多模态行人重识别

1.论文：Deep Learning for Person Re-identification: A Survey and Outlook

基于深度学习的行人重识别的总结和展望

作者：Mang Y e, Jianbing Shen, Senior Member, IEEE, Gaojie Lin, Tao Xiang, Ling Shao and Steven C. H. Hoi, Fellow, IEEE

摘要：

人员重新识别（ReID）旨在通过多个不重叠的摄像头检索感兴趣的人员。该领域分为封闭世界（closed-world）和开放世界（open-world）。我们首先从三个不同的角度对封闭世界的 ReID 进行了全面的概述和深入的分析，包括深度表征学习、深度度量学习和排名优化。我们为 ReID 引入了一个新的评估指标（mINP），表示查找所有正确匹配项的成本。

1.引言

给定一个感兴趣的询问者，ReID 的目标是确定此人是否在不同的时间出现在另一个地方，或者是在不同的时刻出现在同一台相机上。由于公共安全的迫切需求和监控摄像机数量的不断增加，人的身份识别在智能监控系统中势在必行，具有重大的研究影响和现实意义。

由于存在不同的视点、不同的低图像分辨率、照明变化、无约束姿势、遮挡、异构模式[、复杂的相机环境、背景杂波、不可靠的边界框生成等，重新识别是一项具有挑战性的任务。这些都会导致不同的变化和不确定性。

2.主要贡献

新的评价标准 mINP:

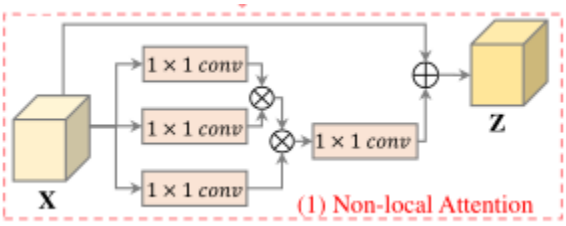
$$NP_i = \frac{R_i^{hard} - |G_i|}{R_i^{hard}}$$

R_i^{hard}表示最难匹配的排名位置，$|G_i|$表示查询 i 的正确匹配总数。

$$\text{mINP} = \frac{1}{n}\sum_i (1 - \text{NP}_i) = \frac{1}{n}\sum_i \frac{|G_i|}{R_i^{hard}}$$

新的基准方法 AGW:

1. Non-local 注意力机制的融合



2. Generalized-mean (GeM) Pooling 的细粒度特征提取

$$\mathbf{f} = [f_1 \cdots f_k \cdots f_K]^T, f_k = (\frac{1}{|\mathcal{X}_k|}\sum_{x_i \in \mathcal{X}_k} x_i^{p_k})^{\frac{1}{p_k}}$$

用一句话说：在最低纬度上，对每个元素的 p 次方求均值，再开 p 次方。

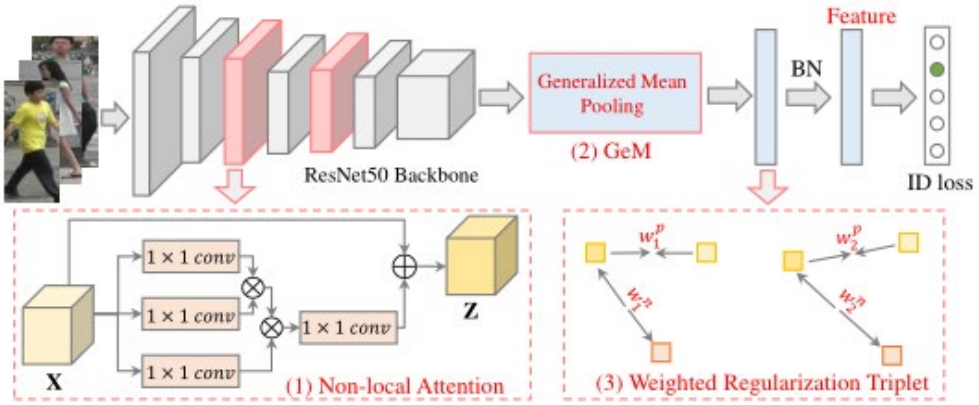
3. 加权正则化的三元组损失（Weighted Regularization Triplet (WRT) loss）

$$\mathcal{L}_{wrt}(i) = \log(1 + \exp(\sum_j w_{ij}^p d_{ij}^p - \sum_k w_{ik}^n d_{ik}^n)).$$

$$w_{ij}^p = \frac{\exp(d_{ij}^p)}{\sum_{d_{ij}^p \in \mathcal{P}_i} \exp(d_{ij}^p)}, w_{ik}^n = \frac{\exp(-d_{ik}^n)}{\sum_{d_{ik}^n \in \mathcal{N}_i} \exp(-d_{ik}^n)}$$

$d^p{}_{ij}/d^n{}_{ik}$表示正负样本对的成对距离。

3.整体框架



2.论文： CMTR: Cross-modality Transformer for Visible-infrared Person Re-identification

CMTR：跨模态的 Transformer 用于多模态行人重识别

作者：Tengfei Liang, Yi Jin1, Yajun Gao, Wu Liu, Songhe Feng, Tao Wang, Yidong Li

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。

1.引言

现有的 ReID 方法主要侧重于 RGB 摄像机下单个可见模态中的人员检索，这限制了方法只能在白天使用。为了实现全天候智能视频监控，基于现有监控摄像机在夜间自动切换到红外模式的机制。VI -ReID 需要能够匹配可见光和红外模式之间相同身份的图像的方法，因为存在巨大的异质差距，这更具挑战性。

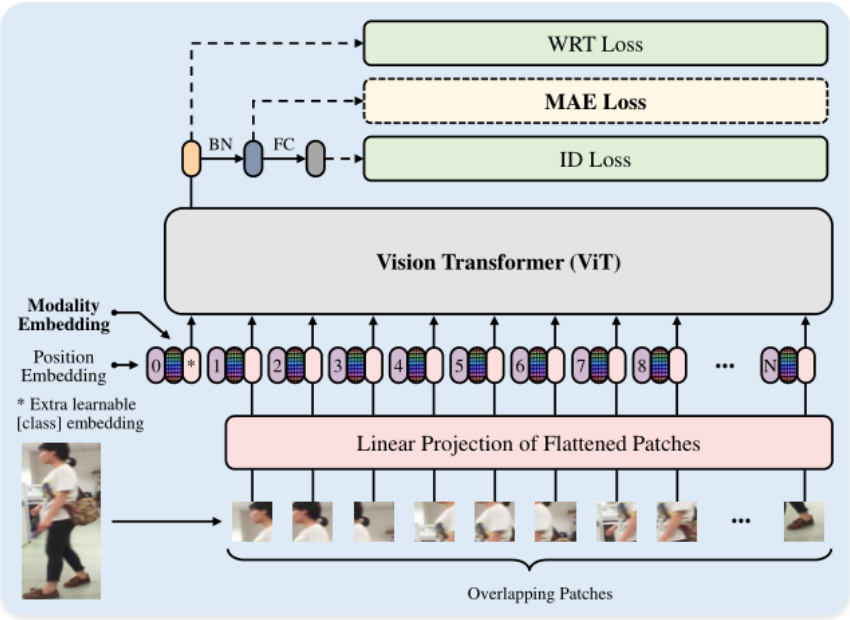
Transformer 模型可以获得具有自注意力模块的全局感受野和完整的空间特征，而不需要汇聚层，但它仍然无法解决跨模态任务的差距问题。本文提出了 cross-modality Transformer（CMTR）模型，该模型可以通过可学习的显式嵌入来捕获模态特征，并在此基础上生成更有效的匹配嵌入。具体来说，为了挖掘模态的特征，我们首先将模态嵌入（ME）引入到我们的方法中。与普通 Transformer 中位置嵌入的想法类似，通过添加补丁的标记嵌入，我们的 ME 可以集成到变压器框架的输入阶段。对应于可见光和红外模式，我们定义了两个可学习的嵌入。它们用于学习每个模态的信息，这有助于后续的模态不变量嵌入学习过程。为了增强对可学习 ME 的约束并优化匹配嵌入的分布，我们进一步设计了一个新的损失函数，即模态感知增强（MAE）损失。它包括模态感知中心损失和模态感知 ID 丢损失以及模态移除过程。通过从 ME 中减去学习到的模态知识，尝试拉近类内特征和区分类间特征。

2.主要贡献

1. 提出了一种新的 cross-modality Transformer（CMTR）网络，这是首次基于 Transformer 的可见-红外行人重识别。
2. 在 CMTR 网络中引入了可学习模态嵌入（ME），它可以直接挖掘模态信息，并可以有效地用于缓解异构图像之间的差距。
3. 设计了一种新的模态感知增强（MAE）损失函数，强制 ME 捕获每种模态的更有用特征，并帮助生成区分特征。

3.方法介绍

3.1 整体框架

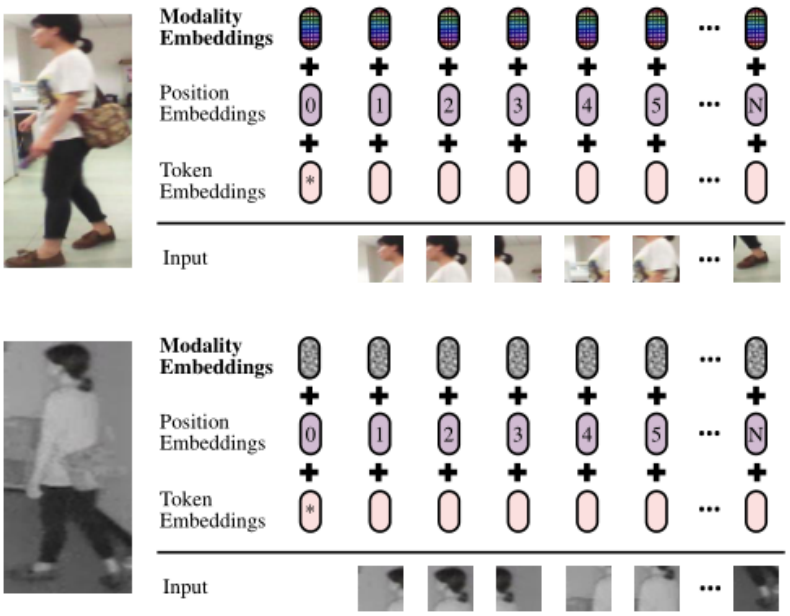


(a) Visible-infrared Cross-modality Transformer (CMTR) Network

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$X \in R^{C*H*W}$，进行切割分块转化为$X \in R^{C*(\frac{H}{P}*\frac{W}{P})*P^2}$，也即是$X \in R^{C*N*P^2}$，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$\in R^{N*D}$，其中 D 代表 P^2*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$X \in R^{(N+1)*D}$。

