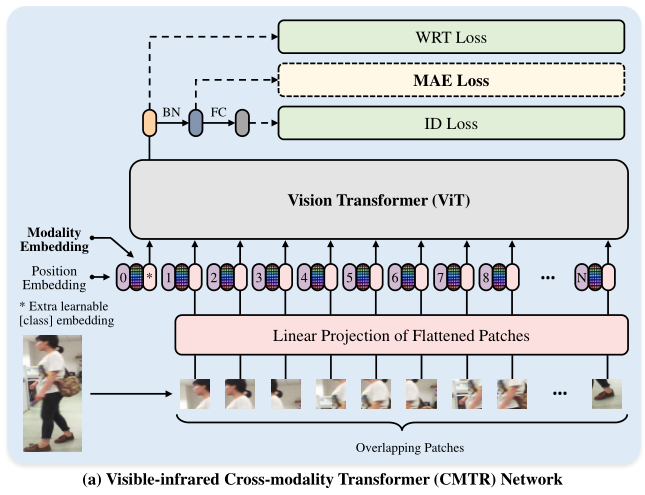
**CMTR: Cross-modality Transformer for Visible-infrared Person Re-identification**

**跨模态的行人重识别Transformer**

**摘要：**

现有的基于卷积神经网络的方法主要面临着对模态信息感知不足的问题，无法学习对身份具有良好鉴别能力的模态不变量。提出了一种基于Transformer的跨模态行人重识别方法（CMTR），该方法可以明确地挖掘每个模态的信息，并基于它生成更好的鉴别特征。具体来说，为了捕获模态的特征，我们设计了新的模态嵌入，并将其与标记嵌入融合，以编码模态信息，并提出了一种新的损失函数。

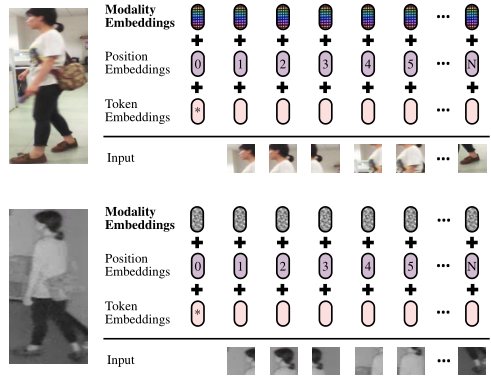
**整体架构：**



对于输入图像，我们让vis和ir表示可见模态和红外模态。在一个训练批次中，有B个图像，vis和ir数量相同。我们的方法从下到上主要包括三个阶段：输入嵌入、特征提取和多重损失约束。

对于输入的图像，进行切割分块转化为，也即是，P代表的是每个切块的宽高，N代表的是序列数。我们通过步长S（S<P）（软分割）生成重叠的块，以增强相邻块之间的相关性。将X从映射为X，其中D代表P2\*C。一个额外的可学习[cls]标记嵌入被合并到序列中，以捕获整个图像的全局注意力。最终。

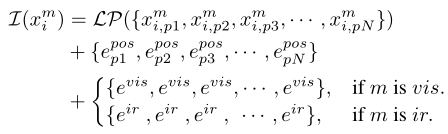
**模态嵌入：**



CMTR的输入嵌入

感知模态特征有助于生成模态不变特征。然而，许多现有方法都忽略了该关键点。为了实现这一点，我们在CMTR模型中引入了模态嵌入（ME），它直接旨在学习和捕获每种模态的固有信息和特征。

我们的CMTR的输入嵌入由部分，即标记嵌入、位置嵌入和模态嵌入。前两种方法与之前的方法一致，每种模态的图像享有相同的模态嵌入。



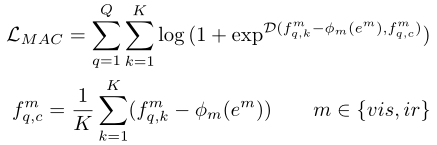
evis和eir分别表示可见模态嵌入和红外模态嵌入，位置嵌入epos因为补丁位置而改变，而模态嵌入em（m∈ {vis，ir}）根据图像的模态形式的不同去感知不同的模态信息。

**模态感知增强损失**

MAE损失作用于批量归一化（BN）后提取的特征，这些特征在测试期间用作匹配特征。表示提取的特征。MAE损失由两部分组成，模态感知中心丢失和模态感知ID丢失。

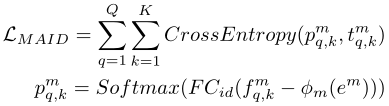


对于LMAC的定义，它侧重于缩小同一身份下不同模态之间的差距，并利用从ME学到的知识来缩小类内特征的距离。



表示从m模态的q身份的k图像中提取的特征，表示挖掘模态嵌入em的映射，在这个公式中，我们让直接减去对应的，以去除模态特定信息并过滤出模态不变特征。表示q身份的中心特征向量，这是模态特定信息去除后图像特征的平均值。

模态感知ID丢失LMAID旨在学习不同身份之间的区别特征，它也基于学习到的ME信息，旨在区别不同ID图像特征之间的距离。



是图像特征去除模态特定信息后的特征。

通过优化模态感知增强损失LMAE，首先，网络可以利用模态去除过程强制ME挖掘更有用的模态特定特征，这是增强ME表示的更直接的方法。其次，基于ME的损失函数可以调整特征嵌入的分布，使其对图像检索更具区分性，并且不受异构跨模态鸿沟的影响。

