行人重识别中的摄像机风格适应

摘要

作为一个跨摄像机的检索任务，行人重识别受不同摄像机引起的图像风格变化影响。现有的技术通过学习摄像机不变的特征描述子空间来隐式地解决这个问题。在这篇论文中，我们通过引入摄像机风格适应方法(CamStyle)显式地思考了这一问题。CamStyle 可以作为数据增强方法平滑不同摄像机风格的分布。具体来说，通过使用 CycleGAN，可以将已有标签的训练图像风格迁移到每个摄像机，并和原始的训练样本一起形成增强训练集。这种方法在增强了数据多样性的同时也会产生相当大的噪声。为了减轻噪声的影响，我们采用了标签平滑正则化方法(LSR)。我们方法的初始版本(无 LSR)在经常出现过拟合的少摄像机系统上表现良好。我们证明了借助 LSR，无论系统的过拟合程度如何，其性能都得到了改进。我们也得到了与现有最好技术相比有竞争力的准确率。可以在 https://github.com/zhunzhong07/CamStyle 获取代码。

1. 引言

行人重识别(re-ID)[43]是一个跨摄像机检索任务。它旨在从一个多摄像机收集的数据中检索到给定感兴趣的人的同一人。在这个任务中，行人图像往往会有外表和背景的剧烈变化。使用多摄像机拍摄图像是这种变化的重要原因(图 1)。通常，摄像机在分辨率，环境光照等方面彼此相异。

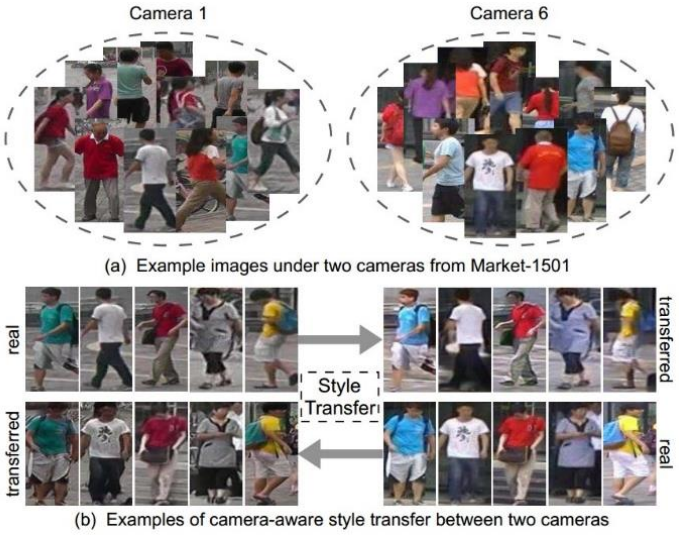

图 1 (a)来自 Market-1501 的例子[42](b)使用我们的方法进行摄像机间风格转化的例子。同一行图像代表同一个行人。

为了解决摄像机变化带来的挑战，以前的文章采取了隐式策略，去学习在不同摄像机下有不变属性的稳定特征表示。传统方法的例子有 KISSME[16]，XQDA[20]，DNS[39]等。深度特征学习方法的例子有 IDE[43]，SVDNet[29]，TripletNet[11]等。

与先前的方法相比，本文从摄像机风格适应的角度采取了显式的策略。我们的灵感主要来自需要大数据容量的基于深度学习的行人重识别方法。为了学得对摄像机变化鲁棒的丰富特征，标注大规模数据有用但代价很高。然而，如果我们能将更多样本加入训练集，以获知不同摄像机间的风格差异，我们就能够 1)解决行人重识别中的数据短缺问题，以及 2)学到跨摄像机的不变特征。更好的是，这个过程不需要更多的人工标注，所以预算能够保持很低。

基于以上的讨论，我们使用摄像机风格适应方法(CamStyle)去规范对行人重识别的 CNN 训练。在初始版本中，我们使用 CycleGAN[51]对每个摄像机对学习了一个图像到图像转化模型。使用习得的 CycleGAN 模型，对一个特定的摄像机拍摄的一张训练图像，我们能生成其他摄像机风格的新的训练样本。在此方法中，训练集由原始图像和风格转化的图像构成。风格转化图像能直接借用原始训练图像的标签。在训练中，我们使用新的训练集按照基准模型对行人重识别 CNN 进行训练。初始方法有利于减轻过拟合和实现摄像机不变属性，但重要的是，我们发现这为系统引入了噪声(图 2)。在有低过拟合风险的相对大量数据的全摄像机系统下，此问题可能降低我们方法的优势。为了缓解这个问题，在改进版本中，我们进一步在风格转化样本中应用了标签平滑正则化(LSR)方法，以让它们的标签在训练时缓和分布。