3.2 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。

$$\begin{aligned}\mathcal{I}(x_i^m) = \mathcal{LP}(\{x_{i,p1}^m, x_{i,p2}^m, x_{i,p3}^m, \cdots, x_{i,pN}^m\}) \\ + \{e_{p1}^{pos}, e_{p2}^{pos}, e_{p3}^{pos}, \cdots, e_{pN}^{pos}\} \\ + \begin{cases}\{e^{vis}, e^{vis}, e^{vis}, \cdots, e^{vis}\}, & \text{if } m \text{ is } vis. \\ \{e^{ir}, e^{ir}, e^{ir}, \cdots, e^{ir}\}, & \text{if } m \text{ is } ir.\end{cases}\end{aligned}$$

e^{vis} 和 e^{ir} 分别表示可见模态嵌入和红外模态嵌入，位置嵌入 e^{pos} 因为补丁位置而改变，而模态嵌入 e^m（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。$f_i^m = BN(V_i^m)$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。

$$\mathcal{L}_{MAE} = \mathcal{L}_{MAC} + \mathcal{L}_{MAID}$$

对于 L_{MAC} 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。

$$\mathcal{L}_{MAC} = \sum_{q=1}^{Q}\sum_{k=1}^{K} \log\left(1 + \exp^{\mathcal{D}(f_{q,k}^{m} - \phi_m(e^m), f_{q,c}^{m})}\right)$$

$$f_{q,c}^{m} = \frac{1}{K}\sum_{k=1}^{K}(f_{q,k}^{m} - \phi_m(e^m)) \qquad m \in \{vis, ir\}$$

$f_{q,k}{}^{m}$表示从 m 模态的 q 身份的 k 图像中提取的特征，$\phi_m(.)$表示挖掘模态嵌入 e^m 的映射，在这个公式中，我们让$f^{m}{}_{q,k}$直接减去对应的$\phi_m(.)$，以去除模态特定信息并过滤出模态不变特征。$f^{m}{}_{q,c}$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 L_{MAID} 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。

$$\mathcal{L}_{MAID} = \sum_{q=1}^{Q}\sum_{k=1}^{K} CrossEntropy(p_{q,k}^{m}, t_{q,k}^{m})$$

$$p_{q,k}^{m} = Softmax(FC_{id}(f_{q,k}^{m} - \phi_m(e^m)))$$

$p_{q,k}{}^{m}$ 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 L_{MAE}，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。

$$\mathcal{L}_{overall} = \mathcal{L}_{ID} + \mathcal{L}_{WRT} + \lambda \cdot \mathcal{L}_{MAE}$$

3.论文：NFormer: Robust Person Re-identification with Neighbor Transformer

NFormer:使用 Neighbor Transformer 使行人重识别具有鲁棒性

作者：Haochen Wang, Jiayi Shen1, Yongtuo Liu, Yan Gao, Efstratios Gavves

摘要：

人物再识别旨在通过不同的摄像机和场景在高度不同的环境中检索人物，在这种情况下，健壮且有区别的表征学习至关重要。大多数研究都考虑从单个图像中学习表征，忽略了它们之间的任何潜在交互作用。NFormer:它对所有输入图像之间的交互进行显式建模，从而抑制了异常特征并导致整体上更健壮的表示。由于建模大量图像之间的交互是一项具有大量干扰因素的艰巨任务，NFormer 引入了两个新模块，即 Landmark Agent Attention 和 Reciprocal Neighbor Softmax。

1. 引言

基于图像的人物再识别(Re-ID)旨在从不同相机和场景拍摄的大量图像中检索出特定的人物。迄今为止，大多数研究主要研究如何从单幅图像中获得更具辨别性的特征表示，或通过注意模块部分表示学习，或 GAN 生成。然而，Re-ID 的一个主要挑战是任何个体在他们的外在因素，如不同的相机设置，照明，视点，闭塞，或内在因素，如衣着改变。因此，在对应于特定个体的表征中存在高度的身份内变异，导致不稳定匹配和异常值的敏感性。

针对身份内部高度差异的一种可能的补救方法是利用来自同一身份的不同图像中存在的知识。部分研究着重于利用一个批次的少量图像的相似性图来建模，然而，这些工作仅侧重于在训练时建模少数图像之间的关系，而在测试过程中，由于计算的限制，它们独立地提取每个图像的表示，这不可避免地会丢失交互，导致训练和测试之间的差距。此外，它们只在每个训练批次中的一小组图像之间建立关系，因此可以相互学习的相关信息有限。

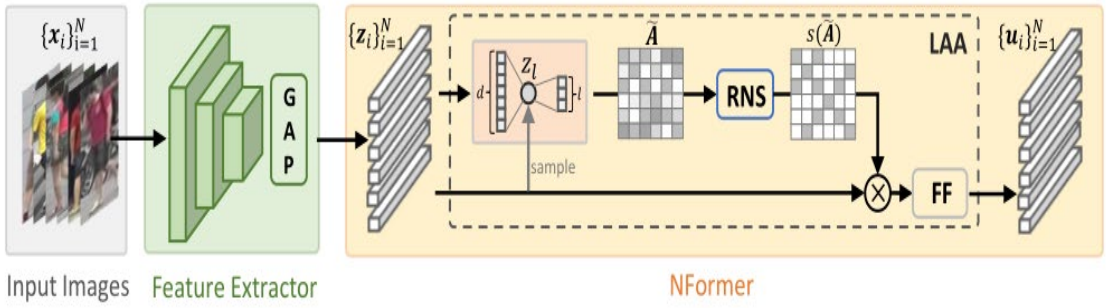
我们提出了一种邻域变压器网络，简称 NFormer，以在训练和测试时有效地建模所有输入图像之间的关系。NFormer 计算表示各个表示之间关系的关联矩阵，然后根据关联矩阵执行表示聚合过程。图像之间关系建模的参与抑制了较高的内部身份差异，并导致更强大的特征。

2.主要贡献：

1. 提出了一个 Neighbor Transformer Network (NFormer)，以有效地在训练和测试时建模所有输入图像之间的关系
2. 提出了一个 Landmark Agent Attention (LAA)，通过在表示空间中引入少量 landmark agent 来减少亲和矩阵的计算量
3. 提出 Reciprocal Neighbor Softmax (RNS)函数，来实现稀疏 attention，只关注计算上可管理的邻居。RNS 显著地约束了不相关个体之间的噪声交互，使表示聚合过程更加有效和高效

3.方法介绍

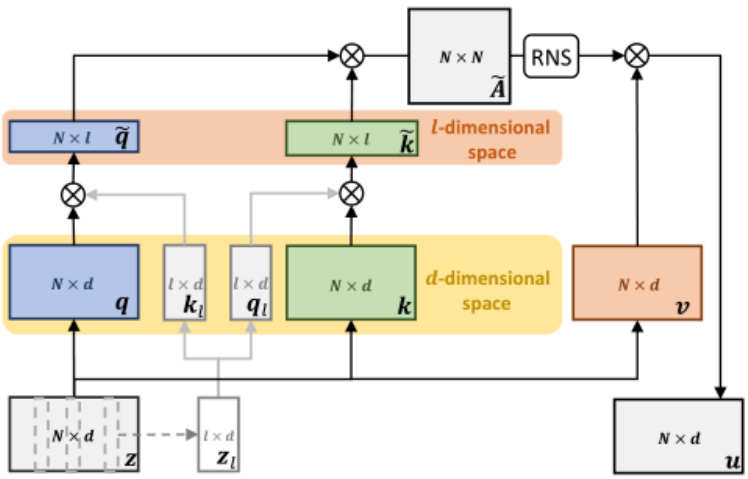
3.1 整体框架



GAP：全局平均池，LAA: Landmark Agent Attention， RNS: Reciprocal Neighbor Softmax，FF：前馈网络

整个架构主要由两部分组成：特征提取器和 NFormer。其中，NFormer 主要由两部分组成：Landmark Agent Attention (LAA)和 Reciprocal Neighbor Softmax (RNS)。

3.2 Landmark Agent Attention



如果依照传统的 Attention 进行计算，由输入 Z 得到$Q,K,V \in R^{n*d}$,这样在计算 q 和 k 获取相似性矩阵时复杂度较高，为 $O(N^2d)$。 为此，作者如下改进：(用 m 代替图中的 l 方便区分数字 1，下面公式中 m 和 l 一致)

(1) 在输入 z 中随机采样 m 个样本得到 z_m，然后生成 k_m 和 q_m，这样特征就从 N×d 降为 m×d。

(2) 将原始的 q 和 k 通过与 k_m 和 q_m 分别相乘，得到$\hat{q} \in R^{N*m}$和$\hat{k} \in R^{N*m}$。

(3) $\hat{q}$和$\hat{k}$计算得到 N×N 的相似性矩阵$\hat{A}$ 。通过上述操作，将复杂度就从 $O(N^2d)$ 降低为 $O(N^2m)$。根据文中描述，m=5,d=256，显著的降低了计算量。

$$\widetilde{\mathbf{A}}_{ij} = (\mathbf{q}\mathbf{k}_l^\top)_i (\mathbf{k}\mathbf{q}_l^\top)_j^\top / \sqrt{d} = \widetilde{\mathbf{q}}_i \widetilde{\mathbf{k}}_j^\top / \sqrt{d}$$

