arXiv:2601.00307v1 [cs.CV] 1 Jan 2026

VisNet：通过α-散度损失、特征融合和动态多任务学习实现高效行人重识别

Anns Ijaz人工智能系管理与技术大学拉合尔，巴基斯坦
annsijaz@outlook.com

Muhammad Azeem Javed博士人工智能系管理与技术大学拉合尔，巴基斯坦 azee
m.javed@umt.edu.pk

摘要—行人重识别（ReID）是监控和移动应用领域中极其重要的方向之一，需要在尽可能低的计算成本下实现较高的准确率。当前最先进的方法能够取得良好的准确率，但需要较高的计算预算。为此，本文提出了VisNet，这是一种计算高效且有效的重识别模型，适用于真实世界场景。该模型融合了多项概念性创新，包括在多尺度上进行特征融合，并对每个尺度应用自动注意力；结合人体解剖分区的语义聚类；用于平衡分类语义正则化的动态权重平均技术；以及使用FIDI损失函数以改进度量学习任务。多尺度特征融合了ResNet50的第1至第4阶段，且未采用并行路径；语义聚类则通过基于规则的伪标签生成引入空间约束。在Market-1501数据集上，VisNet取得了87.05%的Rank-1准确率和77.65%的平均精度均值（mAP），参数量为32.41M，计算量为4.601 GFLOPs。因此，本文提出了一种实用的方法，可用于计算资源有限的监控和移动应用中的实时部署。

I. 引言

智能视频分析的兴起推动了大规模监控网络的发展，使其能够应用于零售分析、法证调查和公共安全等领域。这些系统的核心是行人再识别（re-ID），即在互不重叠的摄像机视角中匹配同一个人的外观。尽管这对可扩展分析至关重要，但由于不同视角下存在显著的外观变化，实现稳健的再识别仍然困难重重。身份的视觉特征常因摄像机角度和人体姿态的大幅变化（视点变化）而受到影响，也可能因拥挤场景遮挡了具有区分性的线索（遮挡）而难以辨认。此外，光照变化造成的光度不一致、摄像机距离不同导致的尺度变化，以及时间间隔，都会显著改变外观。传统的手工设计特征（例如颜色直方图和基于边缘的描述符） [1], 无法捕捉克服此类变化所需的丰富语义，因此需要采用更稳健的特征学习方法。

随着深度学习的兴起，空间划分策略和基于卷积神经网络的方法带来了显著的改进。方法如 PCB 和 AANet 通过划分特征图或引入自适应注意力来捕捉局部语义结构，实现了 76–92% 的 Rank-1 准确率 [2] [3]。然而，这些方法本质上仍受限于对预定义空间区域的依赖。因此，近期研究已转向 Transformer 架构 [4] 和视觉语言模型，以捕捉全局语义上下文。TransReID [5] 和 CLIP-ReID [6] 等方案报告了更高的准确率（Rank-1 准确率超过 88%），但计算成本高得难以承受。例如，TransReID 约需 17.8 GFLOPs 和 8600 万个参数，因此无法部署在边缘设备上。这清楚地表明了效率与准确率之间的权衡：轻量级卷积神经网络方法计算效率高，但缺乏 Transformer 的语义推理能力；而基于 Transformer 的模型虽然提升了准确率，却付出了高昂的计算成本，因而无法部署在资源受限环境中。



图 1：Market-1501 查询：在 19,733 张图像中，所提出的模型成功地在排名前 5 的结果中跨多个视点重新识别出该行人（ID: 0921）。

为了解决这一权衡，需要一种方法，既保留卷积神经网络的空间效率，又能模拟 Transformer 的语义捕获能力。尽管多尺度特征学习在计算机视觉中已广为人知，空间金字塔池化和特征金字塔网络 [7]可提升目标检测性能，但在行人重识别领域仍未得到充分应用。此外，尽管 SE-Net [8] 和 CBAM [9] 等注意力机制作用于通道或空间维度，尺度维度仍是一个尚未充分探索的设计空间，涉及如何确定对于给定图像而言，哪种多尺度特征表示最具信息量。

