}$ 和 $\mathbf{H}_{C\to D}$，然后将它们分别输入权重共享分类器 f_{linear}，其中复制特征是为了对齐输入维度。我们使用身份分类损失作为正则化损失：

$$\hat{y}_D = f_{linear}([\mathbf{H}_{D\to C}, \mathbf{H}_{D\to C}]) \tag{23}$$

$$\hat{y}_C = f_{linear}([\mathbf{H}_{C\to D}, \mathbf{H}_{C\to D}]) \tag{24}$$

$$\mathcal{L}_{inter} = \mathcal{L}_{ID_D} + \mathcal{L}_{ID_C} \tag{25}$$

3.5. 训练损失

目标函数可表示如下，由行人重识别损失、模型内正则化损失和模型间正则化损失组成：

$$L_{total} = L_{ID} + L_{Tri} + \lambda_1 L_{inter} + \lambda_2 L_{intra} \tag{26}$$

其中， λ_1 和 λ_2 是超参数。

4. 实验

4.1. 实验设置

数据集与评估协议。 在本研究中，我们在五个人员再识别数据集上评估我们的方法，包括 Market-1501（Zheng 等人，2015）、MSMT17（Wei 等人，2018）、DukeMTMC（Ristani 等人，2016）、CUHK03-NP（Li 等人，2014）和 Occluded-Duke（Miao 等人，2019）。

表 *1.* 我们实验中数据集的统计信息。

数据集	#ID	图像	Cams
Market-1501	1501	32668	6
MSMT17	4101	126441	15
DukeMTMC	1404	36411	8
CHUK03-NP	1467	13164	2
Occuled-Duke	1404	35489	8

这些数据集的详细信息汇总于表 1。遵循现有的行人再识别设置，我们使用平均精度均值（mAP）和 Rank-1（R-1）准确率来评估性能。

实现细节。 我们采用 DINOv2（Oquab 等人，2024）和 CLIP （ Radford 等人，2021 年）的视觉编码器作为骨干网络。我们为 DINO 编码器选择 ViT-B/14，为 CLIP 编码器选择 ViT-B/16。二者均包含 12 个 Transformer 层，隐藏维度为 768 。所有输入图像均调整为 252×126 （用于 DINO）和 256×128 （用于 CLIP）。可学习令牌的数量默认设为 2。这些令牌以维度为 $\mathbb{R}^{N\times 768}$的随机值初始化。融合Transformer由一个交叉注意力层和两个自注意力层组成。最终分类器是一个简单的线性层，用于将特征映射到分类标签。我们使用 Adam 优化器，学习率设为 5e-6，并在所有数据集中始终将 $\lambda_1 = 0.5$ 和 $\lambda_2 = 5$设为公式 26 中的值。由于余弦距离损失的取值范围为 $[0, 1]$,，因此引入系数 λ_2 将其重新缩放至与其他损失项相同的量级。模型训练 70 轮。整个框架使用 PyTorch 实现，并在配备 24GB 显存的单块 NVIDIA RTX4090 GPU 上运行。

基线。 基线基于预训练 CLIP，其视觉编码器通过两种标准损失进行微调，即三元组损失和身份分类损失。

4.2. 与最先进方法的比较

表 2总结了与基于 ViT 的方法的比较结果。DRFormer 在每个基准上都以较大间隔超越了第二优方法和基线。具体而言，我们的方法在 DukeMTMC 上达到了 85.5% 的平均精度均值和 92.5% 的 Rank-1，分别比 HPL（Zhou 等人，2025）方法高 2.6% 和 2.2%。在 Market-1501 和 MSMT17 数据集上，就平均精度均值而言，我们的方法分别比 FlexiReID（Sun 等人，2025）高 0.8% 和 11.2%。与 PHA（Zhang 等人，2023）相比，DRFormer 在 CUHK03-NP 数据集上的平均精度均值/Rank-1 高出 5.1%/5.1%。实验结果证明了我们的方法在五个数据集上的有效性，并凸显了其强劲性能。

表2. 在Market-1501、MSMT17、DukeMTMC、CUHK03-NP（标注版）和Occluded-Duke数据集上与最先进的基于 ViT 的方法的比较。最佳性能以**粗体**突出显示，次佳结果以下划线标出。

