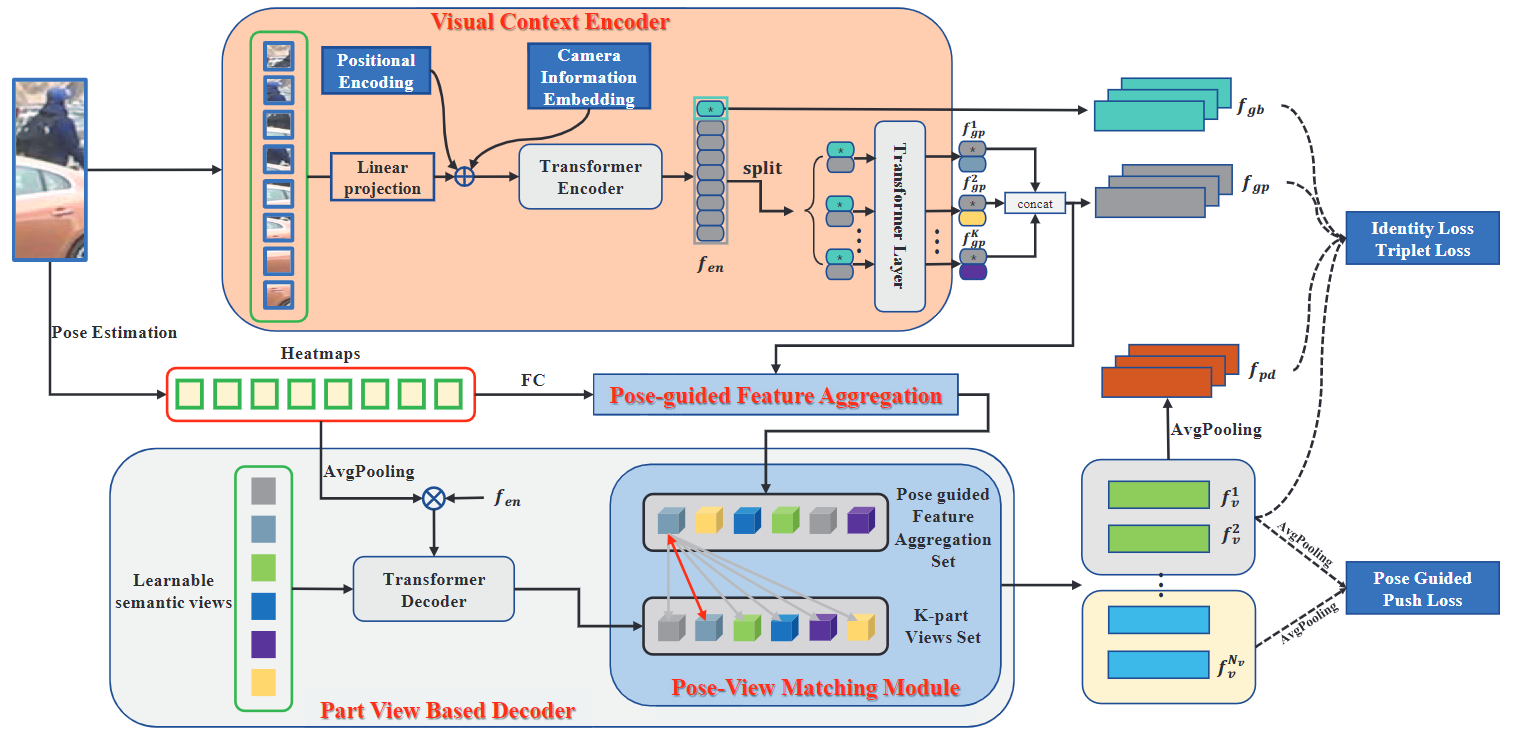


AAAI 2022



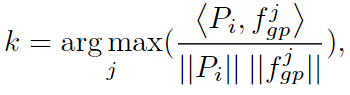
PFD输入：

（λcm=3.0，camera的设置和transreid一致）

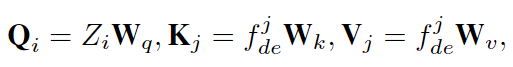
1、Visual context encoder:为了学习更多人体区别性的特征，将经过transformer encoder得到的fpart按顺序分成K组，每个group的大小为(N//K) × d，然后将encoder全局特征fgb（1×D）和每个group concat后送入共享transformer层，学习K组part local feature：fgp 。