在这项工作中，我们提出了 VisNet，这是一种经过系统设计的

行人重识别模型，在保持实际计算效率的同时实现了具有竞争力的准确率。VisNet 没有采用计算开销巨大的 Transformer 架构，而是策略性地结合了经过验证的卷积神经网络技术。它从多个 ResNet50 残差块中提取特征 [10]并将其映射到统一维度。随后，通过一种新颖的定制尺度注意力机制将这些特征融合。该轻量级模块能够学习自适应的逐尺度权重，从而捕捉不同抽象层级的语义信息。为了在不采用 SOLIDER 等开销高昂的教师-学生框架的情况下进一步规范语义学习[11], 我们引入了结合基于规则的伪标签的空间语义聚类（将人体划分为上半身、下半身和鞋子）。最后，为确保度量学习的鲁棒性，我们采用 FIDI损失、语义损失和交叉熵损失，并通过动态权重平均 [12], 平衡身份分类、度量学习和语义正则化任务的收敛速度。

所提出的方法在 Market-1501 基准上进行了验证 [13]，，仅需 4.601G 浮点运算次数和 31.08M 个参数（为 TransReID 的 0.36×），即可达到 87.05% 的 Rank-1 准确率和 77.65% 的平均精度均值。本文的主要贡献如下：

- 一种轻量且准确的基于卷积神经网络的行人重识别模型，采用定制的尺度注意力机制，学习各尺度的自适应权重，以实现多尺度特征融合。
- 展示了一种方法：基于规则的空间伪标签能够有效地规范语义学习，无需开销巨大的教师-学生框架，并可作为近期复杂方法的一种有竞争力且更简单的替代方案。
- 通过广泛的实证验证，量化各组件的贡献，分析效率-准确率权衡，并对结果进行定性分析。
- 一种语义感知的数据增强框架，通过明确地将前景身份与环境杂乱因素解耦，确保背景不变性。

本文结构安排如下：第二节介绍所提出的 VisNet 架构。第三节描述评估性能，结论与未来工作分别见第四节和第五节。

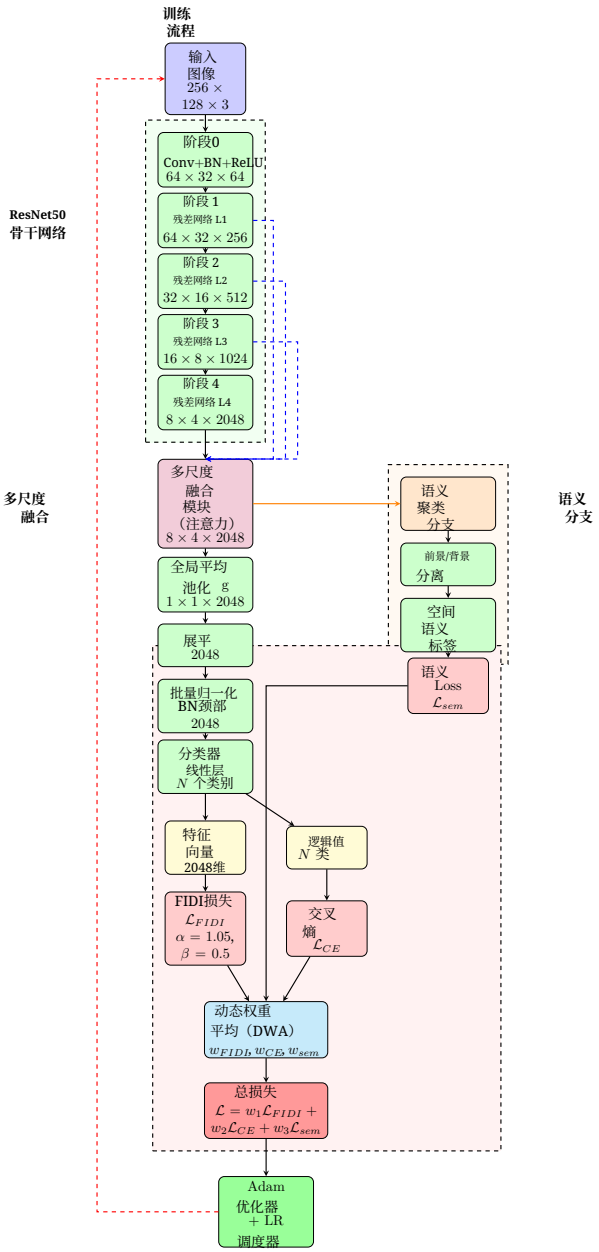


图 2：所提方法的训练流程。

II. 提出的方法

VisNet集成了ResNet50骨干网络、结合学习到的注意力的多尺度特征融合、空间语义聚类，以及
一种动态多任务损失函数。它将输入256 × 128 图像依次通过 ResNet50 骨干网络的五个阶段。在不同的语义层级和空间分辨率下，从阶段 1 至阶段 4 提取多尺度特征。这些特征通过学习到的注意力加权组合进行融合。随后，这些融合特征图被输入两个并行分支，即用于行人重识别的身份分类分支和用于空间正则化的语义聚类头。在推理时，只有身份分类分支参与最终的特征嵌入。在训练过程中，二者都通过采用动态权重平均的多任务学习框架参与总损失的计算。

