有序还是无序：重新审视视频行人重识别

摘要—为了学习视频行人重识别的良好视觉表示，循环网络真的必要吗？在本文中，我们首先表明，采用递归神经网络（RNN）聚合时空特征的通用做法可能不是最优的。具体而言，通过诊断分析，我们表明，递归结构可能无法像我们期望的那样有效地学习时间依赖性，并且隐式地产生了无序表示。基于这个发现，对于视频行人重识别任务，我们提出一个简单但是有效的方法。其中，我们把视频行人重识别任务是做基于图像行人重识别问题的高效无序集合。更具体地说，我们把视频分成独立的图像和重新识别具有图像排名的人员。基于独立同分布假设，我们提供一个错误界限，该错误界限能帮助我们减轻如何改善视频行人重识别任务的负担。我们工作也提出一种方式去弥补视频行人重识别和图像行人重识别之间的差距。全面的实验评估表明，所提出的解决方案在多个广泛使用的数据集（iLIDS-VID, PRID 2011, and MARS）上实现了最先进的性能。

Index Terms—深度学习，集成学习，视频行人重识别。

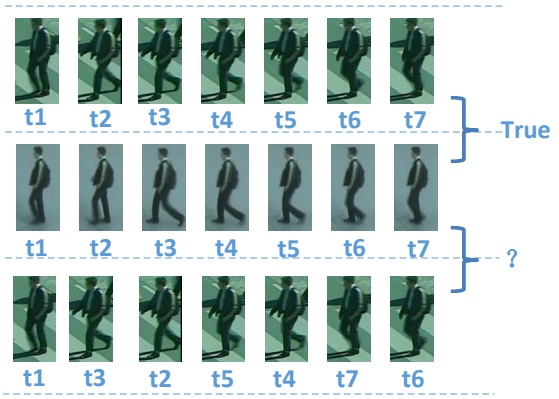


图 1. **出发点**。上面两行: 现有方法采取 RNN 建立时序依赖。下面两行: 人类可以轻松从时序无关的图像中完成视频行人重识别的任务。

1 引言

行人重新识别（Re-id）解决的是在不相交的相机视图之间重新关联人员的问题。在本文中，我们考虑了视频行人重识别的更实际情景。即在一台摄像机中看到的一个人的视频必须与其他非重叠摄像机捕获的视频库进行匹配。由于该任务在诸如视觉监视和取证等问题中的广泛应用，因此视频行人重识别是计算机视觉中一个广受关注的研究主题。自从开拓性的工作 [1] 至今，几个视觉特征 [2], [3], 和学习方法 [4]–[6] 不断提高匹配性能，引导研究方向至解决复杂数据集中更具挑战性的场景 [4], [7], [8]。然而，由于外观，视点，照明和遮挡的变化而导致的重大障碍成为解决问题的途径。

最近，深度卷积神经网络站在了一些视觉领域的前沿，包括图像分类 [9]–[12]、分割 [13], [14]、深度估计 [15]、姿态估计 [16]、人脸识别 [17]、人群统计 [18] 和图像行人重识别 [19]。但是，针对的视频行人重识别深度学习也见证了如此广泛的普及。特别地，之前的方法 [6], [20]–[25] 通常通过使用 RNN 时序性地融合从 ConvNets 提取的空间信息来处理视频。

但是，与其他顺序建模任务 [26] 不同，使用 RNN 对视频行人重识别的时序性地提取空间信息进行建模的好处鲜为人知。尽管从理论上讲引人入胜，但我们表明一种典型的递归结构可能无法像假定的那样有效地捕获时间依赖性。为了简化我们的分析，我们发现使用 RNN 及其任务的后续工作的开创性工作 [20] 导致了无序表示。这激发了我们对视频行人重识别的任务做出重新思考，如 Fig. 1所示。根据可分离的视觉通路假说 [27]，人的视觉皮层包含腹侧流和背侧流，它们分别识别物体及其运动方式。现有的采用 RNN 的方法 [20] 过分强调了与人的动作有关的时间依赖性。不幸的是，这些行为生物识别技术遭受了很大的困扰，使得现有的 RNN 难以推广。在这项工作中，我们假设来自单个静止图像的外观信息在重新关联来自不同相机的人时起着更重要的作用。以 Fig. 1为例，即使第二台摄像机中的图像帧已被手动改组以消除时间依赖性，人类也可以轻松地对两台摄像机中的图像是否来自同一身份进行分类。

