**基于结构感知位置的Transformer多模态行人重识别**

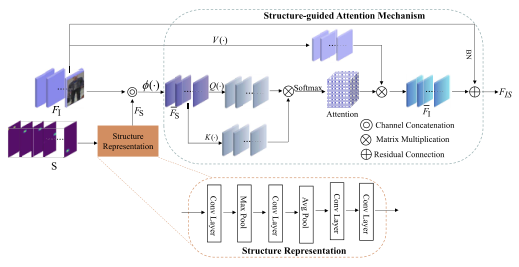
**摘要：**

可见-红外多模态行人重识别（VI-ReID）是一个跨模态检索问题，其目标是在可见光和红外摄像机之间匹配同一行人。由于两种模式之间存在姿势变化、遮挡和巨大的视觉差异，以往的研究主要集中在学习图像级的共享特征。由于它们通常学习全局表示或提取均匀分割的局部特征，因此这些方法对错位敏感。本文提出了一种结构感知的位置变换器（SPOT）网络，利用结构和位置信息去学习共享的模态特征。它由两个主要组件组成：attended structure representation（ASR）和transformer-based part interaction（TPI）。通过ASR和TPI的加权组合，提出的SPOT探索了丰富的上下文和结构信息，有效地减少了跨模态差异，增强了对错位的鲁棒性。

大多数先前的VI-ReID研究建议学习可共享的跨模态外观特征，以应对上述挑战。通常，这些方法首先使用两个相同的非共享权重网络分别提取可见光和红外图像的模态特定特征，以处理跨模态视觉差异。然后，利用共享网络学习模态共享特征并进行跨模态相似性优化。虽然这些方法可以提取某些模态共享特征并提高模型的性能，但它们对失调和背景噪声的鲁棒性是有限的，因为它们只从中学习粗糙的共享特征。

**主要贡献：**

1. SPOT网络将结构相关的外观学习和局部交互学习相结合，以增强VI-ReID的语义共享模态表示。
2. ASR模块学习每个模态的结构和外观特征，以解决复杂的背景噪声。
3. TPI通过建模上下文和位置关系，自适应地组合部分可识别的线索，以提高对姿势变化和遮挡的鲁棒性，从而增强局部特征的区分能力。
4. **Attended Structure Representation(ASR)**