A. 主干网络架构与特征提取

VisNet采用ResNet50作为骨干网络。该网络分为五个阶段：

具体而言，它由以下部分组成：阶段0：首先是一个步幅为 2 的 7×7卷积干网络，随后进行批归一化和 ReLU [14]激活，接着进行最大池化。此阶段输出 64 个特征通道，步幅为 4。

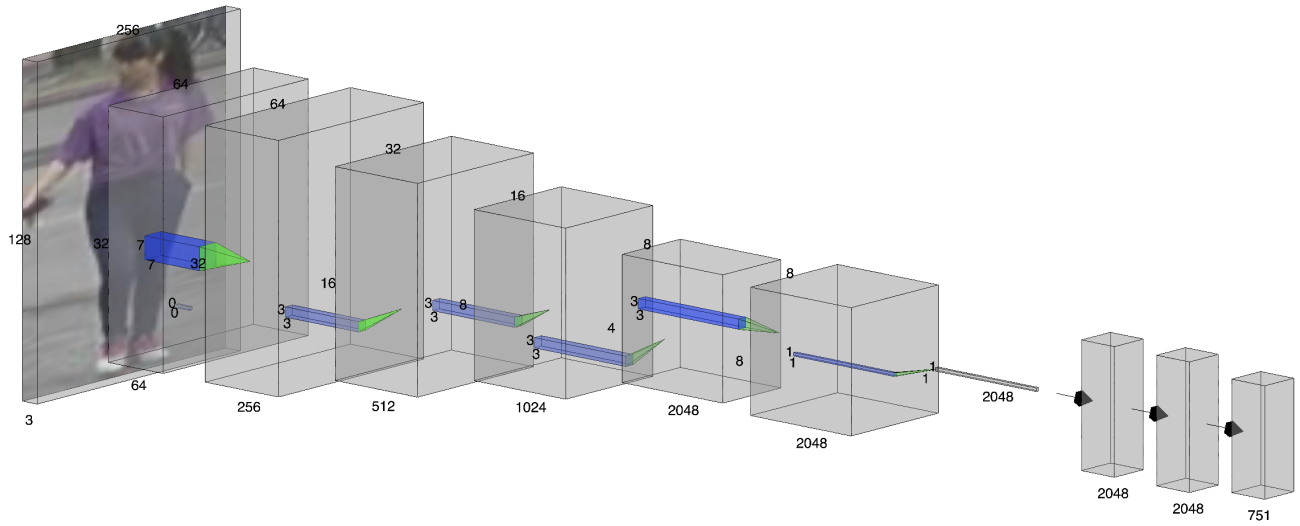


图 3：所提出的 VisNet 架构

• **阶段 1–4：** 与 ResNet50 的 layer1 至 layer4 对应的残差块，逐步提取语义抽象层级不断提升的特征。

各阶段的输出特征维度如下：

- 阶段 1（layer1）：256 个通道，步幅为 4
- 阶段 2（layer2）：512 个通道，步幅为 8
- 阶段 3（layer3）：1024 个通道，步幅为 16
- 阶段 4（layer4）：2048 个通道，步幅为 32

提取全部四个阶段的特征用于多尺度融合，在计算效率与有效捕获细粒度信息和语义级信息的需求之间取得平衡。

B. 采用学习式尺度注意力的多尺度特征融合

VisNet 的一项核心贡献是通过学习得到的逐尺度注意力权重，对多尺度特征进行系统融合；这不同于简单的拼接或相加，能够确定每个 ResNet 阶段的特征对于给定输入图像的相对信息量。

1) 投影至统一维度：阶段 1 至阶段 4 的通道维度各不相同：256、512、1024、2048。所有特征均通过 1×1 卷积投影至 2048 个通道的共同维度，随后进行批归一化和 ReLU 激活：

$$F_i' = \text{ReLU}(\text{BN}(\text{Conv}_{1\times1}(F_i))) \tag{1}$$

其中，F_i 是阶段 i的特征图，F_i' 是其投影表示。训练期间采用标签平滑（$\epsilon = 0.1$），以提高泛化能力。

2) 空间对齐：随后对这些特征进行双线性上采样，使其达到阶段 4 的空间分辨率：步幅为 32。上采样后，所有特征具有相同的空间形状，表示为$[B, 2048, H, W]$，其中 B为批量大小，$H \times W$为该分辨率下的高度和宽度。