方法	参考文献	Market-1501		MSMT17		DukeMTMC		CUHK03-NP		Occluded-Duke	
		mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
TransReID (He 等人, 2021年)	2021年国际计算机视觉大会	88.9	95.2	67.4	85.3	82.0	90.7	79.6	81.7	59.2	66.4
DINO(Caron 等人, 2021年)	2021年国际计算机视觉大会	90.3	95.4	64.2	83.4	-	-	-	-	-	-
DINO+CFS(Luo 等人, 2021年)	Arxiv'21	91.0	96.0	66.1	84.6	-	-	-	-	-	-
DCAL (Zhu 等人, 2022a)	CVPR'22	87.5	94.7	64.0	83.1	80.1	89.0	-	-	-	-
DC-Former(Li 等人, 2023c)	AAAI'23	90.6	96.0	70.7	86.9	-	-	79.4	81.6	-	-
CLIP-ReID（Li 等人，2023b）	AAAI'23	89.6	95.5	73.4	88.7	82.5	90.0	81.6	80.9	59.5	67.1
AAformer（Zhu 等人，2024）	TNNLS'23	88.0	95.4	65.6	84.4	80.9	90.1	79.0	80.3	-	-
PHA（Zhang 等人，2023）	CVPR'23	90.2	96.1	68.9	86.1	-	-	83.0	84.5	-	-
CLIP3DReID(Liu 等人，2024)	CVPR'24	88.4	95.6	61.2	81.5	-	-	-	-	-	-
FlexiReID(Sun 等人，2025)	ICML'25	92.1	96.0	67.5	83.7	-	-	-	-	-	-
HPL (Zhou 等人，2025)	AAAI'26	89.8	95.9	**79.0**	91.0	82.9	90.3	-	-	-	-
基线		86.4	93.3	66.1	84.4	80.0	88.8	80.0	80.5	53.5	60.8
DRFormer		**92.9**	**97.0**	78.7	**91.3**	**85.5**	**92.5**	**88.1**	**89.6**	**65.3**	**72.1**

表3. 在 Market-1501 数据集上对双向 Transformer 的消融实验。

方法	Market-1501		DukeMTMC	
	mAP	R-1	mAP	R-1
基线	86.4	93.3	80.0	88.8
+拼接	89.5	94.7	81.6	90.6
+Transformer	**91.1**	**96.4**	**82.8**	**91.0**

表4. 在 DukeMTMC 数据集上对模型内和模型间正则化器进行的消融研究。

模型内	模型间	Market-1501		DukeMTMC	
		mAP	R-1	mAP	R-1
-	-	91.1	96.4	82.8	91.0
✓	-	91.5	96.3	84.5	92.3
-	✓	92.5	96.8	83.0	91.0
✓	✓	**92.9**	**97.0**	**85.5**	**92.5**

4.3. 消融研究

不同组件的贡献。 如表3所示，我们在 Market-1501 和 DukeMTMC 上进行了实验，以研究各组件的作用。“Con-cat”指直接拼接 CLIP 和 DINO 的输出，而“Transformer”表示经过双向 Transformer 融合后再进行特征拼接。实验结果表明，引入 DINO 相较于基线可带来显著提升，使 Market-1501 和 DukeMTMC 上的平均精度均值分别提高 3.1% 和 1.6%。这表明，使用 DINO 的细粒度表征补充 CLIP特征尤为重要。然而，直接拼接两者的特征会导致性能下降。融合Transformer 使 Market-1501 和 DukeMTMC 上的平均精度均值分别提升 1.6% 和 1.2%。结果表明，融合模块能够减小 DINO 和 CLIP 提取的特征之间的差异，同时保留各自的优势。

正则化器消融研究。 为研究模型间正则项与模型内正则化器的影响，我们在 Market-1501 和 DukeMTMC 上进行了消融实验，如表 4所示。此处，模型内 表示模型内正则化器，且 模型间 表示模型间正则项。有趣的是，在 Market-1501 上，模型间正则项对性能的贡献更大，使平均精度均值提升了 1.4%；而在 DukeMTMC 上，模型内正则化器更为有效，使平均精度均值提高了 1.7%。同时使用两种正则化器在两个数据集上均取得最优结果，在 Market-1501 和 DukeMTMC 上，平均精度均值/Rank-1 分别提升了 1.8%/0.6% 和 2.7%/1.4%。这证明了这两种正则化器的有效性。模型内正则化器增强了可学习令牌所捕获特征的多样性，而模型间正则项使 DINO 与 CLIP 之间的贡献更加均衡。

微调模型消融研究。 为了将 DINO（Caron 等人，2021年）和 CLIP（Li 等人，2023b）的单独微调与我们的方法进行比较，我们在 Market-1501 和 MSMT17 上进行了实验。如表 5 所示，DINO 在 Market-1501 上表现更好，而 CLIP 在 MSMT17 上取得了更优异的结果。相比之下，我们的方法在两个数据集上均表现出色。在 Market-1501 上，我们的方法比 DINO 高出 2.3%。在大规模 MSMT17 数据集上，我们的方法显著优于这两个微调模型，性能分别比 DINO 和 CLIP 高出 14.5% 和 12.6%。实验结果证明了我们的方法在捕捉并融合 DINO 和 CLIP 的互补特征方面的有效性。

