**基于级联聚合的模态协同互补学习的多模态行人重识别**

**Modality Synergy Complement Learning with Cascaded Aggregation for Visible-Infrared Person Re-Identification**

**摘要：**

多模态行人重识别（VI-ReID）在图像检索中具有挑战性。模态差异很容易造成巨大的模态组内差异。现有的大多数方法要么通过模态不变性桥接不同的模态，要么生成中间模态以获得更好的性能。不同的是，本文提出了一个新的框架，称为具有级联聚合的模态协同互补学习网络（MSCLNet）。它的基本思想是协同两种模式来构造不同的身份识别语义表示，并减少噪音。此外，我们提出了用于细粒度优化特征分布的级联聚合策略，该策略逐步聚合子类、类内和类间的特征嵌入。

**引言：**

与ReID相比，VI-ReID面临着巨大的类内差异，主要是因为可见光图像和红外图像之间的巨大的模态差异，模态差异源于由不同波长组成的光的特性。当它们的图像被等效地解析为数字矩阵后，近红外图像更平滑，由于波长更长和散射更多，会丢失纹理细节，它对肤色、反照率和光照变得更加不可知。相似的纹理、散布和颜色可以表示不同的语义。此外，还很难确保摄像机拍摄的视角、行人的服装、遮挡等。这些因素都是VI-ReID面临的巨大挑战。

为了解决上述困难，现有的大多数方法主要关注学习模态不变性，以弥合可见光图像和红外图像之间的差距，或生成中间或相反模态的图像，用于人员检索。基于GAN的方法通常存在计算复杂性和引入噪声的问题。追求模态不变性可能会导致网络忽视语义多样性的特征属性，以及丧失身份鉴别。不同的是，我们考虑了可见光和红外模式之间的不同表示和语义差异。VI-ReID的成功证明，可见光图像的特征对大量身份总是具有足够的辨别力。红外相机倾向于捕捉热物体，而不是非热物体，热敏感性导致语义损失和背景噪声过滤，红外图像表现出相对稳定的相同身份，并且相对不受噪声影响。

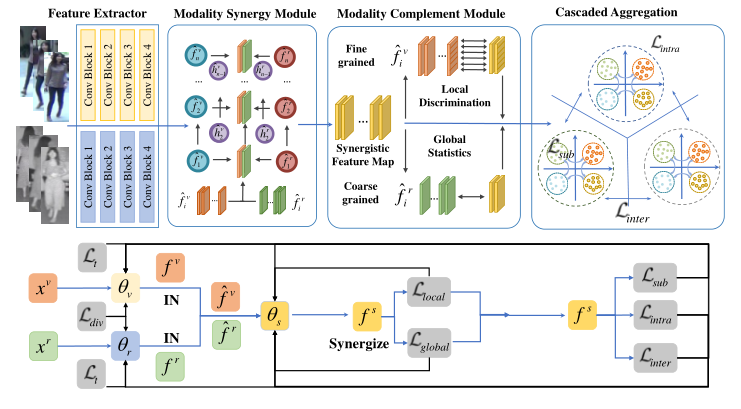
我们的目标是以级联方式在不同层次上进行优化。基本思想是根据相同的拍摄相机将每个身份的实例细分为几个子类。每个子类中的实例更容易聚合，其特征嵌入具有更高的类内相似性。通过这种方式，我们可以逐步限制特征嵌入之间的距离。

因此，我们提出了一个新的框架，即模式协同补充学习网络（MSCLNet）。它旨在减少类内的差异，提高身份差异的表现。首先，它通过与模态协同模块（MS）构建协同特征，保留了可见光和红外模态固有的语义多样性和身份相关性。然后，它通过两种模式的优势增强了协同特征表示。MS和MC极大地提高了网络跨模态身份表示的能力。此外，我们提出了级联聚合策略（CA）来优化特征嵌入的分布。它逐步将样本聚合为子类、类内和不同身份。在级联方式中，属于相同身份的实例倾向于聚合，而属于不同身份的实例则映射为分散。