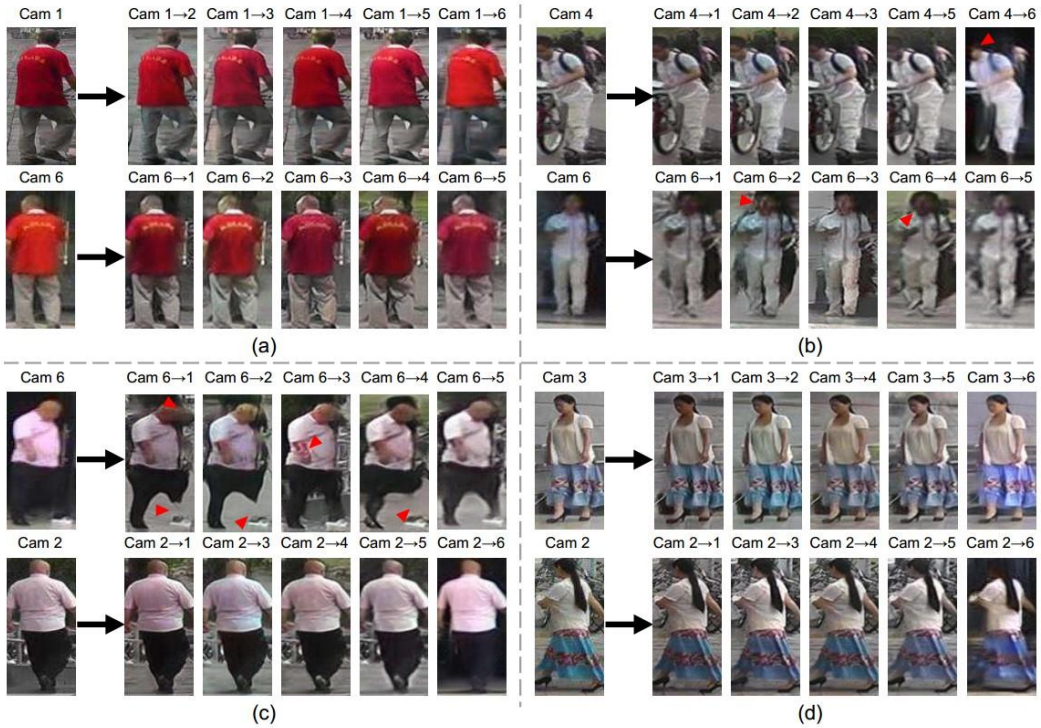

图 2 在 Market-1501 中风格转化的例子。在某一摄像机中拍摄的图像被转化为其他 5 种摄像机风格。除了成功的案例，红色箭头指出的图像到图像转化出现的噪声需要被考虑

我们提出的摄像机风格适应方法，CamStyle，有三个优点。第一，它可以被当作一种数据增强框架，不仅平滑了摄像机风格差异，而且减轻了 CNN 过拟合的影响。第二，通过

整合摄像机信息，它协助网络学到带有摄像机不变属性的行人描述子。最后，CycleGAN 确保其是无监督的，表明其一定的应用价值。总结来说，本篇论文有以下贡献：

• 一个用于行人重识别数据增强的分摄像机的风格转化初始模型。在少摄像机系统中，性能提升达 17.1%.

• 一个在行人重识别训练时应用 LSR 于风格转化样本的改进方法。在全摄像机系统中，连续的提升显而易见。

2. 相关工作

深度学习行人重识别。许多深度学习方法[38, 34, 33, 3, 24]已经在行人重识别领域提出。在[38]中，输入图像对被相对的分成三个重叠的水平部分，并通过一个孪生 CNN 模型用余弦距离学习它们的相似性。随后，Wu 等人[34]使用更小的卷积滤波器来加深网络以得到一个鲁棒的特征。除此之外，Varior 等人[33]将长短时记忆模型融合进一个能序列地处理图像块地孪生网络，这种模型能记住空间信息来提升深度特征的分类性能。

另一个有效的方法是分类模型，这种模型能充分利用行人重识别的标签[43, 35, 29, 18, 36, 44, 41]。Zheng 等人[43]提出身份鉴别嵌入(IDE)，他们将行人重识别模型作为一种图像分类器训练在 ImageNet[17]经过精细调参过的预训练模型。Wu 等人[35]通过将人工设计特征整合入 CNN 特征提出一个特征融合网络(FFN)。最近，Sun 等人[29]使用奇异向量分解迭代地优化全连接层(FC)特征，并生成正交权重。

当一个 CNN 模型相对于训练集规模过于复杂时，可能会发生过拟合。为了解决这个问题，一些数据增强和正则化方法被提出。在[23]中，Niall 等人通过利用背景和线性变换生成不同的样本来提升网络的泛化性能。最近，Zhu 等人[49]用随机值填充的矩形区域随机地擦除输入图像，缓解模型的过拟合并使模型对遮挡鲁棒。Zhu 等人[50]从一个独立的数据集中随机地选择假阳性样本作为额外的训练样本来降低过拟合的风险。关于此工作更进一步的研究，Zheng 等人[47]使用 DCGAN[25]生成无标签样本，并给它们分配一个统一的标签来正则化网络。相对于[47]，此项工作中的风格转化样本是从真实数据中产生的，具有相对可靠的标签。

对抗生成网络。对抗生成网络(GAN)[9]在近些年取得了令人瞩目的成就，尤其是在图像生成[25]方面。最近，GAN 也被应用于图像到图像转化[13，51，22]，风格转化[8，14，6]和跨域图像生成[2，31，5]。Isola 等人[13]使用一种条件性的 GAN 从输入到输出图片中学习一种映射作为图像到图像转化的应用。[13]的主要缺点是它需要一对成对的关联图像作为训练数据。为了克服这个问题，Liu和Tuzel[22]提出了耦合对抗生成网络(CoGAN)，通过应用权值共享网络来学习一个跨域的融合分布。在更近的时候，CycleGAN[51]在 [13]中“像素对像素”框架的基础上引入了循环连续性，从而学习没有配对样本的两个不同域之间的图像转化。风格转化和跨域图像生成也被也被认为是图像到图像的转化，在这种转化中，一种风格(或域)的输入图像被转化为另一种而保留原图像的内容。在[8]中，介绍了一

种分离并重组图像内容和风格的风格转化方法。类似的，[31]提出了域转化网络(DTN)，整合多种 GAN 损失函数以生成未知域的图像，同时保留原始身份。不同于先前方法主要考虑生成样本的质量，本文的工作旨在利用风格转化样本提升行人重识别任务的性能。