3.3Reciprocal neighbor softmax (RNS):

原始的 Softmax 计算是聚合所有的样本，但是不相关样本的显著存在会对最终计算产生负面影响。同时，Softmax 的计算复杂度也是 $O(N^2d)$ 的，因此作者认为对于每张图片只需要取与他最具有相似性的前 k 张图片进行特征聚合即可。

作者提出 Reciprocal Neighbor Softmax (RNS)，用 reciprocal 邻居掩码强制对少数相关的 attention weight 进行稀疏化。根据这一操作，作者首先根据 $\hat{A}$ 将每行的前 k 个最大的位置设置为 1，其他位置设置为 0，得到 M^k。

$$\mathbf{M}_{ij}^k = \begin{cases} 1, & j \in \text{topk}(\widetilde{\mathbf{A}}_{i,:}) \\ 0, & \text{otherwise.} \end{cases}$$

将 M^k 与其转置相乘，就可以得到一个掩码矩阵 M：

$$\begin{aligned} \mathbf{M}_{ij} &= \mathbf{M}^k \circ \mathbf{M}^{k\top} \\ &= \begin{cases} 1, & j \in \text{topk}(\widetilde{\mathbf{A}}_{i,:}), i \in \text{topk}(\widetilde{\mathbf{A}}_{:,j}) \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

$$\text{RNS}(\mathbf{A})_{ij} = \frac{\mathbf{M}_{ij} \exp(-\widetilde{\mathbf{A}}_{ij})}{\sum_k \mathbf{M}_{ik} \exp(-\widetilde{\mathbf{A}}_{ik})},$$

4.论文：Structure-Aware Positional Transformer forVisible-Infrared Person Re-Identification

基于结构感知位置的 Transformer 多模态行人重识别

作者：Cuiqun Chen , Mang Y e , Meibin Qi , Jingjing Wu , Jianguo Jiang, and Chia-Wen Lin ,

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。

1.引言：

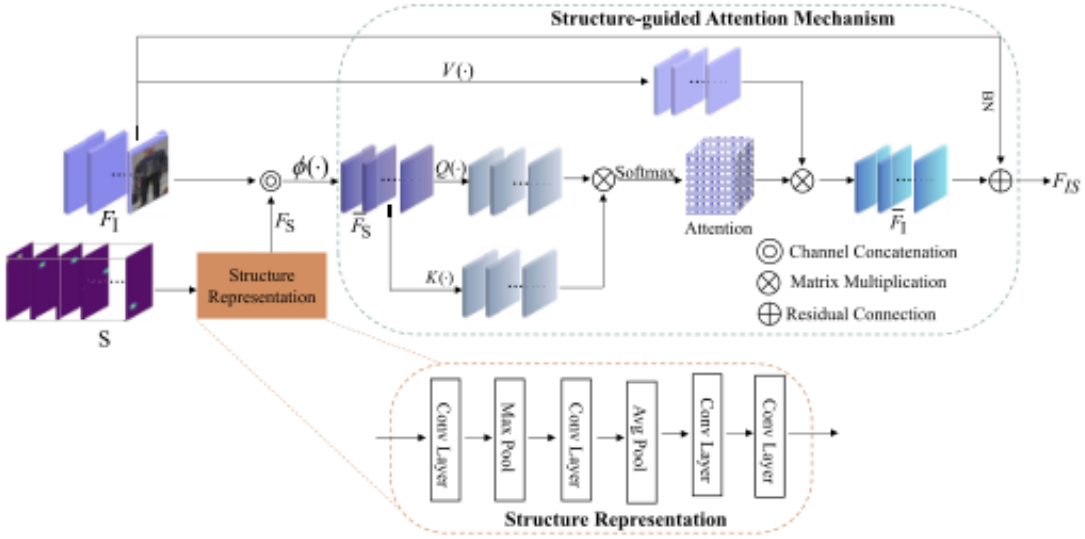
大多数先前的 VI-ReID 研究建议学习可共享的跨模态外观特征，以应对上述挑战。通常，这些方法首先使用两个相同的非共享权重网络分别提取可见光和红外图像的模态特定特征，以处理跨模态视觉差异。然后，利用共享网络学习模态共享特征并进行跨模态相似性优化。虽然这些方法可以提取某些模态共享特征并提高模型的性能，但它们对失调和背景噪声的鲁棒性是有限的，因为它们只从中学习粗糙的共享特征。

2.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。

3.方法介绍

3.1 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：

$$F_S = N_r(S)$$

我们采用 OpenPose 获得关键点热图$S \in R^{K_S * H_S * W_S}$, 其中$K_S$表示关键点的数量，$H_S * W_S$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$F_I \in R^{C_I * H * W}$, 通过模态专用网络 $N_{specific}(\cdot)$ 和模态共享网络 $N_{shared}(\cdot)$计算

$$F_I = N_{\text{shared}}\left(N_{\text{spectfic}}(I)\right)$$

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。

$$\bar{F}_S = W_{\phi 2}\left(ReLU\left(W_{\phi 1}\left([F_I, F_S]\right)\right)\right)$$

其中$F_s \in R^{C_S * H * W}$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。$W_{\phi 1} \in R^{(C_I + C_S) * (C_I + C_S)}$和$W_{\phi 2} \in R^{C_I * (C_I + C_S)}$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$m_{t,j}$为：

$$m_{t,j} = \frac{e^{D(\bar{F}_{S,t}, \bar{F}_{S,j})}}{\sum_{\forall j} e^{D(\bar{F}_{S,t}, \bar{F}_{S,j})}}$$

$$D\left(\bar{F}_{S,t}, \bar{F}_{S,j}\right) = K\left(\bar{F}_{S,t}\right)^T \times Q\left(\bar{F}_{S,j}\right)$$

$D(\bar{F}_{S,t}, \bar{F}_{S,j})$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。

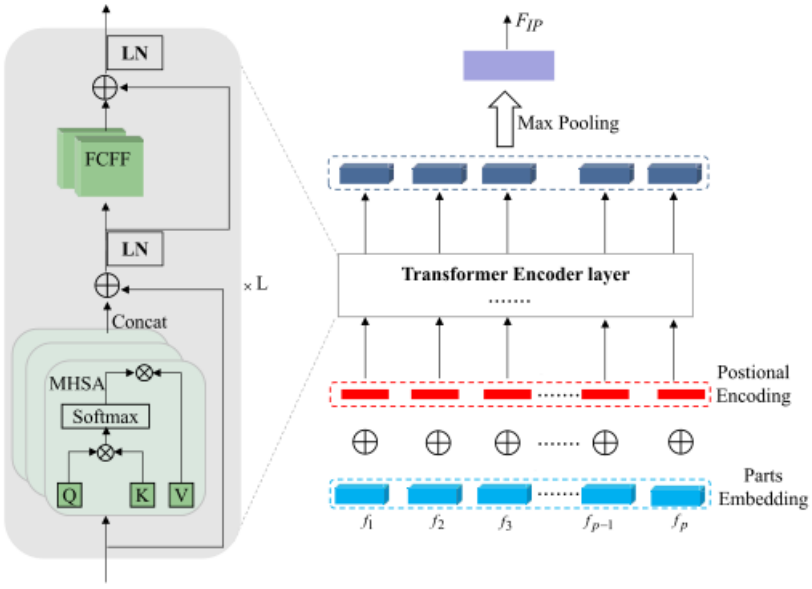
注意力定位区分区域，可用于指导外观特征学习：

$$\bar{F}_I = \sum_{t=1}^{n} m_{j,t} \times V\left(F_{I,t}\right)$$

$V(F_I, t)$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：

$$F_{IS} = \bar{F}_I + BN\left(F_I\right)$$

3.2 Transformer-Based Part Interaction



使用正弦和余弦函数来表示位置编码：

$$PE(pos, i) = \begin{cases} \sin\left(pos / Q^{2k/d_m}\right), & i = 2k \\ \cos\left(pos / Q^{2k/d_m}\right), & i = 2k+1 \end{cases}$$

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：$F_P = \{f_1, f_2 \ldots f_i \ldots f_p | f_i \in R^C, i = \{1,2,\ldots,p\}\}$（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。$M \in R^{p*H*W}$可以被表示：

$$M = W_\theta (F_S)$$

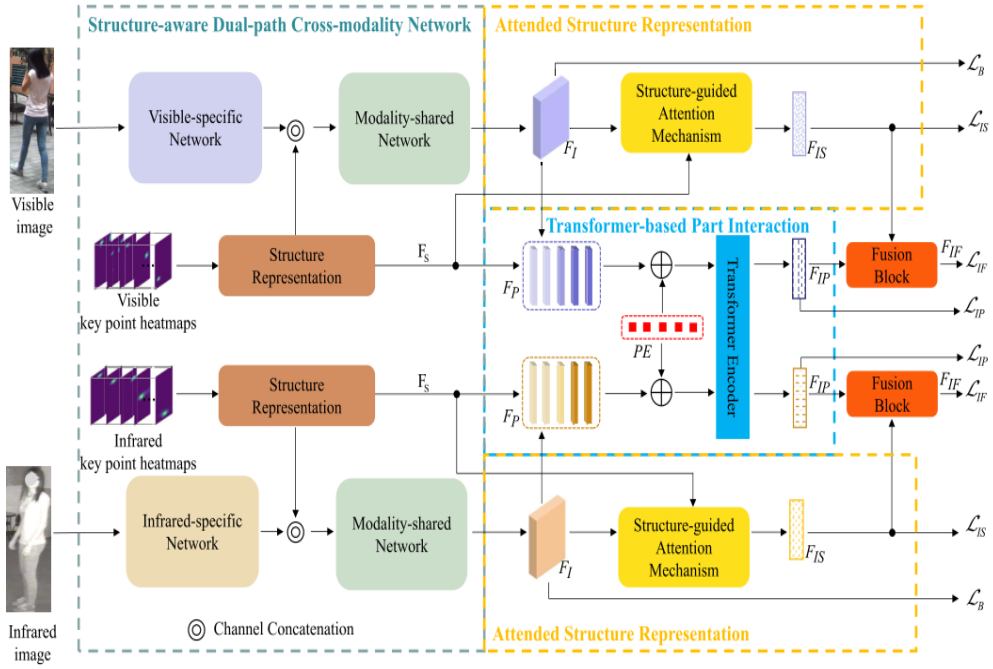
带有参数 W_θ 的 $\theta(\cdot)$是一个嵌入函数，它通过 1×1 卷积层和 Sigmoid 函数对结构特征进行编码，作为部件注意映射。然后，可以通过以下方式获得第 i 个零件特征 f_i：

