Clothes-Changing Person Re-identification with RGB Modality Only （CVPR2022）论文，代码

1. Introduction

行人重识别旨在从跨定位和时间的监控视频中搜索到目标人物。目前的许多工作基于行人在短时间内不会更换衣物的假设，而在长时间跨度中换装问题不可避免，且在真实场景中目标人物通常会通过更换衣物来避免被识别和追踪。换装ReID问题的关键是提取那些衣物不相关的特征，如人脸、发型、身体形状、步态等等。

(a) (b) (c)

Figure 1. The visualization of (a) two original images, (b) the learned feature maps only with identification loss, and (c) the learned feature maps with identification loss and the proposed CAL. Note that all training settings of (b) and (c) are consistent except loss functions. (b) only highlights face as the clothes-irrelevant features, while (c) highlights more clothes-irrelevant features, *e.g.*, face, hairstyle, and body shape. (Since different samples of the same person in the training set mostly wear the same shoes, shoes are also highlighted.)

1. 任务：仅使用RGB模态的图像来解决换装（Clothes-Changing）场景下的Person ReID问题，并通过实验证明其有效性。
2. 现有问题

- 为了避免衣物的干扰，一些换装ReID方法会从多模态输入中（如骨架、轮廓等）对身体形状和步态进行建模。然而，基于多模态的方法需要额外的模型去捕捉多模态信息，且学习特征的解耦表示通常十分耗时。
- 原始RGB模态的图像中包含丰富的衣物不相关信息，而这些信息通常是尚未被现有方法利用到的。对于一些换装ReID方法，没有一个合适的损失函数设计的支撑的话，即使使用了较强的backbone去从原始图像中提取特征，得到的特征图也仅仅会注意到一些简单的衣物不相关的信息（比如人脸），而其他重要的衣物不相关的信息会被忽略。
- 目前多数的换装ReID工作仅关注于基于图片的设定，query和gallery样本都是图片，但现实ReID场景中，很多query和gallery集合都包含大量的视频，这比图片数据有着更丰富的外貌信息和额外的时间序列信息，因此试图从视频中学习到合适的时空模式（如步态）是很有前景的。而目前没有相关的可用公开数据集。

3. 主要贡献

- 提出了基于衣物的对抗损失（CAL），通过惩罚模型在衣物上的表达能力来更好地挖掘RGB模态中的衣物不相关信息。
- 基于一个步态识别的数据集FVG重建了一个新的换装视频ReID数据集，并提供了细粒度的衣物标签。

2. Method

2.1 Framework and Notation

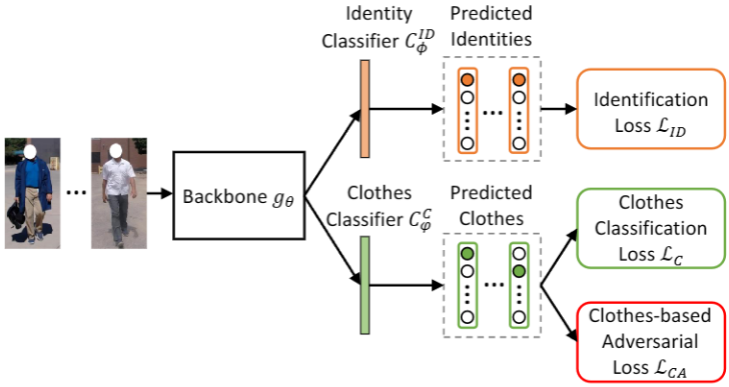


Figure 2. The framework of the proposed method. In each iteration, we first optimize the clothes classifier by minimizing $\mathcal{L}_C$. Then, we fix the parameters of the clothes classifier and minimize $\mathcal{L}_{ID}$ and $\mathcal{L}_{CA}$ to force the backbone to learn clothes-irrelevant features.