3. 提出的方法

在此节中，我们首先在 3.1 小节简单回顾一下 CycleGAN[51]，然后我们将在 3.2 小节阐述使用 CycleGAN 产生分摄像机的数据生成过程。使用 LSR 的基准和训练过程将在 3.3 小节和 3.4 小节独立地阐述。整个网络框架如图 3。

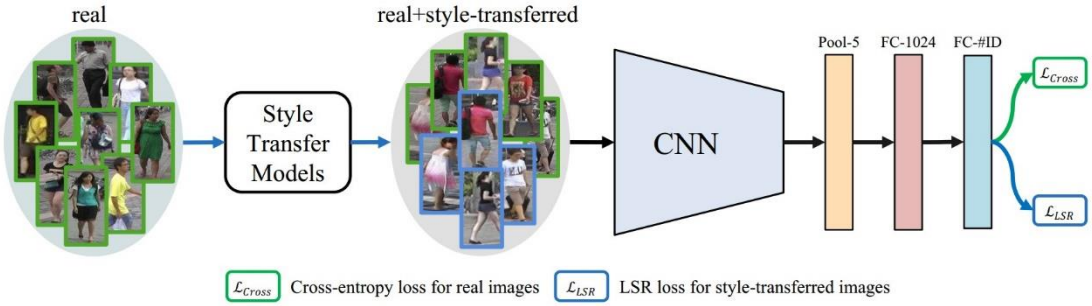

图 3 我们方法的流程。从不同摄像机间的真实图像中学得分摄像机得风格转化模型。对每张真实图像，我们能利用训练好的转化模型生成目标摄像机风格的图像。随后，结合真实图像(绿框)和风格转化图像(蓝框)训练行人重识别 CNN。交叉熵损失和标签平滑正则化损失分别被应用于真实图像和风格转化图像。

3.1 回顾 CycleGAN

给定两个数据集$\{x_i\}_{i=1}^{M}$，$\{y_j\}_{j=1}^{N}$，分别采集自两个不同的域A和B，$x_i \in A$，$y_j \in B$。CycleGAN 的目标是通过使用对抗损失函数学习一个函数映射$G: A \rightarrow B$使得$G(A)$和B的分布一致。CycleGAN 包含两个映射函数，$G: A \rightarrow B$和$F: B \rightarrow A$。两个对抗分类器D_A和用于区分图像是否是另一个域转化而来。CycleGAN 应用 GAN 框架融合地训练分类器和生成器。CycleGAN 的整个损失函数表达式如下：

$$\begin{aligned} V(G,F,D_A,D_B) &= V_{GAN}(D_B,G,A,B) \\ &+V_{GAN}(D_A,F,B,A) \\ &+\lambda V_{cyc}(G,F) \end{aligned} \tag{1}$$

$V_{GAN}(D_B,G,A,B)$和$V_{GAN}(D_A,F,B,A)$分别是对映射函数G和F的损失函数并且是对分类器D_B和D_A的损失函数。$V_{cyc}(G,F)$是循环损失函数迫使$F(G(x)) \approx x$且$G(F(y)) \approx y$，在此约束下，每张图像都能在循环映射后重建。λ调节V_{GAN}和V_{cyc}的比重。在[51]可获得关于 CycleGAN 的更多细节。

3.2 分摄像头的图像到图像转化

在此项工作中，我们使用 CycleGAN 产生新的训练样本，给定采集自L个不同摄像机视角的行人重识别数据集，我们的方法是用 CycleGAN 为每个摄像机对学得图像到图像的转化模型。为了激励风格转化(公式 1)保留输入和输出之间色彩连续性，我们在 CycleGAN 损

失函数中加入了身份映射损失[51]，迫使生成器在使用目标域的真实图片作为输入时逼近一个身份映射。身份映射损失函数的表达式如下：

$$V_{identity}(G,F) = E_{x\sim p_x}[\| F(x)-x \|_1] + E_{y\sim p_y}[\| G(y)-y \|_1] \quad (2)$$

特别的，对于训练图像，我们对每个摄像机对使用 CycleGAN 训练分摄像机的风格转化模型。根据[51]中的训练策略，所有的图像被放缩到256×256。我们对我们的分摄像头风格转化模型使用与 CycleGAN 相同的架构。生成器包含 9 个残差模块和 4 个卷积层，分类器则是一个70×70的 PatchGAN[13]。

对采集自某一摄像机的某张训练图像，我们使用训练好的 CycleGAN 模型生成$L-1$个风格近似于相关摄像机的新的训练图像(例子如图 2 所示)。在这项工作中，我们称生成的图像为**风格转化图像**或**伪图像**。通过这种方法，训练集被扩充为原图像和风格转化图像的结合。由于每个风格转化图像都保留了原图像的内容，新图像被认为与原图像具有相同的身份。这允许我们利用风格转化图像和它们相关的标签连同原有的训练样本一起训练行人重识别 CNN。

讨论。如图 4 所示，我们提出的数据增强方法的工作机制主要包括：1)真实图像与伪(风格转化)图像间相似的数据分布，2)假图像的身份标签被保留。在第一个机制中，伪图像填补了真实数据点之间的空白，并略微扩展了特征空间中类的边界。这保证了增强的数据集通常支持在学习映射期间更好地表征类分布。另一方面，第二个机制支持监督学习[43]的运用，不同于为正则化使用没有标签的 GAN 图像[47]。