表5. 在 Market-1501 和 MSMT17 数据集上对不同微调模型进行的实验。

方法	Market-1501		MSMT17	
	mAP	R-1	mAP	R-1
DINO 微调	90.3	95.4	64.2	83.4
CLIP 微调	86.4	93.3	66.1	84.4
DRFormer	**92.9**	**97.0**	**78.7**	**91.3**

表 6. 不同交互模块对 Market-1501 的影响。Cross 和 self 分别表示交叉注意力和自注意力。

融合Transformer	mAP	R-1
不使用融合模块	89.5	94.7
+1 交叉层	91.1	96.1
+1 交叉层和 1 个自注意力层	92.0	96.8
+1 交叉层和 2 个自注意力层	**92.9**	**97.0**

4.4. 超参数分析

双向 Transformer 的不同结构。 表6 总结了融合模块不同设计的实验结果。如表6 所示，引入单个交叉注意力层可显著提升性能。增加自注意力层后，性能能够稳定提升。结果表明，执行交叉注意力操作能够促进 DINO 和 CLIP 原始表征的融合，从而提升性能。

不同数量的可学习令牌。 如图 4所示，不同数据集对可学习令牌数量的敏感度各不相同。通常，仅使用一个可学习令牌可能会限制模型的表征能力，从而导致次优性能。在 Market-1501 和 DukeMTMC 上，分别使用 2 个和 3 个可学习令牌时取得了最佳性能。上述结果表明，我们应考虑适当的词元数量。设置过多词元并不一定更好。

5. 定性分析

DRFormer 的定性结果。 图 1（c）可视化了我们方法的注意力图，展示了其对人物整体身体的注意力。而图 1 （f）表明，我们的方法增大了困难样本之间的特征距离。这些结果表明，我们的方法有效融合了 DINO 和 CLIP 的能力。

注意力图可视化。图5 展示了我们的模型在 DukeMTMC 和 Occluded- Duke 上的注意力图。具体而言，第一行展示了包含单个行人的图像，第二行展示了包含多个人物或复杂背景的图像。如图5所示，我们的模型能够有效聚焦于人体

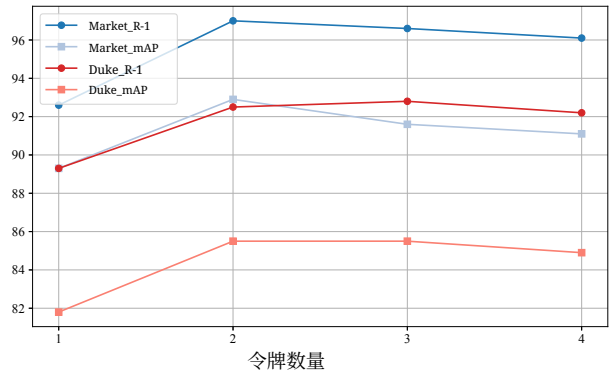


图4. 在 Market-1501 和 DukeMTMC 数据集上，不同数量可学习令牌的实验结果可视化。

并准确定位目标行人。我们的模型生成的注意力图能够清晰地界定遮挡区域与其他人物。

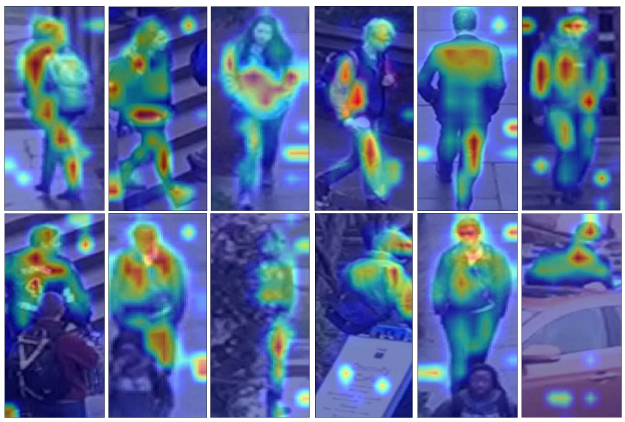


图 5. DukeMTMC 和 Occluded-Duke 数据集上的注意力权重可视化。

6. 结论

本文分析并验证了 DINO 和 CLIP 在行人再识别中的互补作用。DINO 能有效捕捉细粒度细节，而 CLIP 则提供全局语义。为发挥二者的优势，我们提出了一种双向 Transformer 模块，并引入两种正则化器：一种模型内正则化器，用于鼓励每个模型内部进行多样化的特征提取；另一种模型间正则化器，则在理论指导下平衡 DINO 和 CLIP 的贡献。大量定量和定性实验验证了 DRFormer 的有效性。