作者用$g_\theta(\cdot)$来表示拥有参数θ的backbone，用$C_\phi^{ID}(\cdot)$表示拥有参数ϕ的身份分类器，用$C_\varphi^C(\cdot)$表示拥有参数φ的衣物分类器。对于一个样本x_i，它的类别标签表示为y_i^{ID}，衣物标签表示为y_i^C。作者将衣物类别定义为身份类别的一种更细粒度的表示，同一身份的所有样本根据他们的衣物被划分为属于该身份类别的不同衣物类别。

已有的ReID方法将身份损失$\mathcal{L}_{ID}$定义为预测身份$C_\phi^{ID}(g_\theta(x_i))$和身份标签$y_i^{ID}$之间的交叉熵，并通过最小化$\mathcal{L}_{ID}$来训练模型。除了身份分类器外，本论文的框架使用了衣物分类损失$\mathcal{L}_C$来训练一个额外的衣物分类器。而论文中提出的基于衣物的对抗损失（CAL）$\mathcal{L}_{CA}$被用于驱使backbone解耦衣物不相关的特征。

2.2 Clothes-based Adversarial Loss

框架的每一轮训练包含两步优化操作，训练衣物分类器和学习衣物不相关的特征。

- Training clothes classifier

作者定义衣物分类损失$\mathcal{L}_C$为预测衣物$C_\varphi^C(g_\theta(x_i))$和衣物标签$y_i^C$之间的交叉熵，$f_i$为$g_\theta(x_i)$经过$l_2$归一化之后的结果，$\varphi_j$为第$j$个衣物分类器的权重经过$l_2$归一化之后的结果，$N$为batch size，$N_C$为训练集中衣物的类别数量，$\tau$是温度系数。则$\mathcal{L}_C$可以表示为如下公式：

$$\mathcal{L}_C = -\sum_{i=1}^{N} log\frac{e^{(f_i \cdot \varphi_{y_i^C}/\tau)}}{\sum_{i=1}^{N_C} e^{(f_i \cdot \varphi_j/\tau)}}$$

在第一步中，作者提出通过最小化损失$\mathcal{L}_C$来优化衣物分类器。其过程可以公式化为：

$$\min_{\varphi} \mathcal{L}_C(C_\varphi^C(g_\theta(x_i)), y_i^C)$$

- Learning clothes-irrelevant features

在第二步中，作者将衣物分类器的参数固定，并驱使backbone学习衣物不相关的特征，因此需要惩罚模型在衣物上的预测能力。作者希望训练后的衣物分类器不能分辨出有着相同身份但不同衣物的样本，因此基于衣物的对抗损失$\mathcal{L}_{CA}$应该是一个拥有多个正类别的分类损失，也即属于同一身份的所有衣物类别都互为正类。对于一个样本x_i，属于其身份类别y_i^{ID}的所有衣物类别都被定义为正衣物类别。$\mathcal{L}_{CA}$可以表示为如下公式：

$$\mathcal{L}_{CA} = -\sum_{i=1}^{N}\sum_{c=1}^{N_C} q(c) log \frac{e^{(f_i \cdot \varphi_c/\tau)}}{e^{(f_i \cdot \varphi_c/\tau)} + \sum_{j \in S_i^-} e^{(f_i \cdot \varphi_j/\tau)}}$$

$$q(c) = \begin{cases} \frac{1}{K}, & c \in S_i^+ \\ 0, & c \in S_i^- \end{cases}$$

其中S_i^+是和f_i拥有相同身份的衣物类别，S_i^-是和f_i拥有不同身份的衣物类别。K是S_i^+中的类别数量，$q(c)$是第c个衣物类别的交叉熵损失的权重。此时有着相同衣物类别的正类（$c = y_i^C$）和不同衣物类别的正类（$c \neq y_i^C and\, c \in S_i^+$）都有相同的权重$\frac{1}{K}$，而负类的权重为0。这样使得模型不去区分一个人穿什么衣服，而只关注身份本身。

在长期的ReID系统中，衣物不变ReID和换装ReID同样重要，但优化以上公式时会造成换装ReID的精度提升，而衣物不变ReID的精度可能会有所降低。因此可以对该公式做调整如下：

$$q(c) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{K}, & c = y_i^C \\ \frac{\epsilon}{K}, & c \neq y_i^C and\, c \in S_i^+ \\ 0, & c \in S_i^- \end{cases}$$