3) 尺度注意力加权：通过学习得到的逐尺度注意力权重来表示各尺度的重要性。对四个投影特征取平均，计算得到平均特征图：

$$\bar{F} = \frac{1}{4}\sum_{i=1}^{4} F_i' \tag{2}$$

¯轻量级注意力模块对 F 进行处理，该模块由全局平均池化和一个小型多层¯感知机（MLP）组成。具体而言，对 F 进行全局池化降维后，依次通过两个带有 ReLU 激活的全连接层，并在输出端使用 Sigmoid 激活，从而得到逐尺度权重：

$$w = \text{Sigmoid}(\text{FC}_{512}(\text{ReLU}(\text{FC}_{2048}(\text{GAP}(\bar{F}))))) \tag{3}$$

该模块输出四个标量权重，分别对应一个 ResNet阶段，并限制在范围 $[0, 1]$ 内。这些权重彼此独立，且不需要归一化至总和为 1。[1]

4）加权特征求和：将四个投影的输出进行加权求和，得到融合特征图：

$$F_{\text{fused}} = \sum_{i=1}^{4} w_i \cdot F_i' \tag{4}$$

[1]通过 1×1 卷积实现，在全局平均池化后，其功能等同于全连接层。

生成的特征图保留了形状 $[B, 2048, H, W]$，并体现了来自所有 ResNet 阶段的信息的自适应组合。

C. 基于规则的伪标签的语义聚类

通过生成基于规则的伪标签，为每个空间位置提供空间和语义上连贯的表示 [15]。无需使用教师-学生蒸馏方法，即可提供正则化信号。

1）空间划分：基于人体的解剖学先验进行垂直划分。将垂直坐标为 $y \in [0, 1, 2]$ 的每个空间位置分配至一个语义类别：

$$\text{spatial_class}(y) = \begin{cases} 0 & \text{if } y < 0.4 \text{ (upper body)} \\ 1 & \text{if } 0.4 \leq y < 0.8 \text{ (lower body)} \\ 2 & \text{if } y \geq 0.8 \text{ (shoes)} \end{cases} \quad (5)$$

该划分得到三个感兴趣区域，分别对应上半身（顶部 40% ）、下半身（中间 40% ）和鞋类（底部 20% ）。

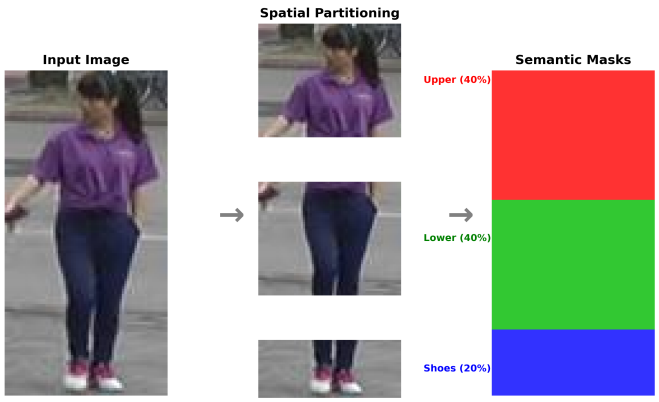


图 4：空间聚类

2）前景-背景分离：通过计算每个空间位置处融合特征的 L2 范数的大小，将前景（即人体）与背景分离。随后对这些大小求平均值，并计算其在所有位置上的标准差：

$$\text{magnitude}(y, x) = \|F_{\text{fused}}(y, x)\|_2 \quad (6)$$

$$\mu = \text{mean}(\text{magnitude}), \quad \sigma = \text{std}(\text{magnitude}) \quad (7)$$

如果满足以下条件，则将该位置分类为前景：

$$\text{magnitude}(y, x) > \mu + 0.5\sigma \quad (8)$$

3) 伪标签生成：每个位置的最终伪标签包含空间类别以及前景-背景信息：

$$\text{pseudo_label}(y, x) = \begin{cases} \text{spatial_class}(y) & \text{if foreground} \\ 3 & \text{if background} \end{cases} \quad (9)$$