影响声明

本文旨在推动机器学习领域的发展。我们的工作可能带来诸多社会影响，但我们认为没有哪一项需要在此特别强调。

参考文献

Cao, Y., Xu, J., Lin, S., Wei, F., and Hu, H. GCNet：非局部网络与挤压激励网络及其扩展相结合。发表于*2019 IEEE/CVF*国际计算机视觉研讨会，*ICCV*研讨会*2019*，韩国首尔，*2019*年*10*月*27—28*日，第1971–1980页。IEEE，2019年。

Caron, M., Touvron, H., Misra, I., J´egou, H., Mairal, J., Bojanowski, P., and Joulin, A. 自监督视觉 Transformer 的新兴特性。发表于*2021*年 *IEEE/CVF*国际计算机视觉会议，*ICCV 2021*，蒙特利尔（加拿大魁北克省），*2021*年*10*月*10—17*日，第9630–9640页。IEEE，2021年。

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. 一张图像胜过16×16个词：用于大规模图像识别的 Transformer。发表于第*9*届国际学习表征会议，*ICLR 2021*，线上会议，奥地利，*2021*年*5*月*3—7*日。OpenReview.net，2021年。

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., 和 Lu, H. 用于场景分割的双重注意力网络。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，第3146–3154页。计算机视觉基金会 / IEEE，2019年。

He, K., Chen, X., Xie, S., Li, Y., Doll´ar, P., 和 Girshick, R. B. 掩码自编码器是可扩展的视觉学习器。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，第15979–15988页。IEEE，2022年。He, L., Liao, X., Liu, W., Liu, X., Cheng, P., 和 Mei, T. Fastreid：用于通用实例重识别的 PyTorch 工具箱。载于 El-Saddik, A., Mei, T., Cucchiara, R., Bertini, M., Vallejo, D. P. T., Atrey, P. K., 和 Hossain, M. S.（编），第*31*届 *ACM* 国际多媒体会议论文集，*MM 2023*，加拿大安大略省渥太华，*2023*年*10*月*29*日至*11*月*3*日，第9664–9667页。ACM，2023年。

He, S., Luo, H., Wang, P., Wang, F., Li, H., and Jiang, W. Transreid：基于Transformer的目标重识别。载于*2021*年*IEEE/CVF*国际计算机视觉会议，*ICCV 2021*，蒙特利尔（加拿大魁北克省），*2021*年*10*月*10*日至*17*日，第14993–15002页。IEEE，2021年。

Hu, B., Wang, X., and Liu, W. Personvit：用于行人重识别的大规模自监督视觉变换器。《机器视觉与应用》，36(2):32，2025。Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. 利用噪声文本监督扩展视觉与视觉语言表征学习。在 Meila, M. and Zhang, T.（编）所编的《第*38*届机器学习国际会议论文集，*ICML 2021*，*2021*年*7*月*18*日至*24*日，线上会议》中，第139卷《style i d='5'>《机器学习研究论文集》，第4904–4916页。PMLR，2021年。

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W., Doll´ar, P., 和 Girshick, R. B. 万物分割。载于*IEEE/CVF*国际计算机视觉会议，*ICCV 2023*，巴黎，法国，*2023*年*10*月*1*日至*6*日，第3992–4003页。IEEE，2023。

Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L. M., 和 Shum, H. Mask DINO：面向目标检测和分割的统一 Transformer 框架。载于 *IEEE/CVF*计算机视觉与模式识别会议，*CVPR 2023*，温哥华，*BC*省，加拿大，*2023*年*6*月*17*日至*24*日，第3041–3050页。IEEE，2023a。

Li, S., Sun, L., and Li, Q. Clip-reid：利用视觉-语言模型进行图像重识别，无需具体文本标签。见 Williams, B., Chen, Y., and Neville, J.（编），第*37*届*AAAI*人工智能会议，*AAAI 2023*，第*35*届人工智能创新应用会议，*IAAI 2023*，第*13*届人工智能教育进展研讨会，*EAAI 2023*，美国华盛顿特区，*2023*年*2*月*7*日至*14*日，第1405–1413页。AAAI出版社，2023b。