平滑后的公式增大了同一身份、相同衣物的权重，减少了同一身份、不同衣物的权重，这就使得模型在关注身份的同时也关注部分衣物特征。

最后是联合优化$\mathcal{L}_{ID}$和$\mathcal{L}_{CA}$，即：

$$\min_{\theta,\phi} \mathcal{L}_{ID}(C_\phi^{ID}(g_\theta(x_i)), y_i^{ID}) + \mathcal{L}_{CA}(C_\varphi^{C}(g_\theta(x_i)), y_i^{C})$$

如果仅使用$\mathcal{L}_{ID}$进行训练，模型会在训练初期从简单样本（同一身份、相同衣物）中学习，并逐渐学习分辨困难样本（同一身份、不同衣物）。即使$\mathcal{L}_{ID}$和$\mathcal{L}_{CA}$的目标都是将拥有相同身份的样本拉近，但作者仍未在训练中丢弃$\mathcal{L}_{ID}$，因为仅仅优化$\mathcal{L}_{CA}$会强制模型在训练初期学习分辨困难样本，从而陷入局部最优。为此，作者在学习率第一次衰减的时候才加入$\mathcal{L}_{CA}$进行联合优化，让模型在前期更趋于把同一衣物的身份归为一类，在后期对不同衣物的同一身份也具有识别功能。

---

3. CCVID Dataset

Table 1. The statistics of our CCVID dataset and other video person re-id and clothes-changing person re-id datasets.

datasets	#identities	#sequences	#bboxes	changing clothes?
PRID	200	400	40,033	✗
iLIDS-VID	300	600	42,460	✗
MARS	1,261	19,608	1,191,003	✗
LS-VID	3,772	14,943	2,982,685	✗
Real28	28	-	4,324	✓
VC-Clothes	512	-	19,060	✓
LTCC	152	-	17,119	✓
PRCC	221	-	33,698	✓
Celeb-reID	1,052	-	34,186	✓
DeepChange	1,082	-	171,352	✓
LaST	10,860	-	224,721	✓
CCVID	226	2,856	347,833	✓

为了解决换装视频场景下的数据集缺失问题，作者建立了一个换装视频行人再识别数据集CCVID，该数据集从步态识别数据集FVG的原始数据经过检测后扩展而来。重建的CCVID数据集包含347883个bounding box，每个序列的长度在27帧到410帧之间，平均长度为122帧。作者也提供了包括上半身、下半身、鞋、手提状态和配饰等细粒度的衣物标签。训练集中有75个身份类别；测试集中有151个身份类别，834个序列用作query集合，1074个序列用于gallery集合。

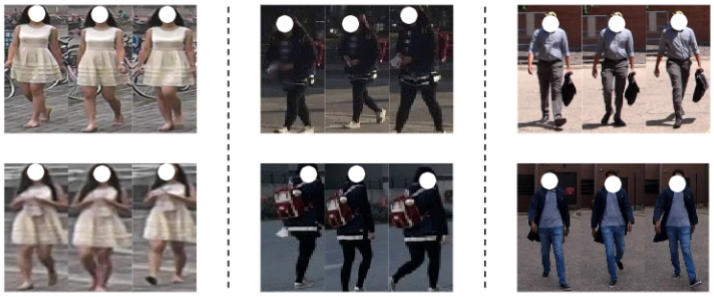


Figure 3. Two different video samples of the same identity on MARS, LS-VID, and CCVID datasets respectively. Only CCVID involves clothes changes.

4. Experiments

4.1 Evaluation Protocol

- 使用Top-1准确率和mAP作为度量指标
- 实验设定分为（1）general setting，换装和不换装的有效样本都参与计算准确率；（2）clothes-changing setting（CC），仅换装的有效样本参与计算准确率；（3）same-clothes setting（SC），仅不换装的有效样本参与计算准确率