因此，可以定义四个语义类别：上半身、下半身、鞋子和背景。

4）语义分类头：逐像素语义分类头以融合特征图作为输入。特征会经过重塑，使每个空间位置都表示一个独立样本，然后由一个紧凑的神经网络预测每个位置所属的语义类别。

我们将语义头设计为一个三层多层感知机：输入维度为 2048，隐藏层分别包含 1024 和 512 个神经元，输出层包含 4 个神经元。层间采用批归一化和 ReLU 激活，并以 0.1 的比率应用丢弃法进行正则化。语义逻辑值定义如下：

$$\text{semantic_logits} = \text{MLP}(F_{\text{fused_flat}}) \quad (10)$$

语义分类损失为预测逻辑值与伪标签之间的交叉熵：

$$L_{\text{semantic}} = \text{CrossEntropy}(\text{semantic_logits}, \text{pseudo_labels}) \quad (11)$$

D. 身份分类分支

与此同时，融合特征图用于身份分类。全局平均池化将空间维度缩减为一个大小为 2048 的 1×1 特征向量：

$$\mathbf{f} = \text{GAP}(F_{\text{fused}}) \quad (12)$$

接下来是一个不带偏置项的批归一化层，这在行人重识别中很常见：

$$\mathbf{f}_{\text{bn}} = \text{BN}(\mathbf{f}) \quad (13)$$

线性分类器将$\mathbf{f}_{\text{bn}}$ 映射到训练集中的身份数量；对于 Market-1501，该数量为751：

$$\text{logits}_{\text{id}} = \text{Linear}(\mathbf{f}_{\text{bn}}) \quad (14)$$

推理时，使用批归一化特征作为行人重识别的嵌入，同时通过单位归一化和欧氏距离处理查询-图库匹配。

E. 损失函数与多任务学习

VisNet 通过三种互补损失进行动态平衡训练。接下来将详细介绍这些损失。

1）身份分类损失：身份分类分支对身份预测使用标准交叉熵损失：

$$L_{\text{CE}} = \text{CrossEntropy}(\text{logits}_{\text{id}}, y_{\text{id}}) \tag{15}$$

其中 y_{id} 包含真实身份标签。

*2) FIDI*度量学习损失： FIDI为度量学习提供了一种替代传统三元组损失的方法。FIDI将度量学习表述为学习得到的分布 $\mathcal{U}$ 与基于真实标注的分布 $\mathcal{K}$之间的对称散度最小化。

FIDI损失基于相对熵构建，相对熵是衡量两个分布之间距离的指标。设 $\mathcal{K}$ 为训练图像对的已知分布，即真实身份标签；设 $\mathcal{U}$ 为我们旨在学习的未知分布，则FIDI损失定义如下 [16]:

$$L_{\text{FIDI}} = D(\mathcal{U}\|\mathcal{K}) + D(\mathcal{K}\|\mathcal{U}) \tag{16}$$

其中，α散度定义如下：

$$D(\mathcal{U}\|\mathcal{K}) = \sum_{p_{ij}\in\mathcal{P}} u_{p_{ij}} \log \frac{\alpha u_{p_{ij}}}{(\alpha-1)u_{p_{ij}} + k_{p_{ij}}} \tag{17}$$

此处， $p_{ij} = \{\mathbf{x}_i, \mathbf{x}_j\}$ 是一对图像样本， $\mathcal{P}$是图像对的集合。若图像对$\mathbf{x}_i$ 和$\mathbf{x}^j$ 属于同一身份，则 $k_{p_{ij}} \in \mathcal{K}$ 和 $k_{p_{ij}} = 1$ ；否则，$k_{p_{ij}} = 0$ 。$u_{p_{ij}}$ 取自未知分布$\mathcal{U}$，该分布表示 $\mathcal{P}$中图像对的特征层级关系分布。

3) 语义聚类损失：语义聚类损失对应于逐像素语义分类任务中的交叉熵损失：

$$L_{\text{semantic}} = \text{CrossEntropy}(\text{semantic_logits}, \text{pseudo_labels}) \tag{18}$$

4) 动态权重平均总损失：不同于固定权重，三个损失通过动态权重平均（DWA） [17]相互平衡：

$$L_{\text{total}} = w_{\text{FIDI}}(t)\cdot L_{\text{FIDI}} + w_{\text{CE}}(t)\cdot L_{\text{CE}} + w_{\text{semantic}}(t)\cdot L_{\text{semantic}} \tag{19}$$