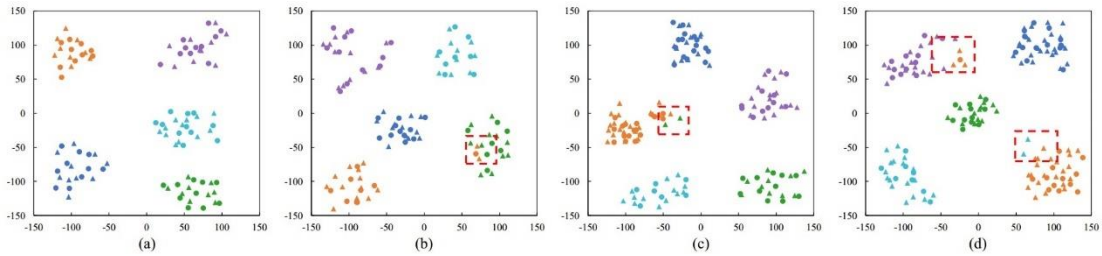

图 4 Barnes-Hut t-SNE[32]在 Market-1501 上的可视化。我们随机选取了 700 个行人身份的真实训练图像训练行人重识别模型，并可视化了其余 20 个行人身份的真实样本(R，点)和他们的伪 (风格转化)样本。在每张图像中，不同的颜色代表不同的行人身份。我们观察到 1)伪样本通常在在真实样本附近，基于它们的数据增强机制；2)噪声伪数据时不时出现(在红框中)，需要正则化技巧如 LSR。

3.3 深度学习行人重识别基准模型

给定拥有身份标签的真实和伪(风格转化)图像，我们使用 IDE[43]训练行人重识别 CNN 模型。使用 Softmax 损失函数，IDE 将行人重识别训练作为一种图像分类任务。我们使用 ResNet-50[10]作为骨干网络，为了优化在 ImageNet[4]上的预训练模型，我们遵循[43]中的训练策略。不同于[43]中提出的 IDE 模型，我们舍弃了最后的 1000 维分类层并加了两个全连接层。第一个全连接层输出有 1024 维，名为“FC-1024”，接着为批量正则化层[12]，

ReLU 层和 Dropout 层[27]。“FC-1024”的添加遵循[29]中的实践，可提高准确性。第二个全连接层的输出是C维，公式维为训练集中的类别数。在我们的实践中，所有的输入图像被放缩到256×128。整个网络如图 3 所示。

3.4 使用 CamStyle 训练

给定一个包含真实和伪(风格转化)图像(带有它们的身份标签)的新的训练集，此节讨论使用 Camstyle 的训练策略。当我们平等地看待真实图像和伪图像，例如，为它们分配一个“独热”标签分布，我们就得到了我们方法的初始版本。另一方面，考虑到伪样本引入的噪声，我们提出了一个包含标签平滑正则化方法(LSR)[30]的完全版方法。

初始版本。在初始版本中，新训练中的每一个样本都有一个单独的身份。在训练中，我们在每一个最小训练批次中随机地选择M个真图像和N个伪图像。损失函数方程可被写作以下形式，

$$\mathcal{L} = \frac{1}{M}\sum_{i=1}^{M}\mathcal{L}_R^i + \frac{1}{N}\sum_{j=1}^{N}\mathcal{L}_F^j \tag{3}$$

$\mathcal{L}_R$和$\mathcal{L}_F$分别为真图像和伪图像的交叉熵损失。交叉熵损失函数形如，

$$\mathcal{L}_{Cross} = -\sum_{c=1}^{C} log(p(c))q(c) \tag{4}$$

C是类别数，$p(c)$是输入属于标签c的预测概率。$p(c)$被 softmax 层归一化，所以$\sum_{c=1}^{C} p(c) = 1$。$q(c)$是真实分布。由于训练集中的每个人都有一个单独标签y，$q(c)$可被定义为，

$$q(c) = \begin{cases} 1 & c = y \\ 0 & c \neq y \end{cases} \tag{5}$$

所以最小化交叉熵等同于最大化真实标签的概率。对于一个有一个身份为y的人，方程 4 中的交叉熵损失可写作，

$$\mathcal{L}_{Cross} = -log(p(y)) \tag{6}$$

由于真伪数据全局数据分布的相似性，初始版本能够提升基准 IDE 在少摄像机系统下的准确率，如 4 节所示。

完全版本。风格转化图像有正向的数据增强功能，但也引入了噪声。因此，虽然初始版本在由于缺少数据，倾向发生过拟合的少摄像机系统中有减少过拟合的优点，在有更多的摄像机的情况下，它的作用会被削弱。原因是当能获得来自更多摄像机的数据时，过度拟合问题会不那么重要，转化产生的噪声问题会开始显现。

转化噪声产自两个方面：1)CycleGAN 没有完美地模拟转化过程，所以在图像生成时会发生错误。2)由于遮挡和检测错误的发生，在真实数据中存在噪声样本，将这些噪声样本转化为伪数据，甚至可能产生更多的噪声样本。在图 4 中，我们在 2-D 空间中可视化了真伪数据的深度特征的一些样本。大多数生成样本分布在原图像的周围。当发生转化错误时(如图 4(c)，图 4(d))，伪样本会成为一个噪声样本并远离真实分布。当一个真实图像是一个噪声样本时((如图 4(b))和图 4(d))，它将会远离有相同标签的图像，所以它的生成样本也会是噪声。在有相对大量数据，低过拟合风险的全摄像机系统中，这个问题减少了生成样本的收益。

为了缓解这个问题我们在风格转化图像上使用了标签平滑正则化(LSR)[30]以平缓它们的分布。换言之，我们对真实标签取更小的置信度并给其他类赋一个小的权重。每个风格转化图像的标签分布的重赋值可写为，

$$q_{LSR}(c) = \begin{cases} 1-\epsilon+\frac{\epsilon}{C} & c = y \\ \frac{\epsilon}{C} & c \neq y \end{cases} \tag{7}$$