4.2 Implementation Details

- backbone：ResNet-50，但去除最后一个降采样，使颗粒度更丰富
- 图片：batch size = 64，每个batch包含8个不同的行人，每个行人有8张图片；视频：batch size = 32，每个batch包含8个不同的行人，每个行人有4段视频切片，每个切片由步长为4的8帧组成
- 优化器：Adam，图片：epoch = 60，$\mathcal{L}_{CA}$在训练25轮次后加入到联合优化中；视频：epoch = 150，$\mathcal{L}_{CA}$在训练50轮次后加入到联合优化中
- 学习率初始化为$3.5e^{-4}$，图片：每经过20轮次就变为原来的$\frac{1}{10}$；视频：每经过40轮次就变为原来的$\frac{1}{10}$
- τ设置为$\frac{1}{16}$，ϵ设置为0.1

4.3 Comparison with State-of-the-art Methods

Table 2. Comparison with state-of-the-art methods on LTCC and PRCC. 'sketch', 'sil.', 'pose', and '3D' represent the contour sketches, silhouettes, human poses, and 3D shape information, respectively.

method	modality	clothes label	extra training data	LTCC				PRCC			
				general		CC		SC		CC	
				top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
HACNN [31]	RGB			60.2	26.7	21.6	9.3	82.5	-	21.8	-
PCB [40]	RGB			65.1	30.6	23.5	10.0	99.8	97.0	41.8	38.7
IANet [20]	RGB			63.7	31.0	25.0	12.6	99.4	98.3	46.3	45.9
SPT+ASE [49]	sketch			-	-	-	-	64.2	-	34.4	-
GI-ReID [28]	RGB+sil.			63.2	29.4	23.7	10.4	80.0	-	33.3	-
CESD [35]	RGB+pose	✓		71.4	34.3	26.2	12.4	-	-	-	-
RCSANet [25]	RGB		✓	-	-	-	-	**100**	97.2	50.2	48.6
3DSL [6]	RGB+pose+sil.+3D	✓		-	-	31.2	14.8	-	-	51.3	-
FSAM [18]	RGB+pose+sil.			73.2	35.4	38.5	16.2	98.8	-	54.5	-
CAL	RGB	✓		**74.2**	**40.8**	**40.1**	**18.0**	**100**	**99.8**	**55.2**	**55.8**

4.4 Ablation Studies

Table 3. The ablation studies of CAL on CCVID, LTCC, and PRCC.

method	CCVID				LTCC				PRCC			
	general		CC		general		CC		SC		CC	
	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP	top-1	mAP
baseline	78.3	75.4	77.3	73.9	65.5	29.4	28.1	11.0	99.8	97.9	45.6	43.3
w/ clothes classifier	58.8	55.8	46.2	45.6	62.3	31.0	21.9	10.9	99.5	99.5	33.1	37.4
CAL	**82.6**	**81.3**	**81.7**	**79.6**	**74.2**	**40.8**	**40.1**	**18.0**	**100**	**99.8**	**55.2**	**55.8**
CAL ($-\mathcal{L}_C$)	52.8	53.0	50.0	49.2	21.5	3.1	9.2	2.3	89.6	67.7	19.3	13.1
Triplet Loss [16]	81.5	78.1	81.1	77.0	71.8	37.5	34.7	16.6	100	99.8	48.6	49.7

- The effectiveness of CAL

baseline方法仅使用身份损失$\mathcal{L}_{ID}$，在baseline的backbone后面添加一个衣物分类器并联合优化$\mathcal{L}_{ID}$和$\mathcal{L}_C$时，SC场景下的精度有所上升，而CC场景下的精度下降明显，因为优化衣物损失$\mathcal{L}_C$会使backbone学习衣物相关的特征。当仅使用$\mathcal{L}_C$训练衣物分类器、使用$\mathcal{L}_{CA}$训练backbone时，CAL在各设定下都超过baseline，CAL帮助了backbone学习到衣物不相关的特征。

- Comparison between different formulations

此处定义$\mathcal{L}_{CA} = -\mathcal{L}_C$，即此时没有多正类，最小化$\mathcal{L}_{CA}$会惩罚模型对所有种类衣物的预测能力，由于衣物类别是身份类别的更细粒度定义，这也会惩罚模型对身份类别的预测能力。