其中，权重 w_i（t）根据近期损失历史在每个批次 t 计算得到。

F. 训练策略：动态权重平均与批次采样

DWA 根据各项损失的下降速率调整权重。对于每项任务，它会跟踪最近 50 个批次中的损失值。在批次 t 计算比率 $r_i(t) = \frac{L_i(t)}{L_i(t-1)+\epsilon}$。权重通过温度为 $T = 2.0$ 的 Softmax 归一化计算得出：

$$w_i(t) = \frac{\exp(r_i(t)/T)}{\sum_j \exp(r_j(t)/T)} \tag{20}$$

这会自动提高当前进展较慢的任务的相对影响力，避免进展最快的目标占据主导地位。温度决定了各批次间权重变化的平滑程度。

该方法还采用PK采样 [18]–[20]，每个批次包含 P 个身份，每个身份包含 K 张图像。更具体地说，我们使用 $P = 8$、 $K = 12$，最终批量大小为96个样本；这也有助于确保批次内的多样性，并支持困难负样本挖掘。三个损失从第0个epoch开始全部启用，这对应于单阶段训练。动态权重平均机制能够自动平衡各个目标，无需手动调参。在测试期间，从身份分类分支中提取经过L2归一化的特征。行人重识别被视为一个检索任务：计算查询图像与图库图像之间的距离，然后依据相似度进行排序。我们报告两个标准指标：累积匹配特征（CMC）和平均精度均值（mAP）。

G. 语义感知数据增强流程

为提高模型对背景变化的鲁棒性，我们实现了一种基于语义分割的增强流程。该模块将前景主体与背景解耦，从而能够针对背景进行扰动，同时保留人物身份特征。

该流程利用 YOLOv8-Seg 进行实时实例分割，以生成二值掩码 $M \in \{0,1\}^{H\times W}$，从而分离人物像素 $I_{person} = I \odot M$。随后，我们仅对背景区域 $I_{bg} = I \odot$ 应用随机变换函数 $\mathcal{T}(\cdot)(1-M)$。最终的增强图像 I_{aug} 重构如下：

$$I_{aug} = I_{person} + \mathcal{T}(I_{bg}) \tag{21}$$

我们执行六种不同的变换类别，以合成多样化的环境条件：

1）色彩空间操作：在 HSV 域中，我们进行随机色调旋转（$\theta \in [30°, 150°]$）、饱和度缩放（$\alpha \in [1.0, 2.0]$）和亮度调制（$\beta \in [0.7, 1.3]$），以模拟不同的光照条件。2）纹理合成：为减少对背景纹理线索的依赖，我们通过 Canny边缘检测和浮雕滤波核进行边缘增强。3）噪声注入：我们引入随机噪声，包括椒盐噪声和加性高斯噪声（μ =0, $\sigma^2 \in [0.01, 0.05]$），以模拟传感器退化。4）模糊模拟：将运动模糊和径向变焦模糊核与背景进行卷积，以模拟相机运动和景深效果。5）图案生成：使用随机颜色绘制合成几何结构，包括网格线、同心圆和对角条纹，以迫使模型忽略结构化背景杂波。6）渐变填充：我们生成线性、径向和角度渐变，并随机设定起始色和结束色，以模拟平滑的无纹理环境。

表 I：VisNet 在 Market-1501 上从头开始训练的准确率表现。

方法	Year	R1 (%)	R5 (%)	R10 (%)	R20 (%)	mAP (%)
TransReID	2021	95.20	98.0	98.7	99.1	89.50
CLIP-ReID	2023	88.10	93.8	95.6	96.9	80.30
VisNet（我们的方法）	**2025**	**87.05**	**93.18**	**95.90**	**97.15**	**77.65**
AANet	2019	82.60	90.5	93.2	95.1	72.20
IDE [21]	2018	79.51	-	-	-	59.87
PSE [22]	2018	87.7	94.5	96.8	-	69.0
TriNet [23]	2017	84.9	-	-	-	69.1
SVDNet [24]	2017	82.3	-	-	-	62.1
MGN（翻转） [25]	2018	95.7	-	-	-	86.9

表II：计算效率分析。VisNet 在准确率有竞争力（82.6% +Rank-1）的各方法中实现了最佳效率权衡：每 GFLOP 可获得 18.91 个准确率百分点，显著优于 TransReID（5.36）和 CLIP-ReID（7.34）。

方法	参数量（M）	浮点运算次数（G）	R1（%）	准确率/GFLOP
VisNet（我们的方法）	**32.41**	**4.601**	**87.05**	**18.91**
CLIP-ReID	63.0	12.0	88.10	7.34
TransReID	86.0	17.8	95.2	5.36
AANet	19.0	2.5	82.60	33.04
MGN（翻转）	68.75	48	95.7	1.99

每种数据增强技术均以可配置的概率 p 和强度 $\lambda \in [0.0, 1.0]$应用，确保生成多样化的训练样本，从而使模型专注于个体特征，而非环境背景。