$$f_i = \frac{1}{K^i} \sum_{h=1}^{H} \sum_{w=1}^{W} \bar{M}_{h,w}^i \otimes F_{I,h,w}$$

$$\bar{F}_P = \text{Transformer}\ (F_P + PE)$$

PE 表示部件序列的位置编码。

3.3Overall Architecture



$$\mathcal{L}_B = \mathcal{L}_{id} + \mathcal{L}_{tri}$$

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，$W_\gamma \in R^{C*2C}$。

$$F_{IF} = \mathrm{W}_\gamma\left([F_{IS}, F_{IP}]\right)$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_B + \lambda_1 \mathcal{L}_{IS} + \lambda_2 \mathcal{L}_{IP} + \mathcal{L}_{IF}$$

5. 论文：Learning Memory-Augmented Unidirectional Metricsfor Cross-modality Person Re-identification

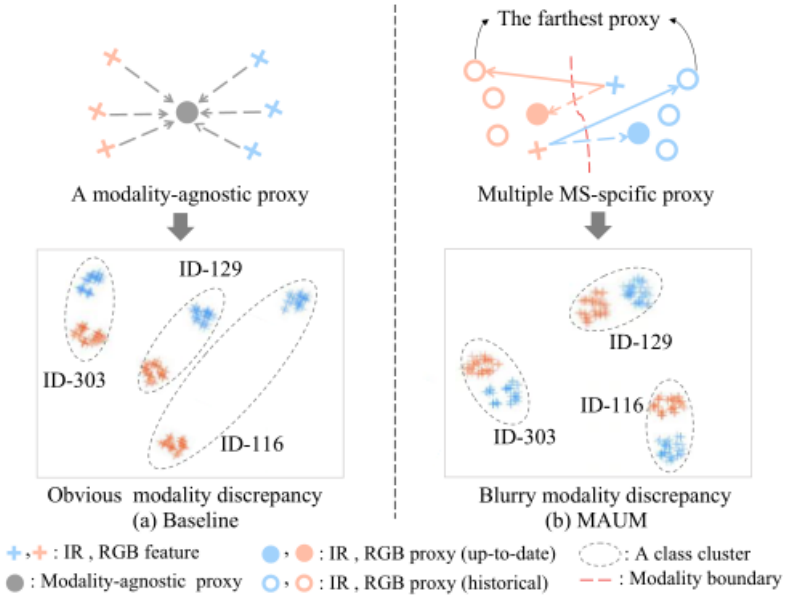
记忆增强单向度量在跨模态行人重识别中的应用

作者：Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, Wenhui Li

摘要：

本文通过抑制模态差异来解决跨模态行人重识别 (re-ID) 问题。在跨模态 re-ID 中，query 和 gallery 图像是不同的模态。给定一个训练 ID，流行的深度分类 baseline 是对两种模态共享相同的代理 (即最后一层的权重向量)。作者发现这样的做法对模态差异有相当大的容忍度，因为共享代理会作为两个模态之间的中间中继。为此，提出了一种记忆增强单向度量学习方法 (MAUM)，包括两种新的设计，即 单向度量 和 基于记忆的增强。具体来说，MAUM 首先在每个模态下独立学习特定模态代理 (MS-Proxies)，之后，MAUM 使用已经学习过的 MS-Proxies 作为静态引用，在对应的模态中关闭特征。这两个单向的指标 (IR 图像到 RGB 代理 以及 RGB 图像到 IR 代理) 共同缓解了中继效应，有利于跨模态联合。通过将 MS-Proxies 存储到 memory banks 以增加参考的多样性，进一步增强了跨模态关联。作者展示了 MAUM 在模态平衡情景下，改善了跨模态 re-ID 的效果，另外对于模态不平衡情景也具有很好的鲁棒性。

引言



(a) 在基线中，每个标识对于两个模态都有一个通用的代理，充当 IR 和 RGB 特征之间的中继。(b) MAUM 有两个特定于模式的代理(MS-Proxies，橙色实点表示 RGB，蓝色实点表示 IR)。每个 MS-Proxy 都是固定的静态引用，用于在对应的模态(虚线箭头)中拉近特性。此外，MAUM 将历史 MS-Proxies (空点)存储到两个内存库中，一个用于 IR 模态，一个用于 RGB 模态。相应地，每个标识都有多个 IR 和 RGB 代理。离模态边界最远的 MS-Proxies 成为硬正引用，因此具有更强的“拉近”效果(实箭头)。

2.主要贡献

1. 提出了一种新的记忆增强单向度量学习方法，用于跨模态 re-ID 问题。它在两个单向上学习显式的跨模态度量，并通过基于内存的增强进一步增强它们；
2. 考虑了模态不平衡问题，这是跨模态 re-ID 中一个重要的现实问题。通过调整特定模态的增益，MAUM 对模态不平衡问题表现出较强的鲁棒性；

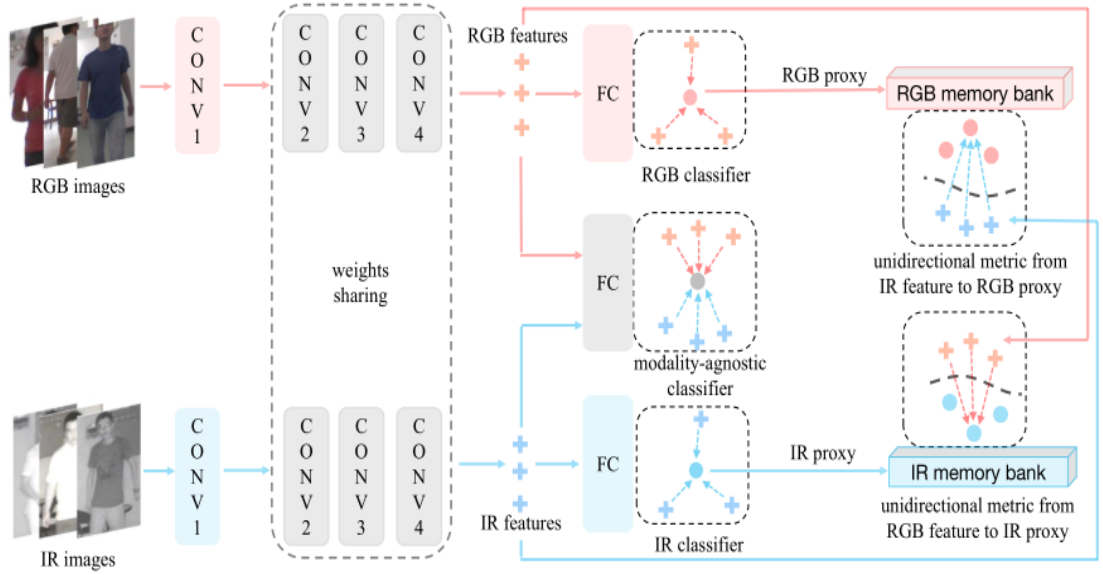
3.相关内容

数据不平衡是深度学习的一个重要挑战。以往的研究大多关注类别失衡问题，主要介绍了两种方法，即重采样和重加权。在训练中对少数类(样本少)或频繁类(样本多)进行过采样或过采样，目的是在每次迭代中平衡头尾数据。重加权为损失函数中的不同类甚至不同样本分配自适应权重。

在 MAUM 中，特定于模式的增强是自然分离的，并允许对特定的模式进行独立的增强。它使 MAUM 对模态不平衡具有较强的鲁棒性。

4.方法介绍

AUM 采用 ResNet50 作为 backbone，接受 RGB 和 IR 图像作为输入。MAUM 将第一个卷积块分成两个独立的分支，以适应特定模态的低级特征形式，一个用于 RGB，另一个用于 IR。为了计算效率，两种模态共享所有的卷积块。对于卷积特征映射，MAUM 使用全局平均池化 (GAP) 为每个输入图像生成深度嵌入。基于这种普遍采用的 backbone 设置，提出的 MAUM 着重于其新的记忆增强单向度量学习方法。



MAUM 采用 ResNet50 作为骨干，两种模式共用“conv2”到“conv4”的参数。将 RGB 和 IR 图像映射到深度嵌入空间中，分别得到 RGB 和 IR 特征。MAUM 有三个分 ID 类器，分别是 RGB 分类器，IR 分类器和通用的分类器。RGB (IR) 分类器只接受 RGB (IR)特征，以便学习到的 MS-Proxies 具有高度特异性，并减轻中继效应。给定已经学习过的 MS-Proxies， MAUM 在每次迭代后将它们存储到两个相应的内存库中。该记忆库具有三个关键功能，即单向度量学习、通过漂移增强和抵抗模态失衡。分类器采用交叉熵损失：

$$\mathcal{L}_{\mathrm{RGB}} = -\frac{1}{N^{\mathrm{R}}} \sum_{i=1}^{N^{\mathrm{R}}} \log \frac{\exp\left(w_{y_i}^{\mathrm{R}} x_i^{\mathrm{R}}\right)}{\sum_k^C \exp\left(w_k^{\mathrm{R}} x_i^{\mathrm{R}}\right)}$$