- Comparison with Triplet loss

Triplet loss仅在一个mini-batch中挖掘难样本，而CAL使用一个衣物分类器去保存所有衣物类别，这样就能从全局角度挖掘到衣物不相关的特征。

Table 4. Comparison on standard datasets without clothes-changing.

method	Market-1501		MSMT17	
	top-1	mAP	top-1	mAP
PCB [40]	93.8	81.6	68.2	40.4
IANet [20]	94.4	83.1	75.5	46.8
OSNet [57]	94.8	84.9	78.7	52.9
JDGL [55]	94.8	86.0	77.2	52.3
CircleLoss [39]	94.2	84.9	76.3	50.2
baseline	92.2	78.7	67.8	43.5
CAL	94.5	87.3	79.7	57.0
baseline (w/ triplet)	94.5	86.6	78.9	57.0
CAL (w/ triplet)	94.7	87.5	79.7	57.3

- Results on standard person re-id benchmarks

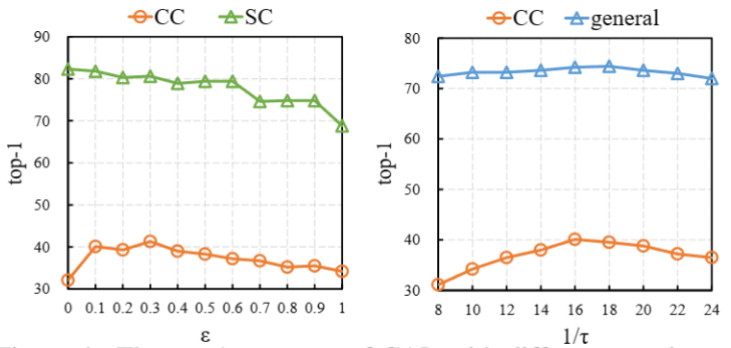


Figure 4. The top-1 accuracy of CAL with different ϵ and τ on LTCC. Note that the abscissa of the second subfigure is the $1/\tau$.

- The influence of ϵ and temperature parameter τ in CAL

随着ϵ的增加，同一身份、相同衣物的正类权重会减少，SC场景下的Top-1精度会下降。对于CC场景，Top-1精度先迅速上升，之后下降，此时已经趋于过拟合了。

4.5 Visualization

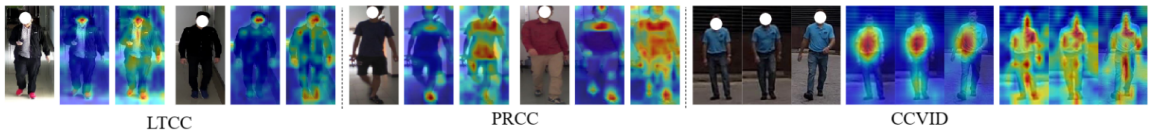


Figure 5. The visualization of feature maps on LTCC, PRCC, and CCVID. In each triplet, the first column presents the original image/video frames. The second and third columns present the feature maps of the baseline method and CAL, respectively.