III. 实验

A. 数据集与评估指标

实验在 Market-1501 数据集上进行，该数据集包含 1,501 个身份、42,872 张（原始为 32,668 张）增强训练图像、3,369 张查询图像和 19,733 张图库图像。输入图像调整为 256×128，四周各填充 10 个像素，然后随机裁剪回 256×128。此外，还以 0.5 的概率进行随机水平翻转，并进行颜色抖动，亮度、对比度、饱和度和色调分别为 0.2、0.15、0.15 和 0.1。此外，以 0.5 的概率进行随机擦除，擦除区域占图像面积的 2–40%，并使用 ImageNet 统计量进行归一化。测试时仅进行尺寸调整和归一化。按照标准协议，我们排除同一摄像机下的匹配结果。我们报告不经重排序计算的 Rank-1、Rank-5、Rank-10、Rank-20 累积匹配特征（CMC）和平均精度均值（mAP），如表 I 所示。

尽管 TransReID 的 Rank-1 准确率达到 95.20%，但它依赖视觉Transformer预训练。同样，CLIP-ReID 利用 CLIP 的大规模视觉语言预训练。VisNet 仅在 Market-1501 上从头开始训练，Rank-1 准确率达到 87.05%，展现出无需外部预训练也能取得的优异性能。我们的方法优于同样采用标准 ResNet50 骨干网络的 AANet，验证了多尺度融合和语义聚类的有效性。

表 III：模型组件分析：VisNet 模型的参数分布

模型组件	参数	百分比
ResNet50骨干网络（阶段0-4）	23,508,032	72.57%
多尺度融合	4,733,444	14.61%
语义聚类头	2,628,100	8.11%
分类器	1,538,048	4.75%
BN颈部	4,096	0.01%
总计	**32,411,720**	**100.00%**

B. 不同指标下的排序质量

除 Rank-1 外，VisNet 在各项排序指标上也表现出色：Rank-5 为 93.18%，Rank-10 为 95.90%，Rank-20 为 97.15%。这表明，经过良好校准的排序对于实际部署很有价值，因为用户需要查看多个候选结果。

表 III中的参数分解揭示了 VisNet 的设计理念。ResNet50骨干网络占总参数的 72.57%，而我们提出的多尺度融合（14.61%）和语义聚类（8.11%）组件则通过高效的正则化机制发挥作用。身份分类器贡献了 4.75% 的参数，用于最终的行人重识别任务。这种架构表明，无需用轻量级替代方案取代骨干网络，通过巧妙的模块设计也能实现具有竞争力的准确率，最终模型总大小为 32.41M 参数。

实验表明，VisNet 具有竞争力的准确率：使用标准骨干网络时，Rank-1分数达到 87.05%，比 AANet 高 4.45%。更重要的是，表 II显示 VisNet 效率出色，每 GFLOP 可取得 18.91 个准确率点，效率超过 $3.5\times$，而 TransReID 为 5.36。AANet 的得分为 33.04，但 R1 仅达到 82.60%。这种精简设计中，语义模块仅使 ResNet50 基础模型的参数开销增加 22.7%，因此 VisNet 适用于资源受限的部署，且不会降低排序性能。

IV. 结论

VisNet 通过将学习到的注意力多尺度融合到每个独立尺度的特征表示中，并利用伪标签引导聚类，在保持效率-准确率权衡的同时实现了具有竞争力的准确率。我们提出的高效方法表明，采用高效轻量的方法也能达到较高准确率；与标准基准模型相比，该方法还可减少 3.87× GFLOPs，因此适用于其他方法过于庞大或资源有限的场景。