Li, W., Zhao, R., Xiao, T., and Wang, X. Deepreid：用于行人重识别的深度滤波器配对神经网络。见*2014*年*IEEE*计算机视觉与模式识别会议，*CVPR 2014*，美国俄亥俄州哥伦布，*2014*年*6*月*23*日至*28*日，第152–159页。IEEE计算机学会，2014。

Li, W., Zou, C., Wang, M., Xu, F., Zhao, J., Zheng, R., Cheng, Y., and Chu, W. DC-Former：用于行人重识别的多样紧凑型Transformer。载于 Williams, B., Chen, Y., and Neville, J.（编），第*37*届*AAAI*人工智能大会（*AAAI 2023*）、第*35*届创新人工智能应用大会（*IAAI 2023*）、第*13*届人工智能教育进展研讨会（*EAAI 2023*），美国华盛顿特区，*2023*年*2*月*7*日至*14*日，第1415–1423页。AAAI出版社，2023c。

Liu, F.、Kim, M.、Ren, Z. 和 Liu, X. 通过双重引导蒸馏 CLIP，以学习判别性人体

体型表示。在 *IEEE/CVF*计算机视觉与模式识别会议，*CVPR 2024*，美国华盛顿州西雅图，*2024*年*6*月*16*日至*22*日，第256–266页。IEEE，2024年。

Luo, H., Wang, P., Xu, Y., Ding, F., Zhou, Y., Wang, F., Li, H. 和 Jin, R. 基于自监督预训练的Transformer行人重识别。*CoRR*，abs/2111.12084，2021年。

Miao, J., Wu, Y., Liu, P., Ding, Y. 和 Yang, Y. 姿态引导的特征对齐，用于遮挡行人重识别。载于*2019*年*IEEE/CVF*国际计算机视觉会议，*ICCV 2019*，首尔，韩国，*2019*年*10*月*27*日至*11*月*2*日，第542–551页。IEEE，2019年。

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., J´egou, H., Mairal, J., Labatut, P., Joulin, A. 和 Bojanowski, P. DINOv2：无需监督即可学习鲁棒的视觉特征。机器学习研究汇刊，2024，2024。

Radford, A.、Kim, J. W.、Hallacy, C.、Ramesh, A.、Goh, G.、Agarwal, S.、Sastry, G.、Askell, A.、Mishkin, P.、Clark, J.、Krueger, G. 和 Sutskever, I. 基于自然语言监督学习可迁移视觉模型。*CoRR*，abs/2103.00020，2021年。

Ristani, E.、Solera, F.、Zou, R. S.、Cucchiara, R. 和 Tomasi, C. 多目标多相机跟踪的性能度量与数据集。载于 Hua, G. 和 J´egou, H.（编），《计算机视觉——*ECCV 2016* 阿姆斯特丹研讨会——荷兰阿姆斯特丹，*2016*年*10*月*8*日至*10*日及*15*日至*16*日，论文集，第二部分》，《计算机科学讲义》第9914卷，第17–35页，2016年。

Shahhosseini, M., Hu, G. 和 Pham, H. 针对回归问题优化机器学习模型的集成权重和超参数。*CoRR*，abs/1908.05287，2019。

Sun, Y., Zheng, L., Yang, Y., Tian, Q. 和 Wang, S. 超越部件模型：使用精细化部件池化进行行人检索（以及一个强大的卷积基线）。载于 Ferrari, V., Hebert, M., Sminchisescu, C. 和 Weiss, Y.（编），计算机视觉——*ECCV 2018*——第*15*届欧洲计算机视觉会议，慕尼黑，德国，*2018*年*9*月*8*日至*14*日，论文集，第四部分，《计算机科学讲义》第11208卷，第501–518页。施普林格，2018年。

Sun, Z., Tan, L., Shen, Y., Cai, C., Sun, X., Dai, P., Cao, L., and Ji, R. FlexiReID：用于多模态行人重识别的自适应专家混合。在第四十二届国际机器学习会议（*ICML*）*2025*年，温哥华，*BC*省，加拿大，*2025*年*7*月*13*日至*19*日。Open-Review.net，2025。Wang, J., Chen, K., Li, Y., Chen, B., Zhao, H., Qi, X., and Tian, Z. 用于增强开放词汇密集感知的广义解耦学习。*CoRR*，abs/2508.11256，2025。

Wei, L., Zhang, S., Gao, W. 和 Tian, Q. 使用行人迁移生成对抗网络弥合行人重识别中的领域差异。载于*2018*年*IEEE*计算机视觉与模式识别会议，*CVPR 2018*，盐湖城，犹他州，美国，*2018*年*6*月*18*日至*22*日，第79–88页。计算机视觉基金会 / IEEE计算机学会，2018年。