5. Conclusion

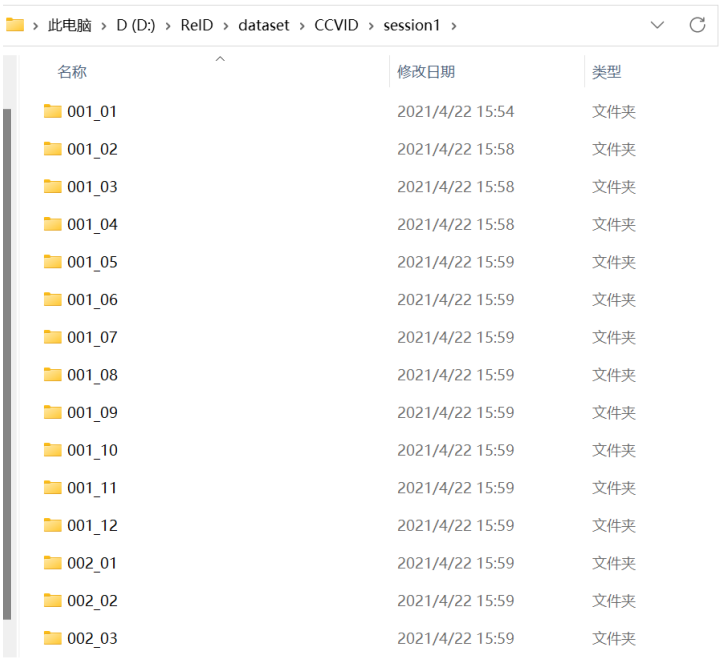
- 论文优点：框架设计simple but effective，因此具有可扩展性；注意到了换装场景下的多正类问题，提出基于衣物的对抗损失（CAL）$\mathcal{L}_{CA}$，第一次使用多正类的分类对多类别的对抗损失进行公式化建模；训练过程中使用了curriculum learning的思想，由易到难，利于模型学习
- 可以改进的地方：目前在换装场景下的数据集标注量过于庞大，可以通过实验发现哪些衣物标注是最有效的，仅关注这些标注可以在减少标注成本的前提下尽量维持原模型的性能，也可以尝试使用机器标注的虚拟数据集；可视化中特征图的衣物部分也被突出，是否可以尝试加实验论证这部分被突出的原因；在算力支持的前提下加入额外分支来处理其他模态的信息，模型就会更具鲁棒性

6. 遗留问题

CCVID数据集中的衣物标签问题。

- CCVID数据集的一个直观展示

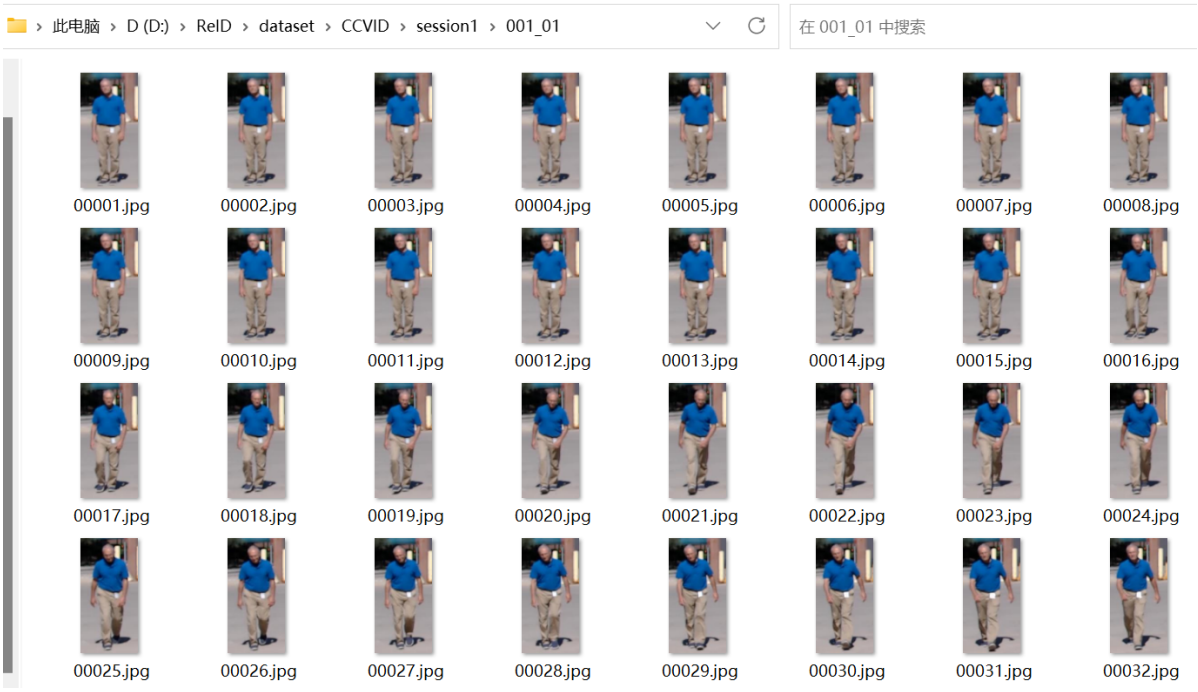
名称	修改日期	类型	大小
session1	2021/4/22 17:21	文件夹	
session2	2021/4/22 18:20	文件夹	
session3	2021/4/22 18:28	文件夹	
gallery.txt	2021/4/26 15:56	文本文档	38 KB
query.txt	2021/4/26 15:56	文本文档	30 KB
readme.txt	2022/4/15 12:35	文本文档	2 KB
train.txt	2021/4/26 15:56	文本文档	34 KB