其中上标“R”表示 RGB 模态，N^R是当前小批量的 RGB 数量，C 是类别数量，

我们使用权重向量w_{y^i}作为 RGB 模态中y^i的代理。

在完全训练特定于模式的代理之后，MAUM 将它们收集到两个相应的内存库中。具体来说，我们使用队列策略来更新内存库。我们将 RGB 模式和 IR 模式的内存大小分别设置为 S_{RGB} 和 S_{IR}。在内存库达到其大小限制后，我们将最新的代理放入队列，并将最老的代理出队列。存储器对 MAUM 有三个关键功能。首先，他们冻结已经学习的 MS-Proxies，并使用它们作为单向度量学习的静态参考。其次，通过累积历史 MS-Proxies，利用模型漂移现象来增加 MS-Proxies 的多样性。第三，它们帮助 MAUM 获得针对模态失衡的额外鲁棒性，因为基于记忆的增强是特定于模态的，可以独立调整以重新平衡 IR 和 RGB 模态的增强。

在特定于模式的分类器中，每个标识只有一个 IR 和 RGB 代理，但将历史 MS-Proxies 存储到内存库中会逐渐增加其数量。因此，在 RGB (IR)内存库中，每个标识都有多个 RGB (IR)代理，为单个 IR (RGB)特性提供多个正引用。具体来说，对于单个 RGB 特征 x^R，我们假设 IR 记忆库中有 N 个正引用$\{u^I{}_1, u^I{}_2, \ldots u^I{}_N\}$和 M 个负引用$\{v^I{}_1, v^I{}_2, \ldots v^I{}_M\}$ (上标“I”表示 IR 模态)。RGB 图像到 IR 代理的单向度量的损失函数定义为:

$$\mathcal{L}_{\mathrm{R}\to\mathrm{I}} = \log\left[1 + \sum_{j=1}^{M}\sum_{i=1}^{N}\exp(\alpha(v_j^{\mathrm{I}}x^{\mathrm{R}} - u_i^{\mathrm{I}}x^{\mathrm{R}} + \delta))\right]$$

特征和代理是 l_2 normalized，α是比例因子，δ是边距参数，在实践中，损失函数是对当前小批处理中的所有 RGB 特性的平均值。IR 图像到 RGB 代理的单向度量的损失函数和 RGB 图像到 IR 代理类似。

6.论文：Cross-modal Local Shortest Path and Global Enhancement for Visible-Thermal Person Re-Identification

跨模态局部最短距离和全局增强行人重识别

作者：Xiaohong WangB, Chaoqi Li, and Xiangcai Ma

摘要：

本文的核心思想是使用局部特征对齐来解决遮挡问题，并通过增强全局特征来解决模态差异。首先，设计了基于注意的双流 ResNet 网络，提取双模态特征并映射到统一的特征空间。然后，为了解决跨模态人体姿态和遮挡问题，将图像水平切割成若干等分以获得局部特征，并使用两个图之间局部特征中的最短路径来实现细粒度局部特征对齐。第三，批处理归一化增强模块应用全局特征来增强策略，从而导致不同类之间的差异增强。多粒度损失融合策略进一步提高了算法的性能。最后，利用局部和全局特征的联合学习机制提高跨模态人员再识别的准确性。

1.引言

大多数工作只关注全局粗粒度特征提取过程，而忽略了提取后特征增强方法的研究，本文设计了一个基于批量归一化的特征增强模块，以增强提取后的全局特征，改善不同行人的特征表示。然而，仅关注全局特征是不够的，局部特征在 VI-ReID 任务中也发挥着重要作用。当人体部位因行人遮挡而丢失时，很难从这些图像中提取全局特征并真实描述此人，这很容易导致错误分类。

考虑到图像中行人的局部信息（例如头部、身体）可以很好地区分，有助于全局特征学习。因此，本文对两种不同模式下的行人图像在水平方向上平均分割，然后使用最短路径算法实现跨模态局部特征对齐。最后，基于局部和全局特征的联合学习机制可以有效地提高算法性能。

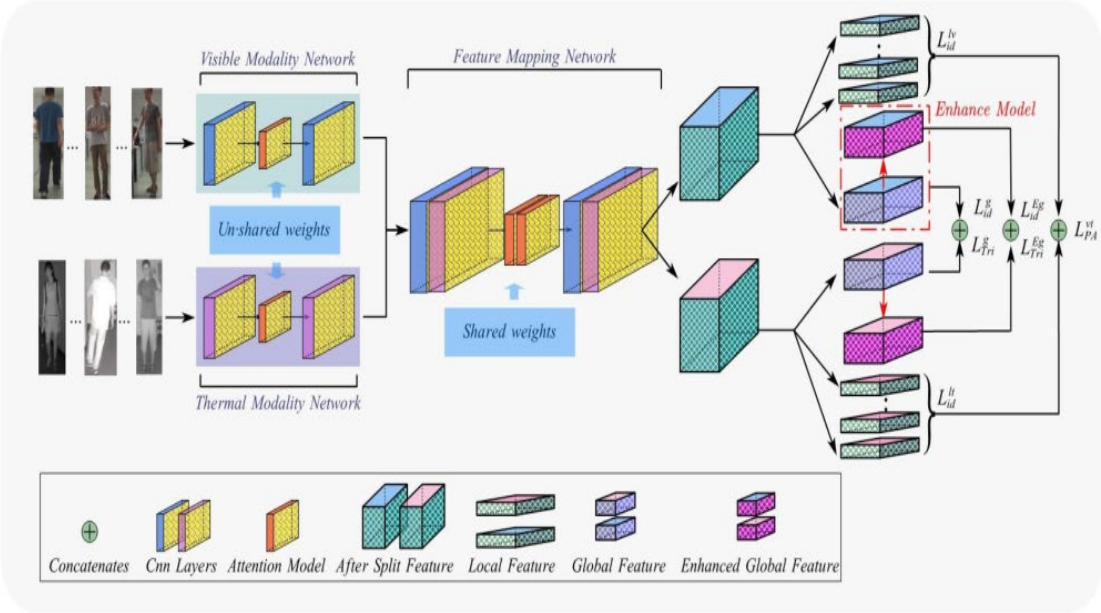
2.主要贡献

1. 我们提出了一种基于注意力的双流 ResNet 网络，用于 VT 跨模态特征获取
2. 提出了一种基于最短路径的跨模态局部特征对齐方法（CM-LSP），有效解决了跨模态行人再识别中的遮挡问题，提高了算法的鲁棒性
3. 设计了一种批量归一化全局特征增强(BN-GE)方法来解决全局特征识别不足

的问题，并提出了一种多粒度损失融合策略来指导网络学习。

3.方法介绍

3.1 网络模型



3. 2 局部特征对齐

将可见光和红外图像平均分成 i 块，定义局部特征表示为:

$$F_{rgb}^{loc} = \{f_r^{\,1}, f_r^{\,2}, \ldots f_r^{\,i}\}, F_{ir}^{loc} = \{f_t^{\,1}, f_t^{\,2}, \ldots f_t^{\,i}\}$$

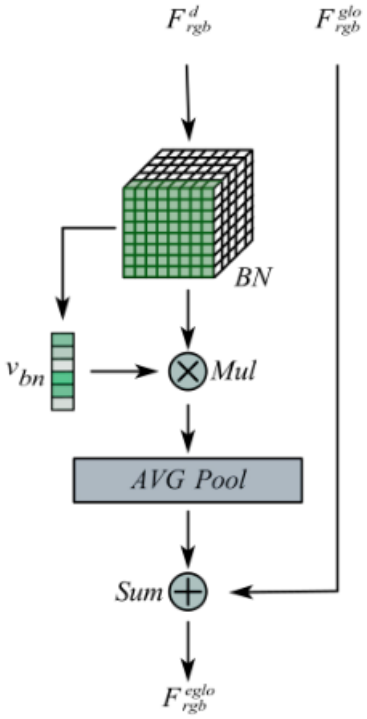
定义了计算两图之间距离的公式如下：

$$d_{i,j} = \left\| \frac{f_r^i - Mean(f_r^i)}{Max(f_r^i) - Min(f_r^i)} - \frac{f_t^j - Mean(f_t^j)}{Max(f_t^j) - Min(f_t^j)} \right\|_1$$

其中 i 和 j∈(1,2,3,…, h)分别是图像的各个部分。$d_{i,j}$为不同模态的局部特征之间的距离，定义$S_{i,j}$为两个图像局部特征之间的总距离，$S_{i,j}$公式如下：

$$S_{i,j} = \begin{cases} d_{i,j} & i = 1, j = 1 \\ S_{i,j-1} + d_{i,j} & i = 1, j \neq 1 \\ S_{i-1,j} + d_{i,j} & i \neq 1, j = 1 \\ \min(S_{i,j-1}, S_{i-1,j}) + d_{i,j} & i \neq 1, j \neq 1 \end{cases}$$

3.3 全局特性增强模块



3.4 多种损失函数

$$L^{g}_{id} = \sum_{i=1}^{N} -q_i \log(p^{g}_{i})$$

$$L^{lv}_{id} = \sum_{j=2}^{S}\sum_{i=1}^{N} -q_i \log(p^{j}_{i})$$

N 是训练数据集中类别的总数，q_i样本是真实的概率分布，S 是水平切片数，$p_i{}^{g}$和$p_i{}^{j}$是预测的样本概率分布

$$L^{g}_{Tri} = \sum_{i=1}^{P}[m_g + \left\|fc^{i}_{v} - fc^{i}_{t}\right\|_2 - \min_{k\in\{v,t\}}\left\|fc^{i}_{v} - fc^{j}_{k}\right\|_2]_{+}$$
$$+\sum_{i=1}^{P}[m_g + \left\|fc^{i}_{t} - fc^{i}_{v}\right\|_2 - \min_{k\in\{v,t\}}\left\|fc^{i}_{t} - fc^{j}_{k}\right\|_2]_{+}$$