宣布无序编码在视频分析中的优势是没有争议的，实际上，我们当然不是第一个将无序编码注入基于视频的任务中的人。即使在诸如人类动作识别之类的其他任务中，每个视频帧之间的时间依赖性都被认为是至关重要的，无序表示 [28]–[30] 仍然可以证明其优于循环结构 [31]。尽管我们从这些工作中获得启发，并在基于图像的人员重新识别方面取得了近期成功，但我们还是第一个深入研究为视频行人重识别建立时间信息建模的方法的人。更具体地说，我们通过将视频划分为单个图像并以基于共享图像的排名的无序集合来执行任务，从而提供了一种简单而出奇的强大方法。我们的研究表明，在独立同分布假设下，当每个基础排名优于随机猜测时，排名的集合是一致的，并且任务的错误界限随图像帧的数量呈指数下降。我们的工作为进一步强调视频行人重识别的外观信息开辟了一条道路。它不仅可以在多个数据集上（iLIDS-VID，PRID 2011 和 MARS）上实现最先进的性能，而且可以缩小

基于视频和图像的人员重新识别之间的差距。

2 相关工作

2.1 视频行人重识别的循环结构

递归结构，例如 RNNS，LSTM，GRU 等，已被广泛用于时序性地聚合从每个视频帧提取的空间信息。Yan *et al.* [6] 采用 LSTM 来融合手动设计的特征。同样的, 循环网络 [20] 被用来融合从视频独立帧中卷积得到的判别特征。Wu *et al.* [21] 在底层考虑卷积激活, 这些嵌入到 GRU 中来捕获时序信息。[32] 中引入双向 RNN 来整合卷积特征。DFGP 被 [33] 提出来同时使用深度特征和 LOMO 人为设计特征。类似的"CNN-RNN" 结构在 [22] 中分别针对全身和部分身体图像序列进行训练, 以从整体和局部角度学习互补表示。空间和时间注意力池化网络 (ASTPN) [23] 通过将池化分解为来自 CNN 的特征图上的空间池，以及将 RNN 的输出上的一个专心的时间池，从而扩展了标准的 CNN-RNN。在 [24] 中提出了空间递归模型和时间注意模型，以选择判别框架并在测量相似性时利用上下文信息。[25] 中介绍了基于循环的竞争性片段相似性聚合和共同关注的片段嵌入。[34] 使用基于循环的新型时空协同残差网络提取时空特征。

2.2 图像行人重识别

与视频行人重识别相比，基于图像的人的重新识别已经得到了更广泛的研究。在 [19] 中设计一个过滤器配对神经网络共同处理错位和几何变换。在 [35] 中提出了一种跨输入差异 CNN，用于基于来自每个输入图像的中间特征捕获两个输入图像之间的局部关系。Xiao *et al.* 引入了领域指导的辍学技术，以缓解来自各种数据集的领域差距 [36]。其他一些工作旨在设计更健壮的损失函数，例如三元组损失 [37]–[39] 来强制图像三元组之间的相对距离的正确顺序或四元组损失 [40] 以结合对比的优势损失和三重损失。[41], [42] 中不同训练图像之间的相似性有助于重新识别。Song *et al.* [43] 提出了一种新颖的区域级三重态损失，以将特征从整个图像和身体区域拉近，而将特征从背景中拉开。

2.3 缩小研究差距

尽管使用 RNN 来完成视频行人重识别的大量工作，但几乎没有人证明使用 RNN 提高视频行人重识别的准确性确实可以追溯到递归结构对图像行人重识别中缺少的时间信息进行编码。相反，研究人员在很大程度上认为循环结构能够成功捕获视频行人重识别中的时间依赖性。要研究对视频行人重识别使用循环结构的效果，我们首次尝试通过对视频行人重识别使用递归结构来研究利用时间信息的效果。请注意其它没有循环神经网络的方法 [44] 使用的是强化学习。但是，他们没有系统地研究帧顺序的影响。另一个例子是 [45]，在该示例中，作者得出了递归连接的近似值，并通过更简单的前馈体系结构获得了更快，更具竞争性的结果。其他方法 [24] 尝试使用递归结构来建模上下文信息，其有效性最近已得到验证 [46]。但是，他们仍然缺乏系统的研究来证明 RNN 的上下文建模在帧顺序方面的能力。如 Sec. 4中所示，我们的诊断分析表明了常用的 "CNN-RNN" 管道的不足：循环结构在建模时间信息方面的效果比我们预期的要差。然后，我们推测无序方法可能更适合该任务，并介绍了一种简单而强大的解决方案来桥接这两个研究领域。我们将视频行人重识别公式化为排名的无序集合，并使用基于图像的人的重新识别方法作为每个单独框架上的基础排名。提供理论分析以进一步保证所提出方法的性能。我们提出的方法与以前的基于递归结构的工作有很大不同，因为它完全忽略了我们的方法不依赖于视频中的时间信息。

