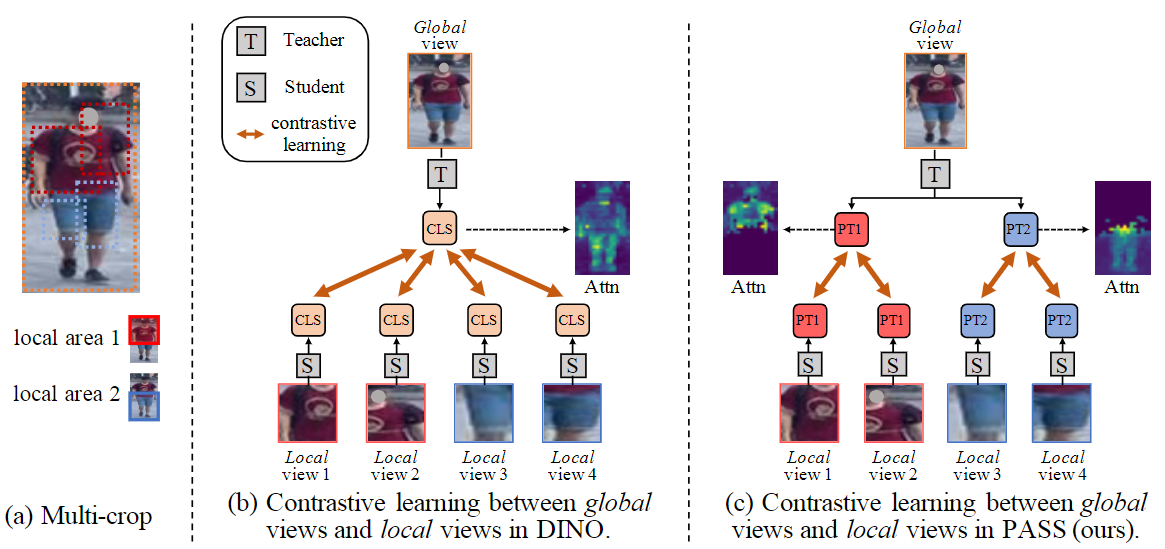
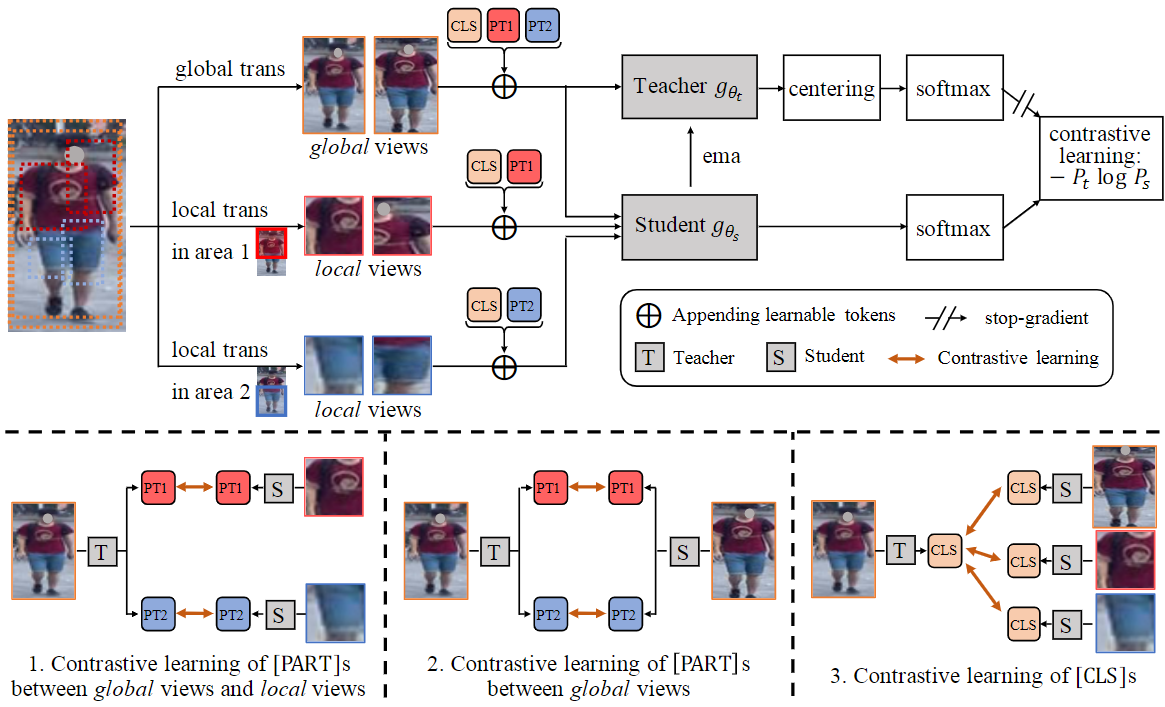
**Part-Aware Self-Supervised Pre-Training for Person Re-Identification**