对随机选取的 P 个人物身份中的每一个，随机抽取 K 张可见图像和 K 张热图像，共为 2 × P × K 图像，$fc_v{}^{i} = \frac{1}{K}\sum_{j=1}^{K} f_{v,j}{}^{i}$，$fc_v{}^{i}$表示第 i 个人的多张可见图像的特征平均，$fc_t{}^{i} = \frac{1}{K}\sum_{j=1}^{K} f_{t,j}{}^{i}$，$fc_t{}^{i}$表示第 i 个人的多张热图像的平均特征。

$$L_{PA}^{vt} = \sum_{i=1}^{P}\sum_{a=1}^{2K}\sum_{j=2}^{H}[m_l + \max_{k\in\{v,t\}}\left\|f_{i.j}^{ka} - f_{i,j}^{kp}\right\|_2 - \min_{k\in\{v,t\}}\left\|f_{i,j}^{ka} - f_{i,j}^{kn}\right\|_2]$$

$f_{i_j}{}^{ka}$是红外/可见光图像的局部特征，$f_{i_j}{}^{kp}$是正样本，$f_{i_j}{}^{kn}$是负样本

总的损失函数如下：

$$L_{total} = \overbrace{L_{id}^{g} + L_{id}^{eg}}^{Globalid} + \overbrace{L_{Tri}^{g} + L_{Tri}^{eg}}^{GlobalTri} + \overbrace{L_{id}^{lv} + L_{id}^{lt}}^{Localid} + \overbrace{L_{PA}^{vt}}^{LocalTri}$$

7.论文：Modality Synergy Complement Learning with Cascaded Aggregation for Visible-Infrared Person Re-Identification

基于级联聚合的模态协同互补学习的多模态行人重识别

作者：Yiyuan Zhang , Sanyuan Zhao, Yuhao Kang, and Jianbing Shen

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。

1.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。

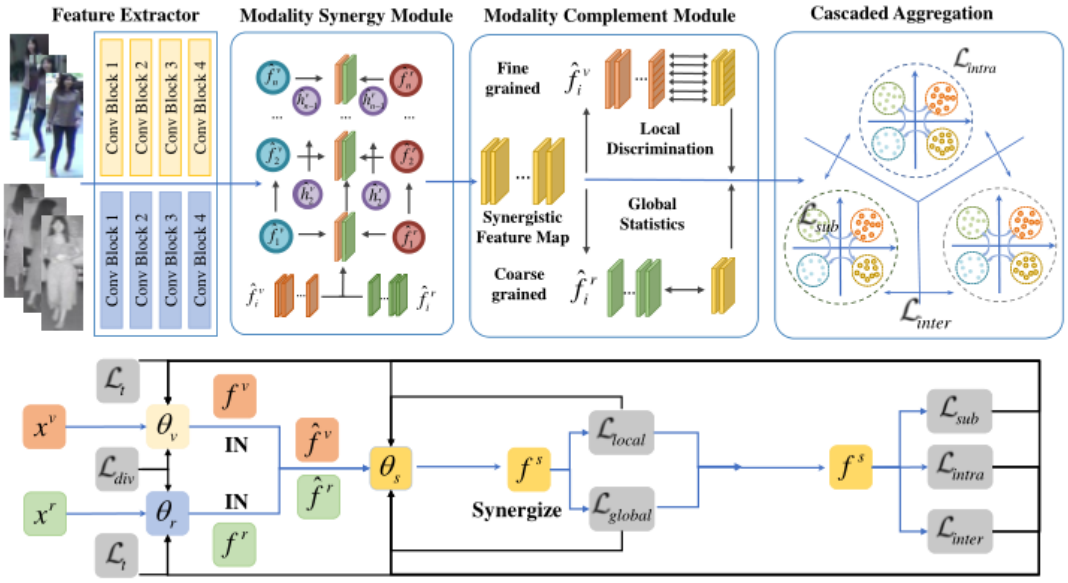
因此，我们提出了一个新的框架，即模式协同补充学习网络（MSCLNet）。它旨在减少类内的差异，提高身份差异的表现。首先，它通过与模态协同模块（MS）构建协同特征，保留了可见光和红外模态固有的语义多样性和身份相关性。然后，它通过两种模式的优势增强了协同特征表示。MS 和 MC 极大地提高了网络跨模态身份表示的能力。此外，我们提出了级联聚合策略（CA）来优化特征嵌入的分布。它逐步将样本聚合为子类、类内和不同身份。在级联方式中，属于相同身份的实例倾向于聚合，而属于不同身份的实例则映射为分散。

2.主要贡献

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。

3.方法介绍

3.1 网络模型



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。

3.2 模态协同模块（MS）

给定一对可见光和红外图像$x_i^v \in V$、$x_i^r \in R$，双流网络提取其特征f_i^v和f_i^r。特征f_i^v和f_i^r通过以下操作进行标准化：

$$\hat{f}_i^v = \frac{f_i^v - \mathrm{E}[f_i^v]}{\sqrt{\mathrm{Var}[f_i^v] + \epsilon^v}} \times \gamma + \beta, \mathrm{E}[f_i^v] = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H} f_{itlm}$$

$$\mathrm{Var}[f_i^v] = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H}(f_{itlm} - \mathrm{E}[f_i^v])^2$$

让 S(·)表示模态协同模块，以在f_i^v，f_i^r的基础上用标签 y_i 构建协同特征f_i^s：

$$f_i^s = \mathcal{S}(\hat{f}_i^v, \hat{f}_i^r, y_i, \theta_s)$$

其中 θ_s 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f_i^s，我们利用 KL-Divergence 来约束可见光和红外特征f_i^v，f_i^r的逻辑分布，其公式如下：

$$\mathcal{L}_{div} = -\mathrm{KL}(\hat{f}^v \parallel \hat{f}^r) = -\frac{1}{N}\sum_{i=1}^{N}(\hat{f}_i^v \cdot \log\frac{\hat{f}_i^v}{\hat{f}_i^r}, \theta_v, \theta_r)$$

引入交叉熵来约束可见光和红外特征：

$$\mathcal{L}_t = -\frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i \cdot \log\hat{p}_i^v(\hat{f}_i^v, \theta_v)] - \frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i \cdot \log\hat{p}_i^r(\hat{f}_i^r, \theta_r)]$$

$$\mathcal{L}_{Synergy} = \mathcal{L}(\theta_v, \theta_r) = \lambda_{div} \cdot \mathcal{L}_{div} + \lambda_t \cdot \mathcal{L}_t$$

3.3 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（f_i^v, f_i^r）。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f_i^s的表示有效性。考虑到细粒度语义，我们利用f_i^v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f_i^r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$f_i^v = [b_1^v, b_2^v \ldots b_n^v], f_i^s = [b_1^s, b_2^s \ldots b_n^s]$。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。

$$\mathcal{L}_{local} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}(cos(b_j^v, b_j^s) + \sqrt{2 - 2cos(b_j^v, b_j^s)})$$

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。

$$\mathcal{L}_{global} = \frac{1}{N}\sum_{i=1}^{N}||C_{y_i}^s - C_{y_i}^r||_2^2$$

其中C_{yi}^s, C_{yi}^r表示协同特征f_i^s，f_i^r的类中心。L_{global}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。

$$\mathcal{L}_{Com}(\theta_s) = \lambda_{local} \cdot \mathcal{L}_{local} + \lambda_{global} \cdot \mathcal{L}_{global}, \hat{\theta}_s = \arg\min_{\theta_s} \mathcal{L}(\theta_s)$$

3.4 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C_{si}表示该子类的中心：

$$\mathcal{L}_{sub} = \frac{1}{N}\sum_{i=1}^{N} ||f_i^s - C_{s_i}||_2^2$$

2. 类内聚合。聚合的公式可以表示如下，其中 N_s 表示每个标识的子类的数量：

$$\mathcal{L}_{intra} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_s} ||C_{s_j} - C_{y_i}||_2^2$$

3. 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：

$$\mathcal{L}_{inter} = -\frac{1}{\binom{N}{2}}\sum_{i=1}^{N}\sum_{j\neq i}^{N} ||C_{y_i} - C_{y_j}||_2^2$$

8. 论文：Channel Augmented Joint Learning for Visible-Infrared Recognition

行人重识别的通道增强联合学习

作者：Mang Ye, Weijian Ruan, Bo Du1, Mike Zheng Shou

摘要：

本文针对可见光红外识别问题，提出了一种强大的信道增强联合学习策略。对于数据增强，大多数现有方法直接采用为单模态可见光图像设计的标准操作，因此在可见光到红外匹配中没有充分考虑图像特性。我们的基本思想是通过随机交换颜色通道来均匀地生成与颜色无关的图像。它可以无缝地集成到现有的增强操作中，而无需修改网络，从而持续提高对颜色变化的鲁棒性。结合随机擦除策

略，通过模拟随机遮挡，进一步丰富了多样性。对于跨模态度量学习，我们设计了一种增强的通道混合学习策略，以同时处理具有平方差的跨模态和跨模态变化，从而获得更强的可分辨性。此外，还进一步提出了一种通道增强联合学习策略，以明确优化增强图像的输出。对两个可见-红外识别任务进行深入分析的大量实验表明，所提出的策略持续提高了识别精度。

1.引言

将红外图像与可见光图像进行匹配是一项重大挑战，会导致模态内和模态间的较大变化。为了消除颜色差异，使用生成对手网络（GANs）生成跨模态图像是一种流行的方法，可以在图像层面上缩小差距。然而，图像生成过程通常需要额外的计算成本，并且不可避免地受到噪声的影响。另一种方法是直接使用灰度图像来执行跨模态匹配，其中假设颜色信息不相关。虽然这种方法确实消除了颜色差异，但它也会丢失颜色通道中的辨别信息。

消除模态差异的直接解决方案是恢复原始的三个颜色通道。然而，将单通道红外图像转换为三通道可见图像是一个具有挑战性的问题，不可避免地会产生噪声。我们建议直接学习可见光图像的每个 R、G 和 B 通道与单通道红外图像之间的关系。这是可见-红外学习过程的通道增强操作，以增强对颜色变化的鲁棒性。我们进一步提出了一种用于遮挡模拟的随机擦除（CRE）技术。结合通道增强，我们的策略在信通道级别执行擦除，以获得更好的分集。此外，我们还包括用于增强的灰度变换，以减少颜色效果。这些增强操作极大地扩大了训练集，带来了更好的通用性。