Pose estimation:用pretrained hrnet v1对一个batch里的图像x，从输入图像中提取M个landmark。然后利用这些landmark生成热图H (M,H/4,W/4)，每个热图下采样到(H/4) × (W/4)。每个热图(H/4,W/4)的置信度值最大的点对应一个连接点。选择阈值γ（消融分析得到0.2）来过滤出高置信度和低置信度的landmark，每个landmark对应热图将分配0和1的二值标签。

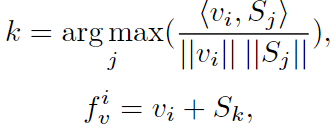
2、Pose guided feature aggregation:为了对姿态信息进行整合，将热图经过卷积后按元素乘fgp，P = [P1, P2，…,PM]。P已经编码了不同身体部位的特征，为了从fgp中找到对身体某一部位贡献最大的part feature。对于每个Pi，在fgp中找到最相似的特征，然后相加。



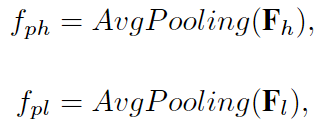
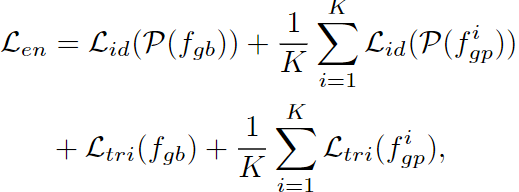
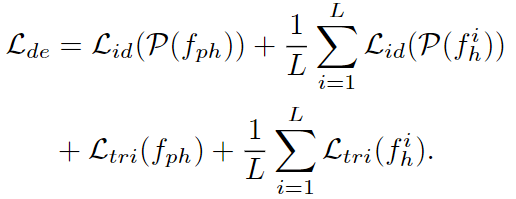


3、Part View Based Transformer Decoder：Nv（消融分析得到17为最优，等于K）个learnable semantic views作为query，heatmaps经过avgpool和fen点乘后，作为key和value，decoder的层数经过实验选择了2.

4、Pose view matching module:transformer decoder更新得到的views和PFA得到的特征根据相似度匹配，最后根据vi对应的Sk找到heatmap对应标签，所有标签为1的特征fvi作为一个集合，剩下的Nv-L个view作为FL



损失函数：λen，λde=0.5

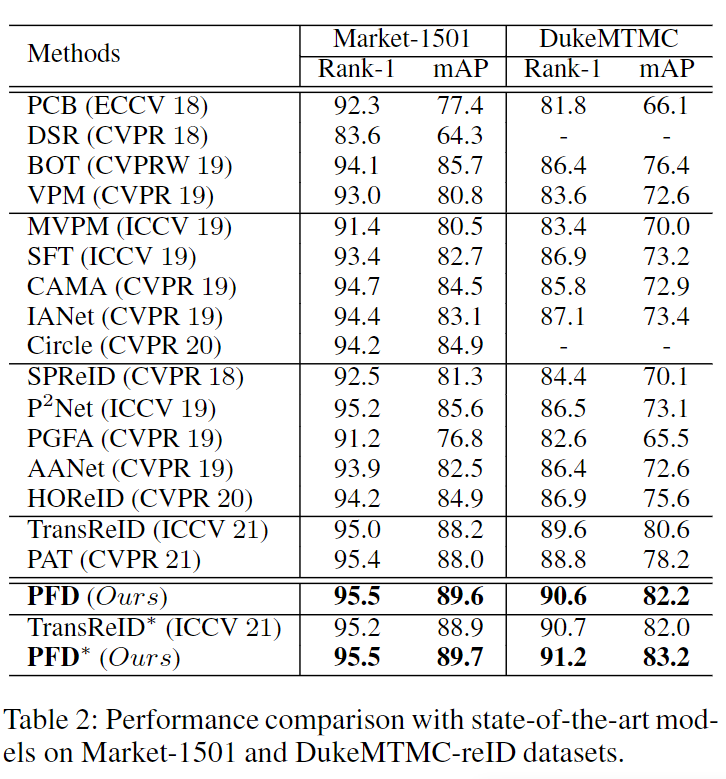
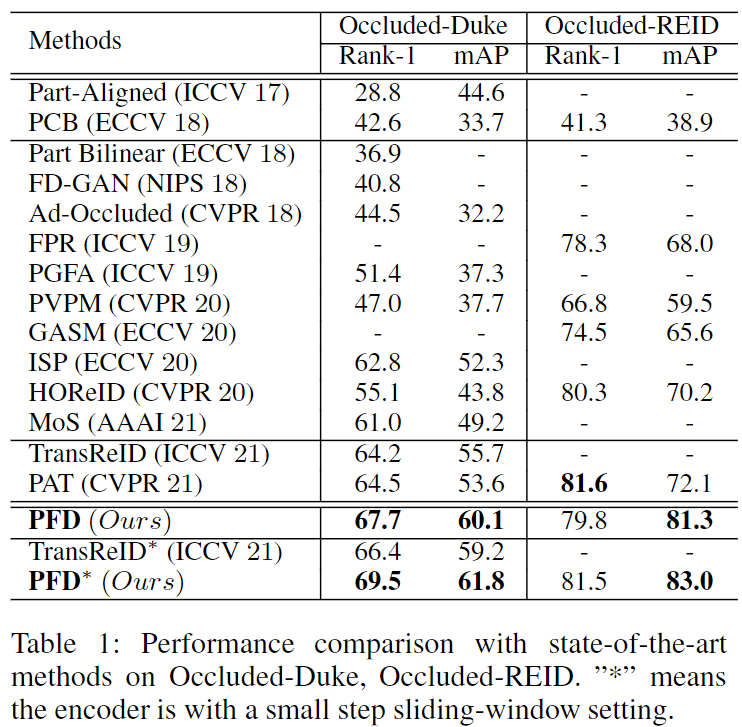
 