<https://arxiv.org/pdf/2203.03931.pdf> （ECCV2022）

由于多数预训练方法在imagenet pretrain，而不是只在人的图像上pretrain，所以效果有限，或者在人上pretrain后，但没有很好的结合part feature，就提出了结合两者的这个预训练方法PASS。

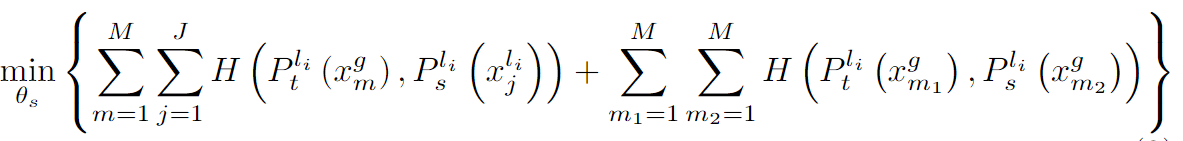


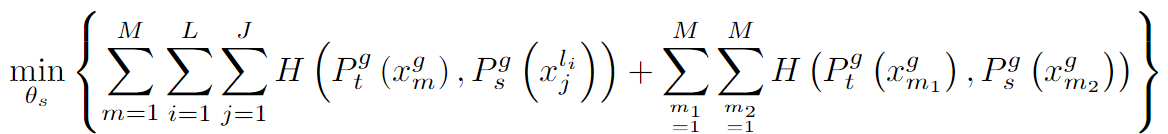
论文采取的Multi-crop，是对一个area随机裁剪不同大小的view，resize为相同大小输入网络，相当于不同分辨率。由于DINO学习匹配所有局部视图和全局视图的输出，难以同时学习到红衣服和蓝裤子特征，丢失很多细节。就提出了PASS，从每个局部区域裁剪的local view用特定的[PT]分配，全局视图后面接所有的[PT]。PASS分别匹配局部视图和全局视图作为输入的相同[PT]上的预测。



在student network中，分配给每个[PART]的局部视图是从特定的局部区域裁剪出来的，因此[PART]可以关注不同的区域。PASS使用student network来更新teacher network，这可以保证teacher network中的[PART]s也关注不同的局部区域，并学习到细粒度的信息。并且由于使用了蒸馏，所以student network学到了更鲁棒的特征。