V. 未来工作

将学习得到的语义原型与教师-学生蒸馏相结合，可能会提升对身体划分的语义理解。此外，从尺度注意力转向自注意力，可能会加强身体区域之间的联系。

参考文献

[1] M. De Marsico、R. Distasi、S. Ricciardi 和 D. Riccio，“行人重识别方法比较”，载于第三届国际模式识别应用与方法会议（ICPRAM）论文集。SCITEPRESS，2014，第 189–198 页。[2] Y. Sun、L. Zheng、Y. Yang、Q. Tian 和 S. Wang，“超越部件模型：通过改进的部件池化进行行人检索”，载于欧洲计算机视觉会议（ECCV）论文集，2018，第 480–496 页。[3] C. P. Tay、S. Fan、K. Tan 和 T. H. Chong，“AANet：用于行人重识别的属性注意力网络”，载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2019，第 7134–7143 页。[4] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，“注意力即一切”，载于神经信息处理系统进展（NeurIPS），2017，第 5998–6008 页。[5] X. Zhang、R. Cai、N. Jiang、M. Xing、K. Xu、H. Yang、W. Zhu 和 Y. Hu，“Te-TransReID：通过局部特征嵌入和轻量级Transformer实现高效行人重识别”，Sensors，第 25 卷，第 17 期，第 5461 页，2025。[6] Y. Li、J. He、T. Zhang 和 X. Zhang，“CLIP-ReID：以视觉与语言为桥梁实现行人重识别”，arXiv预印本 arXiv:2211 13977，2023。。[7] T.-Y. Lin、P. Doll´ar、R. Girshick、K. He、B. Hariharan 和 S. Belongie，“用于目标检测的特征金字塔网络”，载于IEEE计算机视觉与模式识别会议（CVPR）论文集，2017，第 2117–2125 页。[8] J. Hu、L. Shen 和 G. Sun，“压缩与激励网络”，载于IEEE计算机视觉与模式识别会议（CVPR）论文集，2018，第 7132–7141 页。[9] S. Woo、J. Park、J.-Y. Lee 和 I. S. Kweon，“CBAM：卷积块注意力模块”，载于欧洲计算机视觉会议（ECCV）论文集，2018，第 3–19 页。[10] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于IEEE计算机视觉与模式识别会议（CVPR）论文集，2016，第 770–778 页。[11] W. Chen、X. Zhang 和 Y. Huang，“面向以人为中心的视觉任务的语义可控自监督学习框架”，载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2023，第 15 050–15 060 页。[12] S. Liu、E. Johns 和 A. J. Davison，“结合注意力的端到端多任务学习”，载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2019，第 1871–1880 页。[13] L. Zheng、L. Shen、L. Tian、S. Wang、J. Wang 和 Q. Tian，“可扩展行人重识别：一个基准”，载于IEEE国际计算机视觉会议（ICCV）论文集，2015，第 1116–1124 页。[14] S. Ioffe 和 C. Szegedy，“批归一化：通过减少内部协变量偏移加速深度网络训练”，载于国际机器学习会议（ICML），2015，第 448–456 页。[15] Y. Cho、W. J. Kim、S. Hong 和 S.-E. Yoon，“用于无监督行人重识别的基于部件的伪标签细化”，载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2022，第 7308–7318 页。[16] C. Yan、G. Pang、X. Bai、J. Zhou 和 L. Gu，“超越三元组损失：采用细粒度差异感知成对损失的行人重识别”，arXiv预印本 arXiv:2009.10295，2020。[17] S. Liu、E. Johns 和 A. J. Davison，“结合注意力的端到端多任务学习”，载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2019，第 1871–1880 页。[18] S. Liao、X. Zhao、X. Guo 和 S. Gao，“基于图采样的可泛化行人重识别深度度量学习”，载于IEEE/CVF计算机视觉与模式识别会议（CVPR）论文集，2022，第 7975–7985 页。[19] A. Hermans、L. Beyer 和 B. Wang，“为行人重识别中的三元组损失辩护”，载于arXiv预印本 arXiv:1703.07737，2017。[20] F. Schroff、D. Kalenichenko 和 J. Philbin，“FaceNet：用于人脸识别和聚类的统一嵌入”，载于IEEE计算机视觉与模式识别会议（CVPR）论文集，2015，第 815–823 页。[21] Z. Zheng、L. Zheng 和 Y. Yang，“用于行人重识别的判别式学习卷积神经网络嵌入”，ACM多媒体汇刊计算、通信与应用（*TOMM*），第 14 卷，第 1 期，第 13 页，2018 年。[22] M. S. Sarfraz、A. Schumann、A. Eberle 和 R. Stiefelhagen，《一种结合扩展跨邻域重排序的姿态敏感嵌入方法，用于行人重识别》，载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，2018 年，第 1933–1942 页。[23] A. Hermans、L. Beyer 和 B. Leibe，《为行人重识别中的三元组损失辩护》，*arXiv* 预印本 *arXiv:1703.07737*，2017 年。[24] Y. Sun、L. Zheng、W. Deng 和 S. Wang，《用于行人检索的 SVDNet》，载于*IEEE*国际计算机视觉会议（*ICCV*）论文集，2017 年，第 3800–3808 页。[25] G. Wang、Y. Yuan、X. Chen、J. Li 和 X. Zhou，《利用多个粒度学习行人重识别的判别性特征》，*arXiv* 预印本 *arXiv:1804.01438*，2018 年。