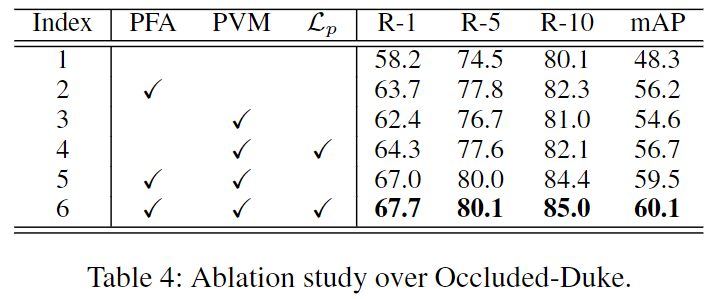
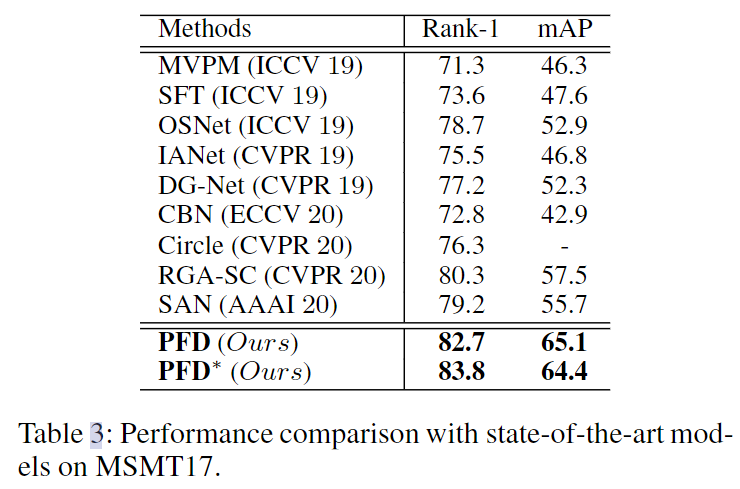


pose guided push loss Lp是想突出人体可见部位的特征，将可见特征和不可见特征尽量分开。

Inference：

结果：





问题：

1. learnable semantic views如何初始化？

0初始化，但是用了PE。PE的初始化是N(0,1)采样的，用nn.embedding作为可学习的参数，可以使不同query学习尽量不同的特征。

1. 为什么在PFA模块和PVM模块都要用集合匹配的方式，选择最相似的part feature相加

在PFA模块P编码了人体不同部位的信息的基础上，从fgp中找到对身体某一部位贡献最大的特征补充了损失的信息。PVM模块是view不知道学习的哪部分特征，所以将带有位姿的特征和视图匹配。

1. Heatmap的标签如何确定为1，heatmap经过avgpool如何和fen进行dot product

Heatmap(H/4,W/4)中的最大值对应的点是landmark，最大值就是landmark对应置信度，根据置信度和阈值的关系给heatmap分配0和1的标签。而一组heatmap的size是(M,H/4,W/4)，resize为（M,H/4\*W/4）经过avgpool得到(1,D)，和fen(N+1,D)点乘。

4、为什么heatmap和group token按元素相乘，那个token如果是管一条条区域的信息的确不太靠谱，硬要解释的话，就只有identityloss训练的时候让这个group token也带了全局的信息，但是后面又根据余弦相似度对当前的pose-guided feature加入了对应的group token