Wei, S., Luo, C. 和 Luo, Y. 通过不平衡学习改进多模态学习。*CoRR*，abs/2507.10203，2025。

Xu, J., Zhao, R., Zhu, F., Wang, H. 和 Ouyang, W. 用于行人再识别的注意力感知组合网络。载于 *2018*年*IEEE*计算机视觉与模式识别会议，*CVPR 2018*，盐湖城，犹他州，美国，*2018*年*6*月*18*日至*22*日，第 2119–2128 页。计算机视觉基金会 / IEEE计算机学会，2018。

Yang, Z., Wu, D., Wu, C., Lin, Z., Gu, J. 和 Wang, W. 一个行人抵得上一个提示：迈向语言引导的行人再识别。载于 *IEEE/CVF*计算机视觉与模式识别会议，*CVPR 2024*，西雅图，华盛顿州，美国，*2024*年*6*月*16*日至*22*日，第 17343–17353 页。IEEE，2024。

Zhai, X.、Mustafa, B.、Kolesnikov, A.和Beyer, L. 面向语言图像预训练的Sigmoid损失。载于*IEEE/CVF*国际计算机视觉会议，*ICCV 2023*，巴黎，法国，*2023*年*10*月*1*日至*6*日，第11941–11952页。IEEE，2023。

Zhang, G.、Zhang, Y.、Zhang, T.、Li, B.和Pu, S. PHA：面向基于Transformer的行人重识别的分块式高频增强。载于*IEEE/CVF*计算机视觉与模式识别会议，*CVPR 2023*，温哥华，*BC*，加拿大，*2023*年*6*月*17*日至*24*日，第14133–14142页。IEEE，2023。

Zhang, W.、Liu, H.、Li, B.、He, J.、Qi, Z.、Wang, Y.、Zhao, S.、Yu, X.、Zeng, W.和Jin, X. 面向自监督单目深度估计的混合粒度特征聚合与由粗到细的语言引导。*CoRR*，abs/2510.09320，2025。

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J.与Tian, Q. 可扩展的行人重识别：一项基准。载于*2015*年 *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，第1116–1124页。IEEE计算机学会，2015。

Zhou, L., Li, S., Dong, N., Tai, Y., Zhang, Y.与Li, H. 面向基于图像和文本的行人重识别的层次化提示学习，2025。Zhou, Q., Zhong, B., Liu, X.与Ji, R. 面向行人重识别的基于注意力的神经架构搜索。*IEEE*神经网络与学习系统汇刊，33(11):6627–6639，2022。Zhu, H., Ke, W., Li, D., Liu, J., Tian, L.与Shan, Y. 面向细粒度视觉分类与目标重识别的双重交叉注意力学习。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，第4682–4692页。IEEE，2022a。Zhu, K., Guo, H., Yan, T., Zhu, Y., Wang, J.与Tang, M. PASS：面向行人重识别的部件感知自监督预训练。载于Avidan, S., Brostow, G. J., Ciss´e, M., Farinella, G. M.与Hassner, T.（编），计算机视觉 - *ECCV 2022* - 第*17*届欧洲会议，以色列特拉维夫，*2022*年*10*月*23*日至*27*日，会议论文集，第*XIV*部分，《计算机科学讲义》第13674卷，第198–214页。施普林格，2022b。Zhu, K., Guo, H., Zhang, S., Wang, Y., Liu, J., Wang, J.与Tang, M. Aaformer：面向行人重识别的自动对齐变换器。*IEEE*神经网络与学习系统汇刊，35(12):17307–17317，2024。

A. 理论推导

在第 3.4 节中，我们简要说 明了引入模型间正则项的动机。在此，我们提供公式的详细
公式的推导 乐彩 n.

双模型学习。 不失一般性，我们考虑两个编码器，分别为 $\varphi_0(\boldsymbol{\theta}_0, \cdot)$ 和 $\varphi_1(\boldsymbol{\theta}_1, \cdot)$，其中 θ_0 和θ_1是编码器的参数。数据集表示为 $\mathcal{D} = \{x_i\}_{i=1,2,...,N}$，其中 $y \in 1, 2, ..., M$， M是类别数量。编码器的表示输出表示为 $\boldsymbol{f}_0 = \phi_0(\boldsymbol{\theta}_0, x_i)$ 和 $\boldsymbol{f}_1 = \phi_1(\boldsymbol{\theta}_1, x_i)$。两个编码器通过某种融合方法，借助表征连接起来。这里，令 $\phi_f(\boldsymbol{\theta}_{\boldsymbol{f}}, \cdot)$ 表示融合模块。$\boldsymbol{\theta}_f$ 是该模块的参数。值得注意的是，我们的方法采用双向交叉注意力进行融合。因此，融合模块输出两个增强的特征表示 $\boldsymbol{z}_0$ 和 $\boldsymbol{z}_1$。

$$\boldsymbol{z}_0 = \phi_f(\boldsymbol{\theta}_{\boldsymbol{f}}, \boldsymbol{f}_0) \tag{27}$$
$$\boldsymbol{z}_1 = \phi_f(\boldsymbol{\theta}_{\boldsymbol{f}}, \boldsymbol{f}_1) \tag{28}$$