由于ReID任务中没有标记结构信息，因此无法直接获取结构特征。我们采用一个具有四个卷积层和两个池化层的关系网络来处理不同关键点热图之间的关系,如下：

我们采用OpenPose获得关键点热图, 其中表示关键点的数量，表示每个关键点热图的大小。

给定可见/红外图像I，外观特征图, 通过模态专用网络Nspecific(·)和模态共享网络Nshared(·)计算

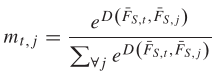


在计算注意矩阵之前，融合了结构特征和外观特征，以减少这两个特征之间语义空间差异的影响。



其中表示最终的结构特征，它是关系网络Nr最后一层的输出，与外观特征具有相同的大小。和是嵌入网络φ(·)的两个参数，通过1×1卷积BN-ReLU层实现。

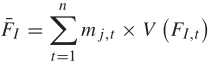
空间关系注意力为：





表示特征节点t和特征节点j之间的相似性，K(·)和Q(·)是两个嵌入函数，由1×1卷积层实现，将结构特征映射到不同的语义空间。

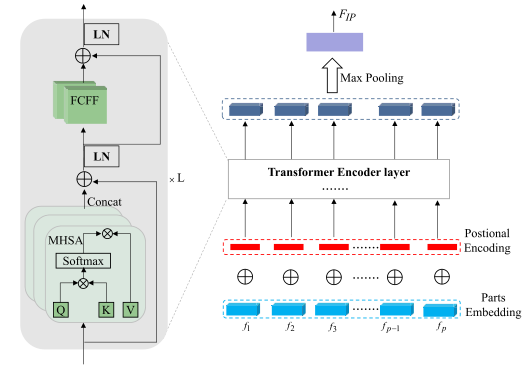
注意力定位区分区域，可用于指导外观特征学习：



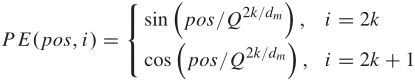
表示带有卷积运算V(·)的嵌入外观特征, 最终结构相关特征由以下公式表示：



1. **Transformer-Based Part Interaction**



使用正弦和余弦函数来表示位置编码：



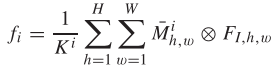
dm和i分别代表一个patch的行和列。Q代表一个常量参数。位置嵌入是Transformer的重要组成部分，它捕获每个元素的位置信息。通过将位置嵌入直接添加到特征嵌入中，每个序列元素的位置信息与其嵌入信息充分集成，并传递给高级特征。此外，通过位置信息的嵌入，可以区分元素之间的距离，有助于序列结构信息的挖掘。

部件交互：（p是部件数）。TPI试图探索区域之间的区别上下文信息和结构关系，以增强零件级特征表示。这个过程包括两个步骤：部件划分和部件交互。

采用结构信息来实现跨模态人员ReID的细粒度局部划分。具体来说，节点级零件图是从结构特征中学习的，并描述每个节点属于特定区域的可能性。可以被表示：



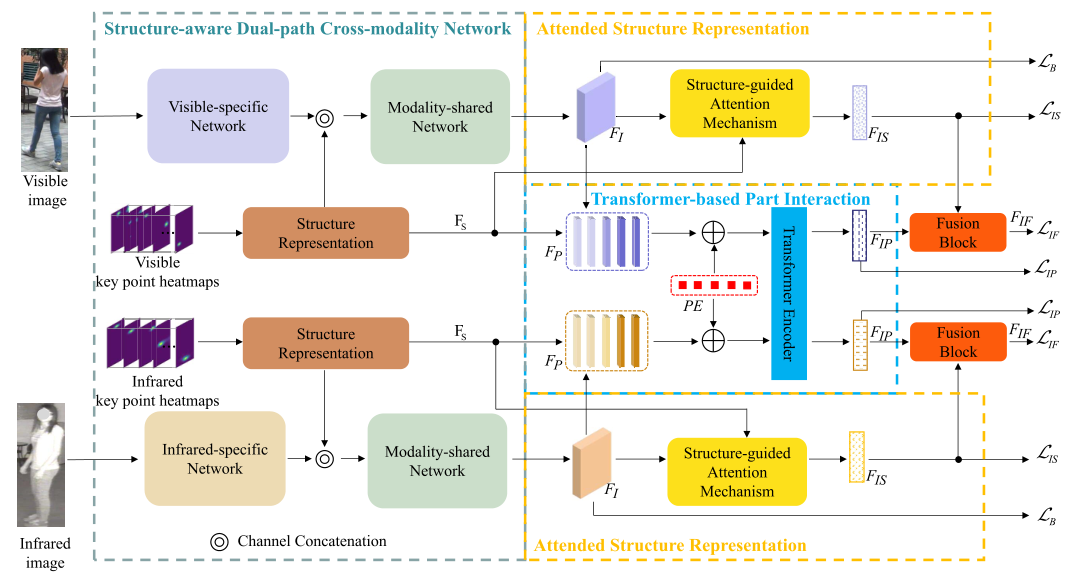
带有参数Wθ的θ(·)是一个嵌入函数，它通过1×1卷积层和Sigmoid函数对结构特征进行编码，作为部件注意映射。然后，可以通过以下方式获得第i个零件特征fi：





PE表示部件序列的位置编码。

1. **Overall Architecture**





**Fusion Block：**在获得与结构相关的外观特征和零件特征后，我们建议通过一个融合块来聚合这两个特征，并输出最终的模态特征进行相似性匹配，。