2.主要贡献：

1. 提出了一种新的通道交换增强用于可见-红外识别，它可以无缝地集成到现有的扩展操作中，而不需要修改网络结构或改变学习策略。
2. 设计了一个增强的信通混合学习方案，以同时处理内模态和跨模态的变化。该算法采用联合学习策略，对通道增强图像进行了显式优化。

3.方法介绍

3.1 随机通道交换增强：

我们明确地学习匹配红外图像和可见图像的颜色通道。具体来说，我们通过挖掘每个通道(R, G 或 B)与单通道红外图像之间的关系，引入了一种通道增强策

略。其主要思想是随机选择一个通道(R、G 或 B)来替换其他通道，将注意力集中在一个通道上，生成一个新的训练图像。这被表述为：

$$\tilde{x}_i^{v,R} = (x_i^R, x_i^R, x_i^R)$$
$$\tilde{x}_i^{v,G} = (x_i^G, x_i^G, x_i^G)$$
$$\tilde{x}_i^{v,B} = (x_i^B, x_i^B, x_i^B)$$

可见光-红外人员再识别中的信道交换增强图

信道交换增强训练的实现简单，计算量小。它可以与其他基本数据扩充操作(随机翻转、随机调整大小和随机裁剪)无缝集成。我们使用单个数据加载器来执行随机通道扩展，这不会增加小批输入的大小。

3.2 通道随机擦除

它的基本思想是在预先确定的擦除概率下，在训练图像中随机选取一个矩形区域I_e，并将其像素值替换为所有三个通道的随机值，模拟不确定遮挡。通道随机擦除(CRE)策略可以丰富训练样本的多样性。

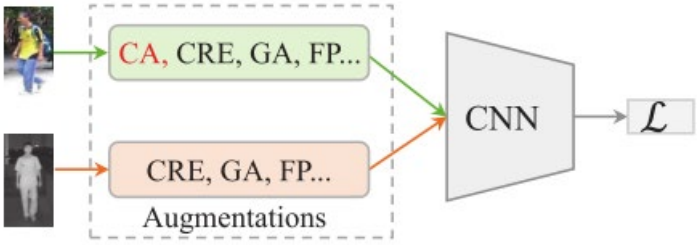
具体来说，假设一个 3 通道的可见训练图像的大小为 W ×H ×3。我们随机选取矩形区域的擦除面积S_e，该区域的擦除面积大小以特定比例为界。在通道增强的同时，我们为不同的通道(R, G 和 B)随机选择擦除区域。在每个通道所选的擦除区域S_e*中，S_e*中的每个像素都被分配给一个特定的预定义值$\alpha*$。

$$\tilde{x}_i^{v,*}(m,n) = \begin{cases} \alpha^*, & (m,n) \in S_e^* \\ \tilde{x}_i^{v,*}(m,n), & otherwise \end{cases}$$

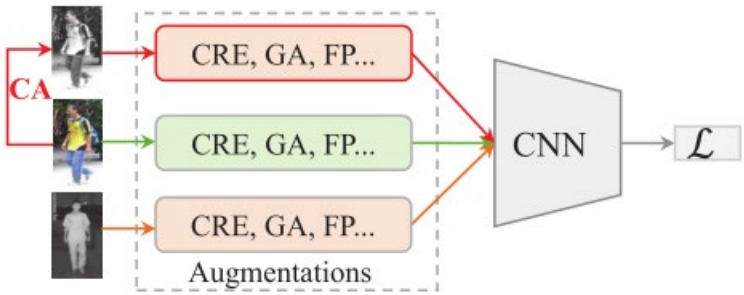
其中 m 和 n 表示像素的坐标位置，$\alpha*$由每个通道的平均值计算。对于单通

道红外图像，在通道随机擦除过程中，我们简单地将其变换为三个复制的单通道图像。

3.3 跨模态度量学习



通道增强混合学习



通道增强联合学习

CA:信道增强，CRE:通道随机擦除，GA:灰度增强，FP:水平翻转

一般的跨模态匹配模型通常采用双向三重损失（可见光到红外、红外到可见光）的变形来指导跨模态特征学习，优化跨模态正、负对之间的相对距离。这种策略的缺点是不能处理模态内的变化，为了同时处理模态内和跨模态的变化，本文提出了通道混合。

Lid 损失：

$$\mathcal{L}_{id} = -\frac{1}{N}\sum_{i=1}^{N}\log(p(y_i|f(x_i);\theta^0))$$

θ^0表示通道增强可见光图像和红外图像在不同数据增强操作下的共享标识分类器。

加权正则化三元组损失：

$$\mathcal{L}_{wrt} = \frac{1}{N}\sum_{i=1}^{N}\log(1+\exp(\sum_{ij} w_{ij}^{p} d_{ij}^{p} - \sum_{ik} w_{ik}^{n} d_{ik}^{n}))$$

$$w_{ij}^{p} = \frac{\exp\left(d_{ij}^{p}\right)}{\sum_{d_{ij}^{p}\in\mathcal{P}_i}\exp(d_{ij}^{p})}, w_{ik}^{n} = \frac{\exp\left(-d_{ik}^{n}\right)}{\sum_{d_{ik}^{n}\in\mathcal{N}_i}\exp(-d_{ik}^{n})}$$

对比难样本挖掘三元组损失

$$L_{\text{hard}} = \frac{1}{P*K}\sum_{A*batch}(maxd_{A,P} - mind_{A,N} + \alpha)$$

难样本挖掘三元组损失只考虑与最难正样本和最难负样本的距离，但是加权正则化三元组损失综合考虑所有的正样本和负样本。给所有的正样本和负样本一个权重，这个权重是根据他们与目标样本的距离计算。

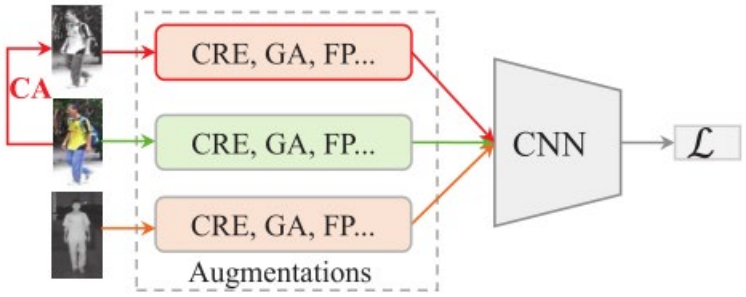
加权策略为 softmax 函数的加权策略，使用这种策略的好处是大大增加了距离较大（较小）的硬样本对正（负）值的贡献

增强平方差：

$$\mathcal{L}_{sq} = \frac{1}{N}\sum_{i=1}^{N}\log(1+\exp(\phi[\underbrace{\sum_{ij} w_{ij}^{p} d_{ij}^{p} - \sum_{ik} w_{ik}^{n} d_{ik}^{n}}_{\mu_i}]))$$

$$\phi[\mu_i] = \begin{cases} \mu_i^2, & \mu_i > 0, \\ -\mu_i^2, & \mu_i < 0. \end{cases}$$

通道增强联合学习



将通道增强可见光图像作为一种辅助模态，与原始的可见光和红外图像一起，制定了一个三模态联合学习框架。虽然通道增强图像作为一种附加模式，但它们与红外和可见光图像共享相同的身份分类器。这种策略可以使模型专注于学习不同的特征表示。

9.论文：Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification

Hi-CMD:可见光-红外行人重识别的分层交叉解纠缠

作者：Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, Changick Kim

摘要：

VI-ReID 存在由不同类型成像系统引起的额外的跨模态差异。为了减少通道内和通道间的差异，我们提出了一种分层交叉通道去纠缠（Hi-CMD）方法，该方法可以自动从可见热图像中分离 ID 鉴别因子和 ID 排除因子。我们只使用身份识别因子进行稳健的跨模态匹配，而不使用身份排除因子，如姿势或光照。为了实现我们的方法，我们引入了一个保持 ID 的人物图像生成网络和一个分层特征学习模块。我们的这一代网络通过生成一个新的具有不同姿势和照明的跨模态图像，同时保留一个人的身份，来学习非纠缠表示。同时，特征学习模块使我们的模型能够明确地提取可见红外图像之间的共同识别特征。

1. 主要贡献

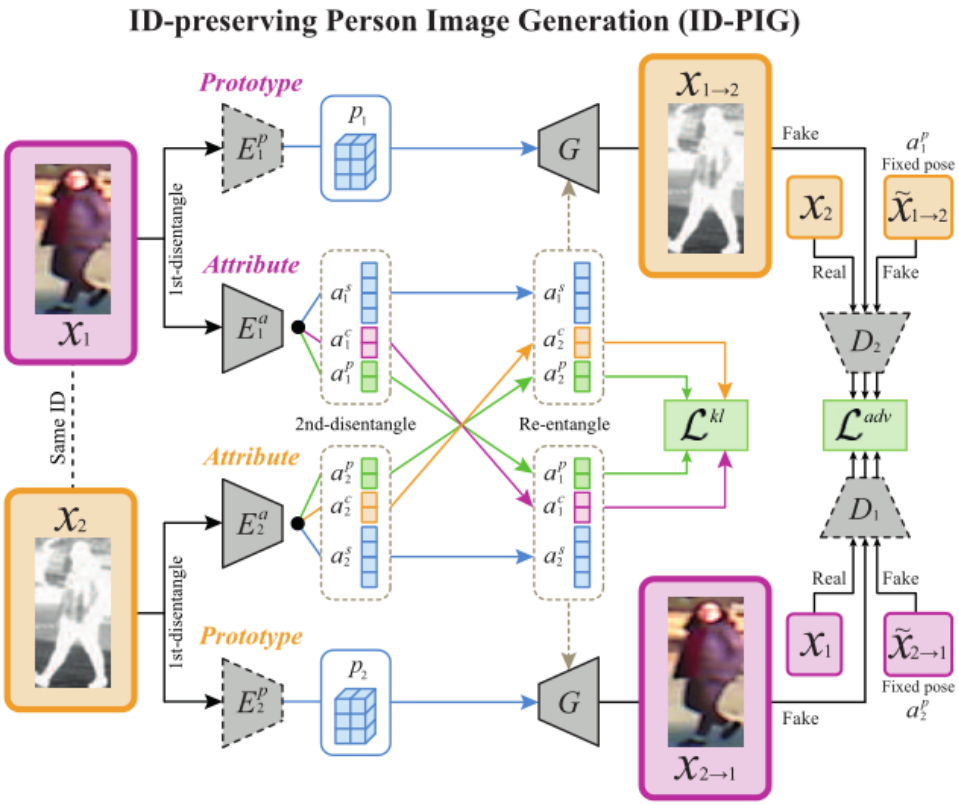
1. 提出了 Hierarchical Cross-Modality Disentanglement (Hi-CMD) 方法，它是一种高效的学习结构，可以提取姿势不变和光照不变的特征进行交叉模态匹配。这是第一个从 VI-ReID 的跨模态图像中同时分离 ID-discriminative factors 和 ID-excluded factors 的工作。
2. 提出 ID-preserving Person Image Generation(ID-PIG)网络，使改变姿势和照明属性同时保持特定人物的身份特征成为可能。