$\epsilon \in [0,1]$，当$\epsilon = 0$，等式 7 将被退化为等式 5。则方程 4 中的交叉熵损失将被重新定义为，

$$\mathcal{L}_{LSR} = -(1-\epsilon)logp(y) - \frac{\epsilon}{C}\sum_{c=1}^{C} logp(c) \tag{8}$$

对于真实图像，我们不使用 LSR，因为他们的标签正确匹配图像内容。更进一步地，我们的实验表明对真实图像使用 LSR 不能在全摄像机系统中提高行人重识别性能。所以对真实图像，我们使用独热标签分布。对于风格转化图像，我们设$\epsilon = 0.1$，损失函数$\mathcal{L}_F = \mathcal{L}_{LSR}(\epsilon = 0.1)$。

讨论。最近，Zheng 等人[47]提出了离群值标签平滑正则化(LSRO)以使用 DCGAN[25]生成地无标签样本。在[47]中，由于生成样本没有标签，生成样本被赋予一个统一的标签，例如$\mathcal{L}_{LSR}(\epsilon = 0.1)$。与 LSRO[47]相比，我们的系统有两项不同之处。1)伪图像根据摄像机风格生成。CycleGAN 的应用保证了生成图像保留行人的主要特点(图 5 提供了一些可视化的比较)。2)我们系统中的标签更可靠。我们使用 LSR 解决一小部分不可靠标签，而 LSRO[47]被用于无标签可得的场景中。

图 5 我们的方法生成的例示样本与 DCGAN[47]中比较

4. 实验

4.1 数据集

我们在 Market-1501[42]和 DukeMTMC-reID[47，26]上评估了我们的方法，因为这两个数据集都是 1)大规模的且 2)为每个图像提供了摄像机标签。

Market-1501[42]采集自 6 个摄像机视角的 1501 个行人的 32668 打过标签的图像。这些图像由可变部件检测模型[7]检测而出。数据集分为两个固定的部分：751 个行人的 12936 张图像用于训练，750 个行人的 19732 张图像用于测试。在测试集中，平均每个行人由 17.2 张

图像。在测试中，750 个行人的 3368 张人工标注数据被用作查询项检索数据库中的匹配行人。我们使用了单查询评估。

DukeMTMC-reID[47]是一个新发布的大规模行人重识别数据集。它采集自 8 个摄像机，包含 1404 个行人的 36411 张图像。与 Market-1501 像素，它分为 702 个行人的张训练图像，702 个行人的 2228 张查询图像和 177661 张数据库图像。我们使用 rank-1 准确度和平均准确度均值在两个数据集上做评估。

4.2 实验预设

分摄像机的风格转化模型。给定 3.2 小节，给定一个采集自L个摄像机视角的训练集，我们为每个摄像机对训练一个分摄像机的风格转化模型。特别地，我们分别为 Market-1501 和 DukeMTMC-reID 训练$C_6^2 = 15$和$C_8^2 = 28$个 CycleGAN 模型。对于所有的实验，在训练中，我们将训练图形放缩到256×256并使用 Adam 优化器从$\lambda = 10$开始训练模型。我们设置训练批次大小为 1，生成器的学习率为 0.0002，分类器的学习率为 0.0001，并且两者的学习率在最初的 30 轮训练中不变，在剩余的 20 轮训练中线性减小到 0。在摄像机风格转化中，对每个训练图像，我们生成$L-1$个(Market-1501 5 个，DukeMTMC-reID 7 个)额外的伪图像作为增强的训练数据，并保留它们的原身份。

行人重识别基准 CNN 模型。为了训练出这一基准，我们遵循[43]中的训练策略。特别地，我们保持所有图像的横纵比，并把他们放缩到256×128。在训练中，我们使用了两种数据增强方法，随机裁剪和随机水平翻转。dropout 概率p公式设为 0.5。我们使用 ResNet-50[10]作为骨干网络，第二个全连接层相对 Market-1501 和 DukeMTMC-reID 分别有 751 和 702 个神经元。我们使用 SGD 算法训练行人重识别模型，并设训练批次大小为 128。在 40 轮训练后，学习率变为初始值的十分之一，我们总计进行 50 轮训练。在测试中，我们提取 Pool-5 层的输出作为图像的描述子(2048 维)，使用欧氏距离计算图像间的相似度。

训练使用 CamStyle 的 CNN。我们使用与训练基准模型时相同的配置，除了我们在每个最小训练批次挑选M个真实图像和N个伪(风格转化)图像。如果没有特殊说明，我们设$M = N$。注意，由于伪图像的数量大于真实图像，在每轮训练中，我们使用所有的真实图像和伪图像的$\frac{N}{M} \times \frac{1}{L-1}$。

4.3 参数分析

Camstyle 涉及一个重要的比例$\frac{M}{N}$，M和N分别时真实和伪(风格迁移)的训练样本在最小训练批次中的数量。这个参数代表伪样本在训练中的比例。通过调节这个比例，我们的实验结果如图 6 所示。可以看出，有不同的$\frac{M}{N}$ 的 CamStyle 相对于基准模型有持续性的改进。当在每个最小训练批次中使用比真实图像更多的伪图像($M : N < 1$)时，我们的方法在 rank-1 准确率上取得了大概 1%的提升。相反，当$M : N > 1$，我们的方法在 rank-1 准确率产生了超过 2%的提升。当$M : N = 3 : 1$时，性能达到最好。