总结:主要贡献可以总结如下：

1. 提出了一个新的框架，名为模态协同补充学习网络（MSCLNet），用于VI-ReID的级联聚合。为了获取更具区分性的语义，它通过不同的语义以及可见光和红外模态的特定优势来学习增强的特征表示。
2. 提出了一个模态协同模块（MS），它创新性地挖掘了模态特有的不同语义。
3. 提出了一个模态补充模块（MC），它通过两个模态特有优势的并行指导进一步增强了特征表示。它们为进一步的高级身份表示提供了参考。
4. 设计了级联聚合策略（CA），以在细粒度级别优化特征嵌入的分布。它以级联方式逐步聚合整体实例，并增强身份的区分能力。

**模态协同补充学习**

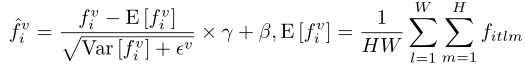


MSCLNet框架它主要包含三个主要组件：模态协同模块（MS）、模态补充模块（MC）和级联聚合策略（CA）。我们利用MS来协同来自提取器的模态特定的不同语义，然后在两种模态优势的指导下使用MC来增强特征表示。为了优化特征的分布并聚合相同身份的实例，我们利用CA以细粒度和渐进的方式约束特征的分布。最后，我们总结了建议的损失函数。

现有的方法提取模态共享特征的代价是丢弃能够很好地描述人的模态特定的多样性语义。因此，我们考虑到这些内在的不同语义和每种情态的特殊优势，以学习更精确和更好的身份鉴别表示。

**模态协同模块（MS）**

给定一对可见光和红外图像、，双流网络提取其特征和。特征和通过以下操作进行标准化：



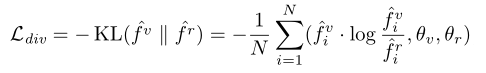


让S(·)表示模态协同模块，以在，的基础上用标签yi构建协同特征：

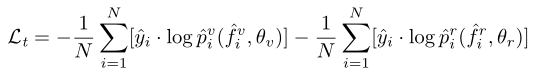


其中θs作为模态协同模块S(·)的参数。我们利用Mogrifier LSTM作为协同特征编码器，以最大化模态协同学习的效果

为了构造具有不同语义的，我们利用KL-Divergence来约束可见光和红外特征，的逻辑分布，其公式如下：



引入交叉熵来约束可见光和红外特征：

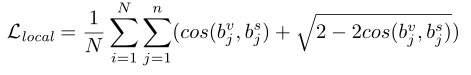




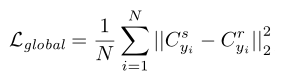
**模态补充模块**

尽管协同表示包含更多与身份相关的不同语义，但尚不确定协同特征是否优于可见光和红外特征的组合Concat（）。由于红外图像包含较少噪声的全局行人信息，而可见图像包含细粒度的区分语义，因此我们从两个方面增强了协同特征的表示有效性。考虑到细粒度语义，我们利用在局部方面的优势来增强协同特征。考虑到粗粒度语义，我们利用红外特征的全局优势增强了协同特征。

在细粒度级别上，我们将可见和协同特征拆分为n=6个部分，作为MPANet，并获得单独的特征块，如。协同特征的局部区分可以通过可见形态的细微区域来增强。余弦相似性cos(·,·)用于优化过程。



同时，在粗粒度级别上，协同特征的全局信息可以通过红外模态的中心一致性来优化。

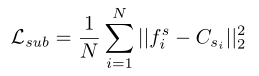


其中表示协同特征，的类中心。Lglobal有助于协调协同特征和红外特征的语义，并过滤协同表示的无关身份。

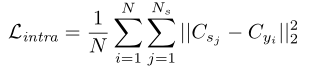


**级联聚合策略**

1. 子类级别的聚合。我们将每个图像的拍摄相机的身份作为自然子类，因为由同一台相机拍摄的同一个人的图像彼此具有高度的相似性，其中表示该子类的中心：



1. 类内聚合。聚合的公式可以表示如下，其中Ns表示每个标识的子类的数量：



1. 类间级别的聚合。我们的聚合方法不仅最大化类内实例的相似性，而且从整体上最大化类间实例的差异性。从形式上讲，不同身份之间的分散可以表示为：