令 $\boldsymbol{W} \in \mathbb{R}^{M \times (d_0 + d_1)}$ 和 $\boldsymbol{b} \in \mathbb{R}^M$ 表示用于产生 logits 输出的线性分类器参数。双模型中输入 x_i 的输出可表示如下:

$$f(x_i) = \boldsymbol{W}\boldsymbol{z}_f + \boldsymbol{b} \tag{29}$$
$$\boldsymbol{z}_f = [\boldsymbol{z}_0; \boldsymbol{z}_1] \tag{30}$$

令 $\boldsymbol{s}_i^0 = \boldsymbol{W}_0 \cdot \boldsymbol{z}_0 + \boldsymbol{b}_0$表示模型 $\varphi_0(\boldsymbol{\theta}_0, \cdot)$ 的逻辑值输出， $\boldsymbol{s}_i^1 = \boldsymbol{W}_1 \cdot \boldsymbol{z}_1 + \boldsymbol{b}_1$表示模型 $\varphi_1(\boldsymbol{\theta}_1, \cdot)$ 的逻辑值输出。因此，最终输出是 $\boldsymbol{s}_i^0$与 $\boldsymbol{s}_i^1$之和。从而，梯度由 $\boldsymbol{s}_i^0$和 $\boldsymbol{s}_i^1$决定。

不同模型的最优贡献比例。我们需要找出不同模型的最优贡献比例。一般而言，最优贡献比例应使泛化误差最小化，因此我们将问题表述如下，

$$\min_{w_0, w_1} g = \frac{1}{N} \sum_{i=1}^{i=N} \mathcal{L}(f(x_i), y_i) \tag{31}$$
$$f(x_i) = w_0 s_i^0 + w_1 s_i^1 \tag{32}$$

其中 $w_0 > 0$ 和 $w1 > 0$ 分别表示两个模型的贡献，且 $w_0 + w_1 = 1$。 $\mathcal{L}(.)$ 衡量预测与真实值之间的泛化误差。为简化起见，我们定义 $E[f(x)] = \frac{1}{N} \sum_{i=1}^{i=N} f(x_i)$。根据偏差-方差分解（Shahhosseini 等人，2019），预测与真实值之间的误差可改写如下。这里我们主要关注平方项。

$$\begin{cases} g = (Bias(f(x), y))^2 + Var(f(x)) + Var(\epsilon) \\ Bias(f(x_i), y) = \mathbb{E}[f(x) - y] \end{cases} \tag{33}$$

然后，在约束 $w_0 + w_1 = 1$下，我们得到 w_0 和 w_1 的解如下，

$$w_0 = \frac{Bias(s^0, y)}{Bias(s^0, y) - Bias(s^1, y)} \tag{34}$$
$$w_1 = \frac{-Bias(s^1, y)}{Bias(s^0, y) - Bias(s^1, y)} \tag{35}$$

此处，数值解毫无意义，因为 w_0 或 w_1 必有一个小于 0，这与 $w_0 > 0$ 和$w_1 > 0$相矛盾。因此，当固定 $Bias(s^0, y)$和 $Bias(s^1, y)$时，我们无法找到合理的 $w0$ 或 $w1$ 组合来最小化 $Bias(。)^2$。因此，最小化 $Bias(。)^2$ 的唯一方法是最小化 $Bias(s^0, y)$和 $Bias(s^1, y)$。

B. 分类结果可视化

困难行人样本判别。图6 展示了若干身份不同、但外观和身体姿势高度相似的行人图像。红色虚线标出了不同图像中行人细粒度细节的差异。我们的模型成功区分了这些具有挑战性的样本，展现了其强大的图像理解能力。