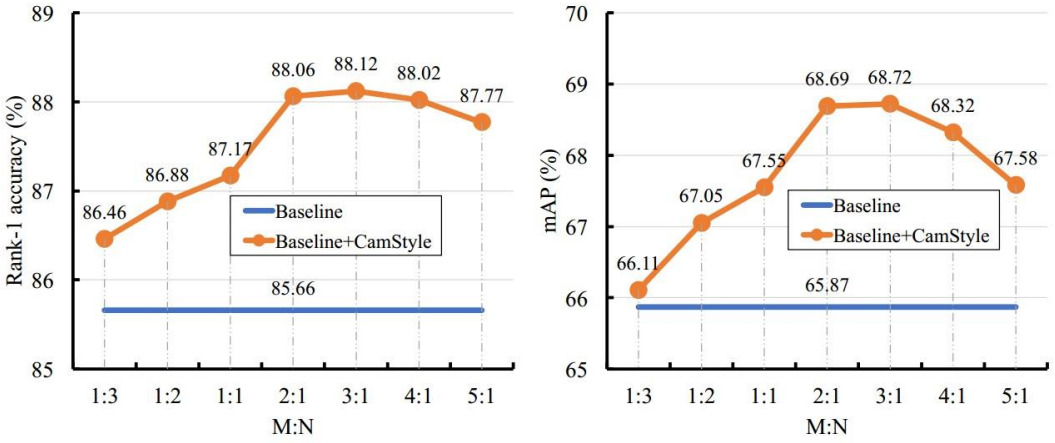

图 6 Market-1501 中的一个小型训练集在不同真实数据和伪数据比例(M: N)下的评估

4.4 变构评估

基准评估。为了全面地展现 CamStyle 地效果，我们的基准模型分别包括 Market-1501 的 2，3，4，5，6 号摄像机和 DukeMTMC-reID 的 2，3，4，5，8 号摄像机。例如，在带有 3 个摄像机的一个系统中，训练集和测试集都含有 3 个摄像机。在图 7 中，随着摄像机数量的增多，rank-1 准确率增加。这是因为 1)获得了更多的训练数据，2)当更多的真实数据出现在数据库中，更容易找到一个 rank-1 的真匹配。在全摄像机(Market-1501 中的 6 个，DukeMTMC-reID 中的 8 个)的基准系统中，Market-1501 上的 rank-1 准确率为 85.6%，DukeMTMC-reID 上的 rank-1 准确率为 72.3%。

初始 CamStyle 提升了少摄像机系统的准确率。我们首先在图 7 和表 1 中评估了初始版方法(无 LSR)的效果。首先，在有两个摄像机的系统中，初始 CamStyle 相对于基准 CNN 产生了显著的提升。在有两个摄像机的 Market-1501 上，提升达+17.1%(从 43.2%到 60.3%)。在有两个摄像机的 DukeMTMC-reID 上，rank-1 准确率从 45.3%提升至 54.8%。这表明，少摄像机系统由于缺少训练数据倾向于过拟合，所以我们的方法出现了显著系统性能提升。

第二，随着系统中的摄像机数量增加，初始 CamStyle 的提升开始减慢。例如，在 Market-1501 的 6 摄像机系统中，rank-1 准确率的提升仅为+0.7%。这表明，1)在全系统中过拟合问题并不严重，2)CycleGAN 带来的噪声开始影响系统的准确性。

表 1 使用不同损失函数在 Market-1501 上的性能评估。CrossE：交叉熵，LSR：标签平滑正则化[30]

Training data	$\mathcal{L}_R$	$\mathcal{L}_F$	Rank-1	mAP
Real	CrossE	None	85.66	66.87
Real	LSR	None	85.21	65.60
Real+Fake	CrossE	CrossE	86.31	66.02
Real+Fake	CrossE	LSR	**88.12**	**68.72**

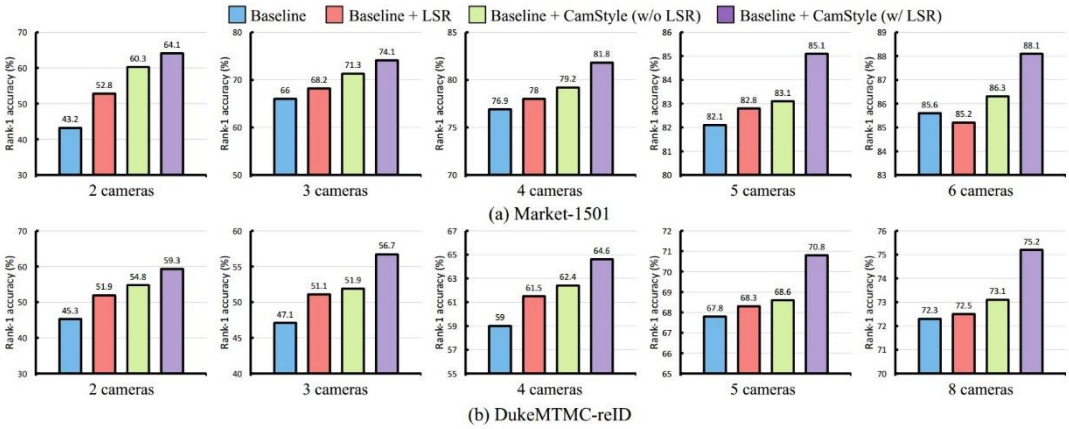

图 7 在 Market-1501 和 DukeMTMC-reID 上不同方法的比较，基准，基准+LSR，初始版基准+CamStyle(无 LSR)，基准+CamStyle(有 LSR)。展示了 rank-1 准确率。展示了 5 个系统，分别有 Market-1501 的 2，3，4，5，6 号摄像机，DukeMTMC-reID 的 2，3，4，5，8 号摄像机。CamStyle(有 LSR)比基准模型产生了持续性提升。