2.方法概述：

我们将可见光图像和红外图像分别表示为$x_1 \in R^{H\times W\times 3}$和$x_2 \in R^{H\times W\times 3}$，其中 H 和 W 分别为图像的高度和宽度。图像$x_1$和$x_2$中的每一个都对应一个标识标签 y∈{1,2，…， N}，其中 N 是人的身份数。在训练阶段，使用多模态图像集x_1和x_2在特征提取网络 φ(·)中训练。在测试阶段，给定一种模态的查询图像，计算另一种模态的图库集中的排序列表。两个特征向量 φ(x1)和 φ(x2)之间的距离由欧氏距离计算。

在 VI-ReID 任务中，最具挑战性的问题是可见光和红外图像之间同时存在跨模态和内模态差异。Hi-CMD 方法旨在从跨模态图像中分离 ID-discriminative factors 和 ID-excluded factors，同时减少跨模态和内模态的差异。提出了 ID-preserving Person Image Generation (ID-PIG)网络和 Hierarchical Feature Learning (HFL)模块

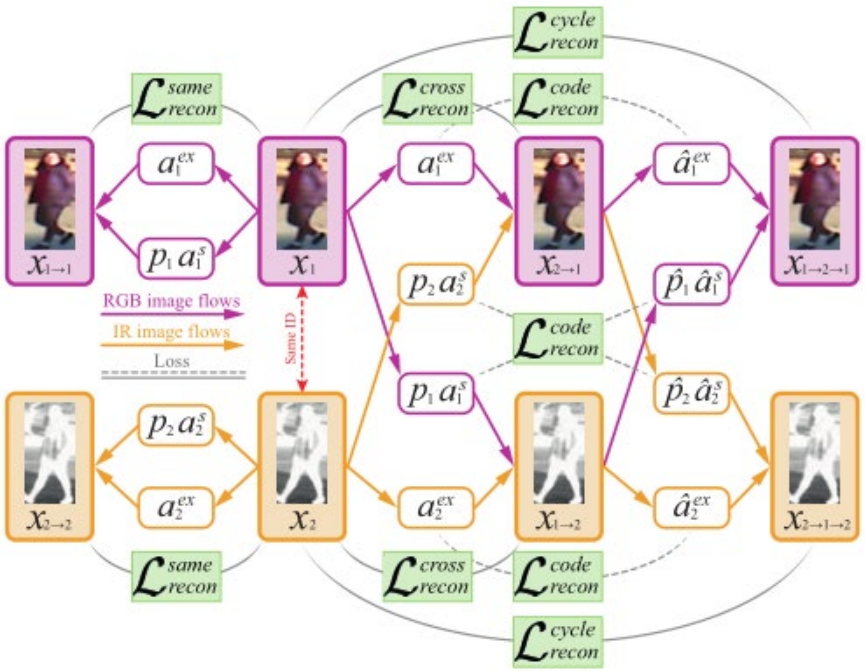
2.1ID-PIG 网络



ID-PIG 网络采用二阶层次化。第一层次采用原型编码器和属性编码器，原型编码器对图像中行人的体态等外貌特征进行编码，属性编码器对衣服风格、姿态、光照这类可变属性进行编码。第二层次属性编码划分成三类 $ai=[a_i^s;a_i^c;a_i^p]$，具体为：风格属性编码、光照属性编码、姿态属性编码。这三种编码分别表示：风格属性是对行人的衣服结构进行编码；光照属性编码对应模态之间的差异，把不同 RGB、IR 摄像头的视觉差异定义为光照属性；姿态属性对应模态内的差异，理解为同一个模态内行人的多种姿态。

最终光照、姿态属性作为 ID-excluded 编码，而风格属性、原型编码作为 ID-discriminative 编码。

2.2 损失函数说明



1. 跨模态重构损失函数:

在图像生成过程中，我们的主要策略是通过交换 ID 相同的两幅图像的 ID-excluded 来合成一对跨模态图像。保证原始图像的体态（原型编码）和衣服结构（风格属性），替换模态（光照属性）和姿态（姿态属性），重构生成的图像要与对应模态的样本图像尽可能接近。形式上，此跨模态重构损失表述如下:

$$\mathcal{L}_{recon1}^{cross} = \mathbb{E}_{\substack{\boldsymbol{x}_1 \sim p_{data}(\boldsymbol{x}_1), \\ \boldsymbol{x}_2 \sim p_{data}(\boldsymbol{x}_2)}} [\|\boldsymbol{x}_1 - G(\boldsymbol{p}_2, \boldsymbol{a}_2^s, \boldsymbol{a}_1^{ex})\|_1]$$

$$a_i^{ex} = [a_i^c; a_i^p]$$

2. 同模态重构损失函数：

除了对不同模态图像的重构损失外，我们还应用了对相同模态图像的重构损失。对同模态的四个编码，重新生成原始图像。

$$\mathcal{L}_{recon1}^{same} = \mathbb{E}_{\boldsymbol{x}_1 \sim p_{data}(\boldsymbol{x}_1)} [\|\boldsymbol{x}_1 - G(\boldsymbol{p}_1, \boldsymbol{a}_1^s, \boldsymbol{a}_1^{ex})\|_1]$$

3. cycle 重构损失函数：

两次跨模态重构

$$\mathcal{L}_{recon1}^{cycle} = \mathbb{E}_{\substack{\boldsymbol{x}_1 \sim p_{data}(\boldsymbol{x}_1), \\ \boldsymbol{x}_2 \sim p_{data}(\boldsymbol{x}_2)}} [\|\boldsymbol{x}_1 - G(\hat{\boldsymbol{p}}_1, \hat{\boldsymbol{a}}_1^s, \hat{\boldsymbol{a}}_1^{ex})\|_1]$$

4. 编码损失函数:

同模态的编码需要尽可能接近

$$\begin{aligned}\mathcal{L}_{recon1}^{code} =& \mathbb{E}_{\substack{\boldsymbol{x}_1\sim p_{data}(\boldsymbol{x}_1),\\ \boldsymbol{x}_2\sim p_{data}(\boldsymbol{x}_2)}}\left[\|\boldsymbol{a}_1^s-\hat{\boldsymbol{a}}_1^s\|_1\right]\\ &+\mathbb{E}_{\substack{\boldsymbol{x}_1\sim p_{data}(\boldsymbol{x}_1),\\ \boldsymbol{x}_2\sim p_{data}(\boldsymbol{x}_2)}}\left[\|\boldsymbol{a}_1^{ex}-\hat{\boldsymbol{a}}_1^{ex}\|_1\right]\end{aligned}$$

总的重建损失函数

$$\mathcal{L}^{recon}=\lambda_1\mathcal{L}_{recon}^{cross}+\lambda_2\mathcal{L}_{recon}^{same}+\lambda_3\mathcal{L}_{recon}^{cycle}+\lambda_4\mathcal{L}_{recon}^{code}$$

5. KL 散度损失

$$\mathcal{L}_1^{kl}=\mathbb{E}_{\boldsymbol{x}_1\sim p(\boldsymbol{x}_1)}\left[\mathcal{D}_{KL}(\boldsymbol{a}_1^{ex}\|N(0,1))\right]$$

6. 对抗损失

$$\begin{aligned}\mathcal{L}_1^{adv} =& \mathbb{E}_{\substack{\boldsymbol{x}_1\sim p_{data}(\boldsymbol{x}_1),\\ \boldsymbol{x}_2\sim p_{data}(\boldsymbol{x}_2)}}[\log(1-D_1(G(\boldsymbol{p}_2,\boldsymbol{a}_2^s,\boldsymbol{a}_1^c,\boldsymbol{a}_1^p)))]\\ &+\mathbb{E}_{\substack{\boldsymbol{x}_1\sim p_{data}(\boldsymbol{x}_1),\\ \boldsymbol{x}_2\sim p_{data}(\boldsymbol{x}_2)}}[\log(1-D_1(G(\boldsymbol{p}_2,\boldsymbol{a}_2^s,\boldsymbol{a}_1^c,\boldsymbol{a}_2^p)))]\\ &+\mathbb{E}_{\boldsymbol{x}_1\sim p(\boldsymbol{x}_1)}\left[\log D_1(\boldsymbol{x}_1)\right],\end{aligned}$$

将 ID-discriminative 编码加权级联得到判别向量，再传入全连接层，得到最终的特征向量。ReID 损失函数包含交叉熵损失和三元组损失，训练数据采用训练得到的生成器进行数据增强，生成相同 ID 但不同姿态、光照属性的行人图片