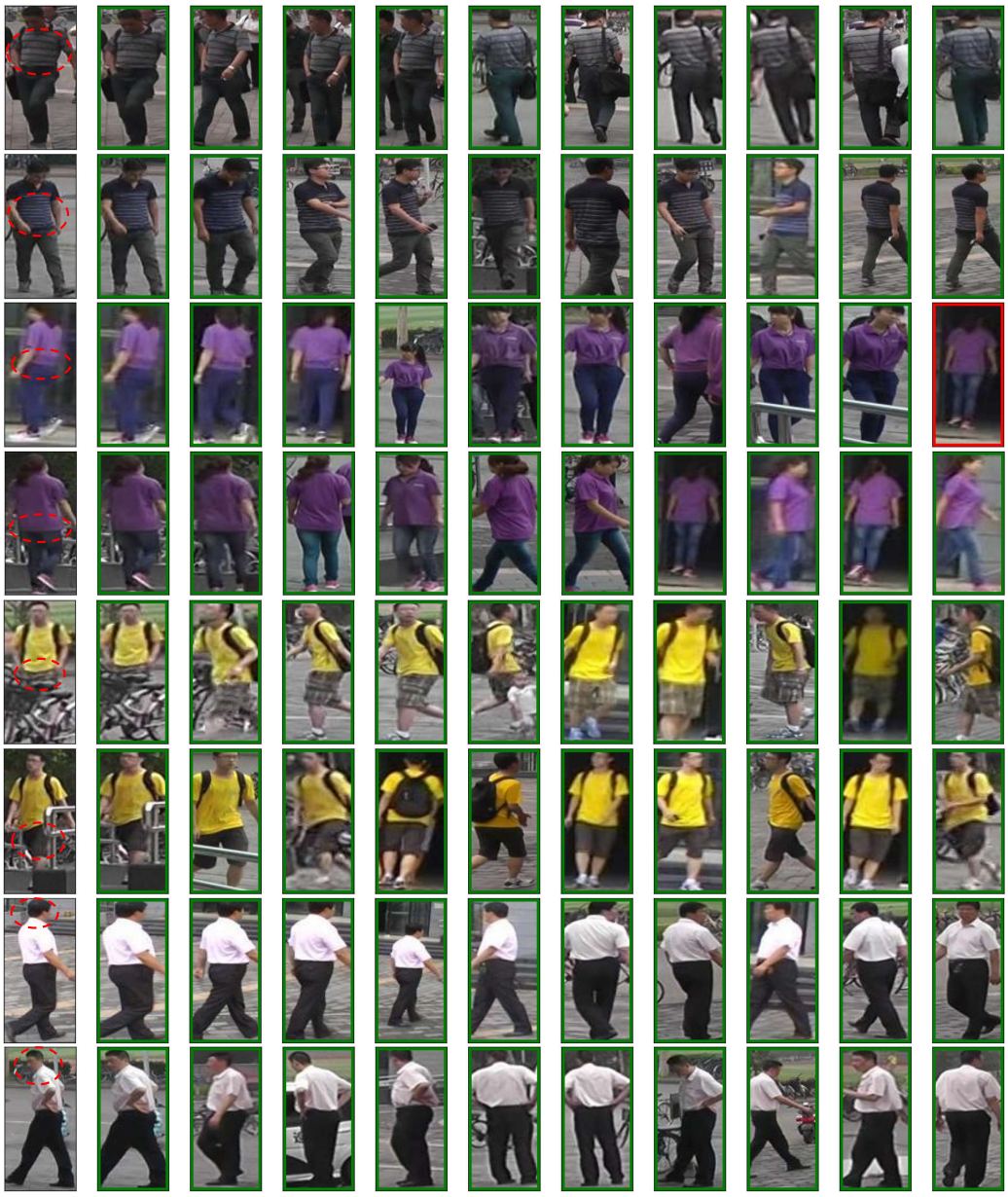


图*6.* Market-1501 上高度相似样本的分类结果。

C. 模型内正则化器消融研究。

模型内正则化器组件的消融实验。 模型内正则化器首先应用 SIE，然后最大化词元之间的余弦距离。我们在 DukeMTMC 和 Occluded-

*Ta*表 *7.* DukeMTMC 和 Occluded-Duke 数据集上模型内正则化器各组件的消融实验 ts.

方法	DukeMTMC		Occluded-Duke	
	mAP	R-1	mAP	R-1
不使用模型内正则化器	83.0	91.0	64.0	72.4
+SIE	84.2	91.5	64.3	72.1
+余弦损失	85.5	92.5	65.3	72.3

Duke 数据集上进行了消融研究。如表 7所示，SIE 分别使 DukeMTMC 和 Occluded-Duke 上的平均精度均值提高了 1.2% 和 0.3%。余弦损失进一步使性能提升了 1.3% 和 1.0%。总体而言，余弦损失对性能的影响更大。实验结果验证了模型内正则化器的有效性。