LSR 对 CamStyle 是有效的。正如先前的描述，当在一个有超过 3 个摄像机的系统中测试时，初始 CamStyle 相比两摄像机系统得到更少的提升。我们在图 7 和表 1 中显示在伪图像上使用 LSR 损失相比交叉熵损失取得了更高的性能。如表 1 所示，在 Market-1501 的全摄像机系统对风格转化数据使用交叉熵会提升 rank-1 准确率到 86.31%。用 LSR 替换用于伪数据的交叉熵，rank-1 准确率将达 88.12%。

特别地，图 7 和表 1 显示在真实数据上单独使用 LSR 帮助不大，甚至会降低全摄像机系统的性能。因此，事实上使用 LSR 的 CamStyle 提升基准性能不是 LSR 单独的特性，而是伪图像和 LSR 的相互作用。在这项实验中，我们证明了在伪图像上使用 LSR 的必要性。

使用不同的摄像机对训练分摄像机风格转化模型的影响。在表 2 中，我们展示了使用更多的摄像机去训练分摄像机风格转化模型，rank-1 准确率从 85.66%提升至 88.12%。尤其是，我们的方法在仅使用 1 号和 2 号摄像机训练分摄像机的风格转化模型的情况下取得了 rank-1 准确率+1.54%的提升。除此以外，当使用 5 个摄像机训练摄像机风格转化模型时，我们取得了 87.85%的 rank-1 准确率，相比 6 摄像机低了 0.27%。这表明，即使只是用一部分摄像机训练分摄像机的风格转化模型，我们的方法能取得与使用全摄像机相似的结果。

CamStyle 是对不同数据增强方法的补足。为了更深入地验证 CamStyle，我们将它与两种数据增强方法，随机翻转加随机裁剪(RF+RC)和随机擦除(RE)[49]比较。RF+RC 是 CNN 训练中常见方法，可提高对图像翻转和目标转化的鲁棒性。RE 被用来提升对遮挡的不变性。

如表 3 所示，当未使用数据增强时，rank-1 准确率为 84.15%。当单独应用 RF+RC，RE，或 CamStyle，rank-1 准确率分别提升至 85.66%，86.83%，85.01%。当我们将 CamStyle 与 RF+RC 或 RE 任一种相结合，我们发现与单独应用他们相比有持续性提升。当三种数据增强方法同时应用时，会达到最佳性能。因此，尽管三个直接的数据增强方法聚焦于 CNN 不变性的不同方面，我们的结果表明，CamStyle 是其他两种很好的补足。尤其是，结合这

表 2 使用不同的摄像机对在 Market-1501 上训练 CycleGAN 的影响的分析。我们采用了 6 摄像机系统。我们起初使用 6 摄像机，随后逐渐加入其他摄像机。

Method	Rank-1	mAP
Baseline	85.66	65.87
Baseline+CamStyle(1+2)	87.20	67.64
Baseline+CamStyle(1+2+3)	87.32	68.53
Baseline+CamStyle(1+2+3+4)	87.42	68.23
Baseline+CamStyle(1+2+3+4+5)	87.85	68.51
Baseline+CamStyle(1+2+3+4+5+6)	**88.12**	**68.72**

表 3 在 Market-1501 上不同的数据增强组合的比较。**RF+RC**：随机翻转+随机裁剪，**RE**：随机擦除[40]

Method	RF+RC	RE	CamStyle	Rank-1	mAP
Baseline				84.15	64.10
	✓			85.66	65.87
		✓		86.83	68.50
			✓	85.01	64.85
	✓	✓		87.65	69.91
	✓		✓	88.12	68.72
		✓	✓	87.89	69.10
	✓	✓	✓	**89.49**	**71.55**

三种方法，我们取得了 89.49%的 rank-1 准确率。

4.5 与最高水准方法的比较

在表 4 和表 5 中，我们分别在 Market-1501 和 DukeMTMC-reID 上与现有最好方法进行了比较。首先，使用我们的基准训练策略，我们在两个数据集上得到了一个较强的基准模型(IDE*)。特别的，IDE*在 Market-1501 取得了 85.66%的 rank-1 准确率，在 DukeMTMC-reID 取得了 72.31%的 rank-1 准确率。与已发布的 IDE 应用[19，47，43]相比，IDE*在 Market-1501 有最高的 rank-1 准确率。

然后，当将 CamStyle 应用于 IDE*，我们与现有最好方法不相上下的结果。尤其是，我们在 Market-1501 上的 rank-1 准确率=88.12%，在 DukeMTMC-reID 上的 rank-1 准确率=75.12%。在 Market-1501 上，我们的方法与 PDF[28]，TriNet[11]，DJL[19]相比有更高的 rank-1 准确率。另一方面，我们的方法在 Market-1501 上的 mAP 比 TriNet[11]略微低了 0.42%,在 DukeMTMC-reID 比 SVDNet[29]低了 3.32%。

表 4 与在 Market-1501 上现有最好方法的比较。IDE*指使用文中训练程序增强版的 IDE。**RE**：随机擦除[49]。

Method	Rank-1	mAP
BOW[42]	34.40	14.09
LOMO+XQDA[20]	43.79	22.22
DNS[39]	61.02	35.68
IDE[43]	72.54	46.00
Re-rank[48]	77.11	63.63
DLCE[45]	79.5	59.9
MSCAN[18]	80.31	57.53
DF[40]	81.0	63.4
SSM[1]	82.21	68.80
SVDNet[29]	82.3	62.1
GAN[47]	83.97	66.07
PDF[28]	84.14	63.41
TriNet[11]	84.92	69.14
DJL[19]	85.1	65.5
IDE*	85.66	65.87
IDE*+CamStyle	88.12	68.72
IDE*+CamStyle+RE[49]	**89.49**	**71.55**