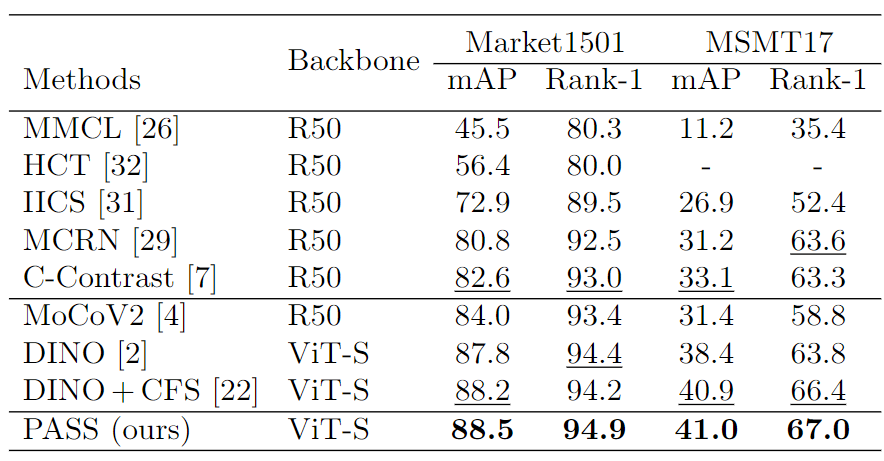
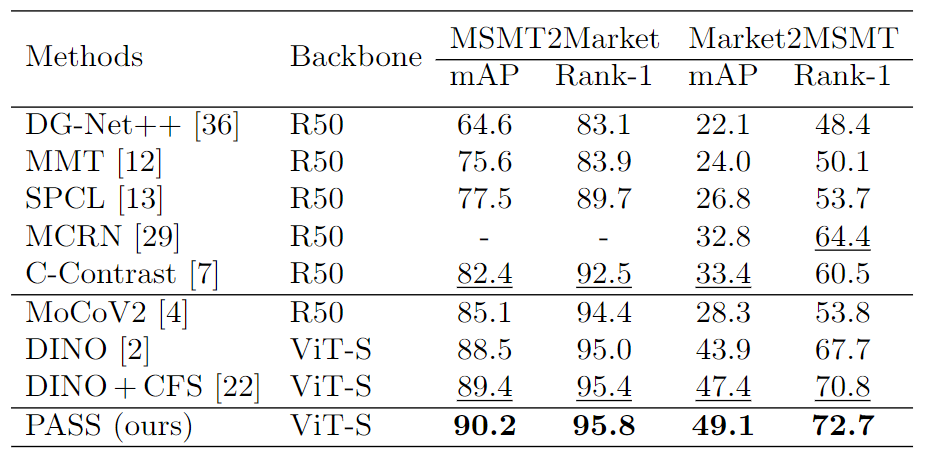
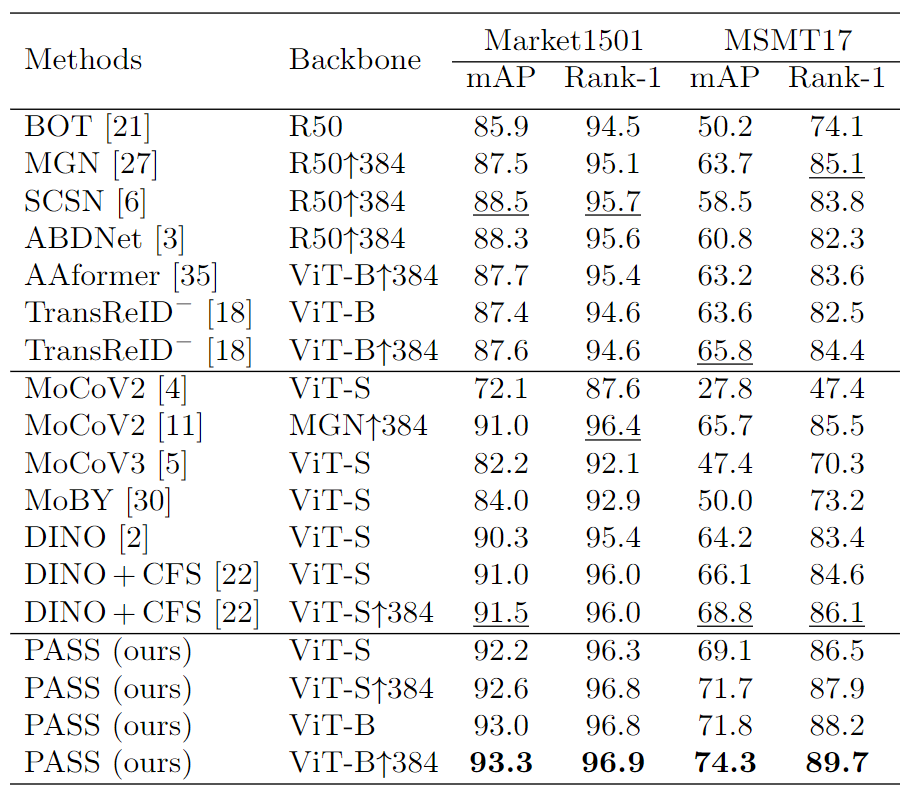
总体流程：global trans从整个图像中裁剪出global view，两种local area分别从两个不同的局部区域裁剪出局部视图。来自每个局部区域的local view被分配一个特定的[PART]。Global view之后接入所有局部区域的[PART]。所有view接了[CLS]。所有的view都输入给了student，而输入给teacher的只有global view。每个网络用了vit结构在LUPperson数据集上通过卷积在[CLS]和[PART]s上预测成K维的特征并进行softmax。然后，用交叉熵损失测量student和teacher输出的相同[PART]/[CLS]之间的相似性。梯度只更新student，teacher参数用student参数的指数移动平均(ema)来更新。

Part上的损失：

Cls上的损失：

消融：选择划分为3个local view，在supervised reid（transreid的baseline，使用了triplet loss和cross entropy loss），UDA，USL上都选择了代表图片

结果：微调阶段，使用teacher network。第一个表是supervised learning，第二个表是unsupervised domain adaptation,第三张表是USL。三张表的第一组都是imagenet pretrained好的，第二组是在LUPerson数据集上用自监督学习pretrained，第三组是pass预训练好的。



问题：

1. 为什么论文里说“一个局部视图占整个图像的比例高达40%，因此当L = 2时，每个局部区域应占图像的70%(与原始图像宽度相同，高度70%)，以保证局部视图可以从图像中的任何地方裁剪出来。”

因为PASS预训练阶段global view大小为256 × 128，local view大小为96 × 48，高占比将近40%。当L=2的两块局部区域每个高占70%，中间会交叉一个40%，正好能保证30%-70%任何地方都可以取到。而如果取了50%，没有交叉区域，那就无法取到跨越了50%的边界的局部区域。同样的道理，当L = 3时，每个局部区域应该占据图像的50%。

1. vit的每个image patch sequence前有个[cls]和这边论文在patch embedding后提到加入[cls]是一致的吗？

是一致的，都是1X1XC的可学习的tokens，pass用的[cls]实际上代码实现也是在patch embedding前面，只是[cls]后面多concat了part tokens，并且part tokens也有可学习的position embedding。

1. Centering模块的作用？

避免模型坍塌，centering模块就是计算出teacher输出特征的中心特征，并通过动量移动的方式在每次训练中更新。