本发明的工作流程如图所示，具体步骤：

1.获取带有身份标签的可见光VIS行人图像和红外IR行人图像，将它们**输入至辅助模态生成器中以生成对应的辅助模态图像**。

2.生成的辅助模态图像VtA、ItA和原始的可见光图像VIS和红外图像IR一同**输入至主干网络中提取模态共享特征**。

3.将从主干网络提取到的特征图水平划分若干块，**进行局部特征学习**，帮助模型获取行人图像的局部信息。

4.联合使用分布一致性损失、身份损失和三元组损失**优化**跨模态行人重识别模型

5.使用训练好的模型进行可见光图像和红外图像间的跨模态行人重识别。

**步骤1所述辅助模态生成器结果如图所示，**采用编码器-解码器架构，由两个编码器和两个解码器组成。生成辅助图像的具体过程为：原始的可见光图像和红外图像输入至辅助模态生成器中，可见光图像经过可见光模态信息编码器进行编码后输入至可见光模态解码器，解码器会将可见光图像投影到一个合适的辅助模态空间，以生成可见光的辅助模态。红外图像同理，经过红外模态信息编码器进行编码后输入至红外模态解码器以生成红外图像的辅助模态。

**步骤1所述辅助模态生成器结果如图所示，**采用编码器-解码器架构，由两个编码器和两个解码器组成。生成辅助图像的过程为：原始的可见光图像和红外图像输入至辅助模态生成器中，可见光图像经过可见光模态信息编码器进行编码后输入至可见光模态解码器，解码器会将可见光图像投影到一个合适的辅助模态空间，以生成可见光的辅助模态图像。红外图像同理，经过红外模态信息编码器进行编码后输入至红外模态解码器以生成红外图像的辅助模态图像。

具体来说，首先对输入图像进行预处理，将输入尺寸调整为3×384×192，由于可见光图像和红外图像的通道数不同，因此将红外图像的通道数设置成3通道，与可见光图像的通道数保持一致。然后两种模态的图像分别输入至对应的模态信息编码器进行编码。在编码过程中，首先会通过一个卷积核为3\*1\*1的卷积层，将三通道图像转换为单通道图像，然后通过卷积核为1\*1\*1的卷积层，减少计算量，最后通过批归一化层进行数据归一化。然后将归一化后的数据输入至对应的解码器进行解码，将编码后的单通道图像通过卷积核为1\*3\*1的卷积层转换为三通道图像，以生成对应的辅助模态图像。

**步骤2所述的主干网络**采用ResNet50作为双流网络的基础网络，前三个网络层参数独立，用以提取模态特定特征。后面的网络层参数共享，用以提取模态共享特征。其中，在网络层3和网络层4后面嵌入的CBAM（卷积注意力机制）模块，该模块采用了混合注意力机制，先后集成了通道注意力机制模块和空间注意力机制模块，能帮助模型排除背景噪声的干扰，关注行人图像间的细微差异，挖掘更具鉴别性的特征信息。CBAM模块的计算公式所下所示：



通道注意力机制可以帮助模型了解哪个通道的内容比较重要，表示特征图F经过通道注意力模块后的输出。其定义如下所示：



表示Sigmoid激活函数。对于输入特征图F，首先分别经过平均池化和最大池化得到最大池化特征和平均池化特征，然后经过一个共享的多层感知机MLP网络得到通道注意力特征图。

空间注意力机制可以帮助模型了解哪些区域比较重要，表示特征图F’经过空间注意力模块后的输出。其定义如下所示：



表示Sigmoid激活函数，表示卷积核尺寸为7x7的卷积运算。对于输入特征图F‘，首先分别经过平均池化和最大池化得到最大池化特征和，将两个特征融合后由一个标准的卷积层作卷积操作，最后经过激活函数输出得到空间注意力特征图。

步骤三所述的局部特征学习具体流程为：首先对主干网络提取的特征图进行广义平均池化，然后将其平均匀切分成四块，进行局部特征学习，这样可以帮助模型关注图像细节信息，减少背景噪声的干扰。广义均值池化，可以帮助模型获取图片的细粒度信息。它的定义如下所示：



表示池化层的输出，表示池化层的输入。是一个超参数，在反向传播中不断优化。当=1时，GeM广义均值池化近似于全局平均池化；当时，GeM广义均值池化近似于全局最大池化。

步骤4中提到的联合损失公式如下所示：



其中，是分布一致性损失,用于优化辅助模态生成器所生成的图像，缓解模态模态差异。它的定义如下所示：



N是训练阶段每一批次的训练样本的总和。和表示由可见光图像和红外图像分别生成的辅助模态图像。表示辅助模态图像经过全连接层后的输出。表示取均值。

其中，是交叉熵损失函数，它的定义如下所示：

，

N表示训练过程中的行人身份个数，y表示图像的行人身份标签，表示预测是类别i的可能性，是一个常量，这里设置为0.1.

是三元组损失函数，它能帮助模型拉近不同模态、相同身份的行人特征间的距离，推远相同模态、不同身份的行人特征间的距离。在本发明中，四种模态间的三元组损失定义如下所示：



可见光图像和红外图像两种模态间的三元组损失定义如下：



N表示每种模态图像在一批次的训练样本中的数量总和，表示欧式距离，是一个阈值常量，这里设置为0.3。其他模态间的三元组损失和上面所定义的公式相似。