表 5 与在 DukeMTMC-reID 上现有最好方法的比较。IDE*指使用文中训练程序增强版的 IDE。**RE**：随机擦除[49]。

Method	Rank-1	mAP
BOW+kissme [42]	25.13	12.17
LOMO+XQDA [20]	30.75	17.04
IDE [43]	65.22	44.99
GAN [47]	67.68	47.13
OIM [37]	68.1	47.4
APR [21]	70.69	51.88
PAN [46]	71.59	51.51
TriNet [11]	72.44	53.50
SVDNet [29]	76.7	56.8
IDE*	72.31	51.83
IDE*+CamStyle	75.27	53.48
IDE*+CamStyle+RE [49]	**78.32**	**57.61**

进一步将 CamStyle 与 Random Erasing 数据增强[49]结合(RF+RC 已被应用于基准模型中)，我们最终的 rank-1 准确率将在 Market-1501 上达 89.49%，在 DukeMTMC-reID 上达 89.49%。

5. 总结

在这篇论文中，我们提出了 CamStyle，一种对深度行人重识别的摄像机风格适应模型。CycleGAN 用于从原始图像生成新的训练图像，使用 CycleGAN 为每个摄像机对学习一种风格转化模型。真实图像和风格转化模型形成新的训练集。此外，为了缓解 CycleGAN 引起的噪声等级的提升，标签平滑正则化(LSR)被应用于生成样本。在 Market-1501 和 DukeMTMC-reID 数据集上的实验表明，结合 LSR，我们的方法能有效减小过拟合的影响，产生高于基准的持续提升。此外，我们也展现了我们的方法对其他数据增强方法的补足效果。未来，我们将扩展 CamStyle 在 one view 学习和域适应中的应用。

参考文献

[1] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In CVPR, 2017. 8

[2] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In CVPR, 2017. 2, 3

[3] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In CVPR, 2017. 2

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. FeiFei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009. 4

[5] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and
J. Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person reidentification. In CVPR, 2018. 2

[6] X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Style aggregated network for facial landmark detection. In CVPR, 2018. 2

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained partbased models. IEEE TPAMI, 2010. 5

[8] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In CVPR, 2016. 2, 3

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In NIPS, 2014. 2

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016. 4, 6

[11] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, 2017. 1, 8

[12] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In ICML, 2015. 4

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In CVPR, 2017. 2, 4

[14] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In ECCV, 2016. 2

[15] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In ICLR, 2015. 6

[16] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and¨ H. Bischof. Large scale metric learning from equivalence constraints. In CVPR, 2012. 1

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012. 2, 8

[18] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In CVPR, 2017. 2, 8

[19] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In IJCAI, 2017. 8

[20] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In CVPR, 2015. 1, 8

[21] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. arXiv preprint arXiv:1703.07220, 2017. 8

[22] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In NIPS, 2016. 2

[23] N. McLaughlin, J. M. Del Rincon, and P. Miller. Dataaugmentation for reducing dataset bias in person reidentification. In AVSS, 2015. 2

[24] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue. Multiscale deep learning architectures for person re-identification. In ICCV, 2017. 2

[25] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In ICLR, 2016. 2, 5

[26] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multicamera tracking. In ECCV workshop, 2016. 5

[27] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. JMLR, 2014. 4

[28] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian. Posedriven deep convolutional model for person re-identification. In ICCV, 2017. 8

[29] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In ICCV, 2017. 1, 2, 4, 8

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In CVPR, 2016. 2, 4, 5, 7

[31] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised crossdomain image generation. In ICLR, 2017. 2, 3

[32] L. Van Der Maaten. Accelerating t-sne using tree-based algorithms. JMLR, 2014. 4

[33] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human reidentification. In ECCV, 2016. 2

[34] L. Wu, C. Shen, and A. v. d. Hengel. Personnet: Person re-identification with deep convolutional neural networks.
arXiv preprint arXiv:1601.07255, 2016. 2

[35] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng. An enhanced deep feature representation for person re-identification. In WACV, 2016. 2

[36] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In CVPR, 2018. 2

[37] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In CVPR, 2017. 8

[38] D. Yi, Z. Lei, S. Liao, S. Z. Li, et al. Deep metric learning for person re-identification. In ICPR, 2014. 2

[39] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In CVPR, 2016. 1, 8

[40] L. Zhao, X. Li, J. Wang, and Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. In ICCV, 2017. 8

[41] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In ECCV, 2016. 2

[42] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In ICCV, 2015. 1, 2, 5, 8

[43] L. Zheng, Y. Yang, and A. G. Hauptmann. Person reidentification: Past, present and future. arXiv preprint arXiv:1610.02984, 2016. 1, 2, 4, 6, 8

[44] L. Zheng, H. Zhang, S. Sun, M. Chandraker, and Q. Tian. Person re-identification in the wild. In CVPR, 2017. 2

[45] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person re-identification. arXiv preprint arXiv:1611.05666, 2016. 8

[46] Z. Zheng, L. Zheng, and Y. Yang. Pedestrian alignment network for large-scale person re-identification. arXiv preprint arXiv:1707.00408, 2017. 8

[47] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In ICCV, 2017. 2, 4, 5, 8

[48] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In CVPR, 2017. 8

[49] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. arXiv preprint arXiv:1708.04896, 2017. 2, 7, 8

[50] F. Zhu, X. Kong, H. Fu, and Q. Tian. Pseudo-positive regularization for deep person re-identification. Multimedia Systems, 2017. 2

[51] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired imageto-image translation using cycle-consistent adversarial networks. In ICCV, 2017. 1, 2, 3