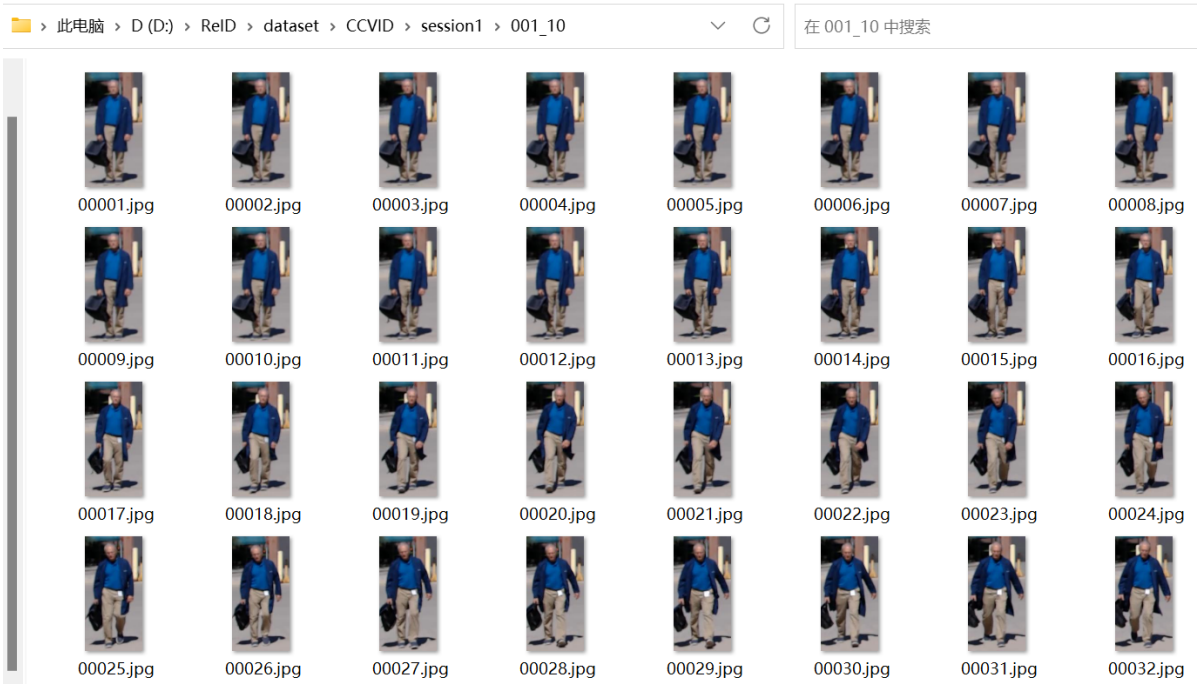


此电脑 › D (D:) › ReID › dataset › CCVID › session1 ›

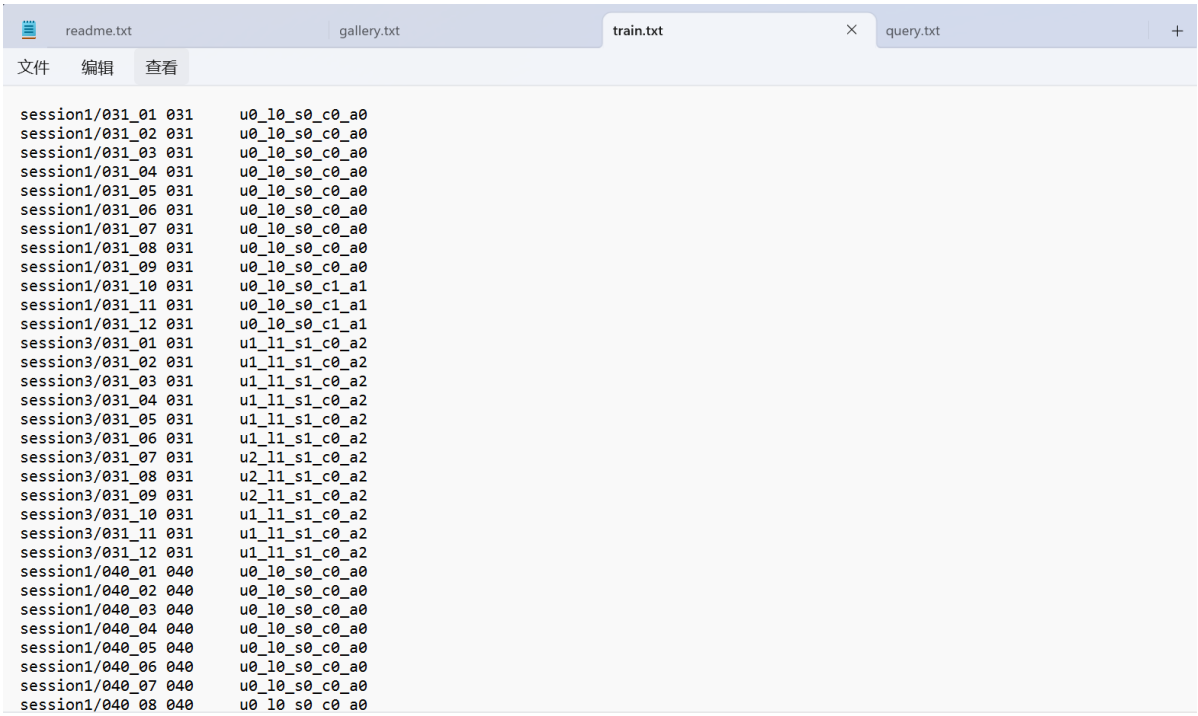
名称	修改日期	类型
001_01	2021/4/22 15:54	文件夹
001_02	2021/4/22 15:58	文件夹
001_03	2021/4/22 15:58	文件夹
001_04	2021/4/22 15:58	文件夹
001_05	2021/4/22 15:59	文件夹
001_06	2021/4/22 15:59	文件夹
001_07	2021/4/22 15:59	文件夹
001_08	2021/4/22 15:59	文件夹
001_09	2021/4/22 15:59	文件夹
001_10	2021/4/22 15:59	文件夹
001_11	2021/4/22 15:59	文件夹
001_12	2021/4/22 15:59	文件夹
002_01	2021/4/22 15:59	文件夹
002_02	2021/4/22 15:59	文件夹
002_03	2021/4/22 15:59	文件夹

```
session1/001_01 001     u0_l0_s0_c0_a0
session1/001_02 001     u0_l0_s0_c0_a0
session1/001_03 001     u0_l0_s0_c0_a0
session1/001_04 001     u0_l0_s0_c0_a0
session1/001_05 001     u0_l0_s0_c0_a0
session1/001_06 001     u0_l0_s0_c0_a0
session1/001_07 001     u0_l0_s0_c0_a0
session1/001_08 001     u0_l0_s0_c0_a0
session1/001_09 001     u0_l0_s0_c0_a0
session1/001_10 001     u1_l0_s0_c1_a0
session1/001_11 001     u1_l0_s0_c1_a0
session1/001_12 001     u1_l0_s0_c1_a0
session1/002_01 002     u0_l0_s0_c0_a0
session1/002_02 002     u0_l0_s0_c0_a0
session1/002_03 002     u0_l0_s0_c0_a0
session1/002_04 002     u0_l0_s0_c0_a0
```





- 衣物标签的处理



每一行分别表示图片路径、行人身份标签、衣物标签。以衣物标签u0_l0_s0_c0_a0为例，u表示上半身衣物，l表示下半身衣物，s表示鞋，c表示手持状态，a表示配饰。

具体处理时，为每一个衣物标签赋予一个数字标签值（0，1，2.....），在训练衣物分类器时使用数字标签值进行训练。