3 问题的定义

令 $\{X^{p_i}, X^{g_i}\}_{i=1}^{N}$，为行人视频的集合，这里 X^{p_i} (测试视频) 和 X^{g_i} (训练视频) 表示两个不相交的相机 p 和 g 捕捉到的第 i^{th} 行人视频，N 是视频的数量。[1] 每个视频被定义为一组连续的帧，*i.e.*, $X^{p_i} = [\mathbf{x}_1^{p_i}, \mathbf{x}_2^{p_i}, \cdots, \mathbf{x}_{T^{p_i}}^{p_i}]$, 这里 T^{p_i} 是相机 p 捕获到的第 i^{th} 个视频的数量。为了方便起见，我们将忽略下标和上标，将通用数据点表示为 $\mathbf{x}$。同样的，通用视频输入序列由 $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$ 表示。我们认为视频帧存在于 $H \times W \times C$ 空间维度: $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^{H \times W \times C}$, 这里 H, W, 和 C 分别表示输入 $\mathbf{x}$ 的高度，宽度，和通道数。

现在，我们介绍视频行人重识别的目标和一些以后将经常使用的定义。

视频行人重识别的目标. 我们的目标是为了分类视频 X^p 和 X^g 是否来源于同一身份。这通常是通过学习映射函数 $\mathbf{g}$ 来实现的:

$$\begin{aligned} &\mathbf{d}(\mathbf{g}(X^{p_i}), \mathbf{g}(X^{g_i})) < \mathbf{d}(\mathbf{g}(X^{p_i}), \mathbf{g}(X^{g_j})), \\ &\forall j \neq i \text{ and } i,\ j = 1, 2 \cdots N, \end{aligned} \tag{1}$$

其中 $\mathbf{d}$ 表示距离测量值，例如常用的欧几里得距离。

定义. 以下网络称为典型 RNN:

$$o_{(t)} = W_{in} f(\mathbf{x}_t) + W_s r_{(t-1)}, r_{(t)} = Tanh(o_{(t)}), \tag{2}$$

这里在每个时间戳 t 的输出 o 是输入 $f(\mathbf{x}_t)$ 和 RNN 前一时刻的状态 $r_{(t-1)}$ 的线性组合。f 表示特征提取器。$\mathbf{x}_t$ 表示视频序列中的第 t^{th} 帧。一个有序的视频序列定义如下:$X^{p_i} = [\mathbf{x}_1^{p_i}, \mathbf{x}_2^{p_i}, \cdots, \mathbf{x}_{T^{p_i}}^{p_i}]$。一个无序的视频序列定义如下: $X^{p_i} = [\mathbf{x}_{\mathcal{S}(1)}^{p_i}, \mathbf{x}_{\mathcal{S}(2)}^{p_i}, \cdots, \mathbf{x}_{\mathcal{S}(T^{p_i})}^{p_i}]$, 其中 $\mathcal{S}$ 是随机排列运算符，用于随机排列（或随机排列）视频序列中的帧顺序。

1. 为简单起见，在这里我们考虑只有两个摄像机视图可用，并且扩展到更多视图很简单。

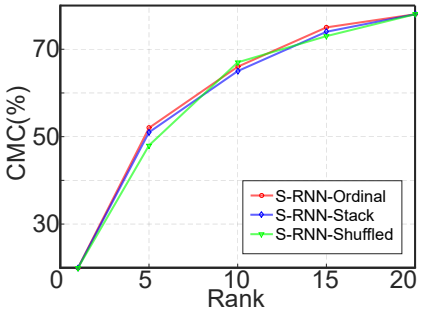


图 2. **一个典型 "CNN-RNN" 框架的诊断分析** [20]。我们考虑不带时序平均池化的单一 RNN。"S-CNN-RNN-Ordinal" 和 "S-CNN-RNN-Shuffled" 分别表示有序和打乱。"S-CNN-RNN-Stack" 表示一堆相同的帧。

	Train		Test		CMC Rank			
	Ordinal	Shuffled	Ordinal	Shuffled	1	5	10	20
[20]	✓		✓		58%	84%	91%	96%
	✓			✓	**59%**	**88%**	**94%**	**97%**
		✓	✓		58%	87%	92%	**97%**
		✓		✓	58%	87%	92%	**97%**
RNN	✓		✓		46%	74%	85%	94%
	✓			✓	46%	74%	86%	94%
		✓	✓		54%	**78%**	**88%**	**95%**
		✓		✓	**56%**	**78%**	87%	94%
LSTM	✓		✓		41%	68%	82%	91%
	✓			✓	44%	72%	84%	91%
		✓	✓		49%	73%	85%	92%
		✓		✓	**52%**	**76%**	**86%**	**93%**
GRU	✓		✓		44%	74%	83%	93%
	✓			✓	46%	**76%**	85%	92%
		✓	✓		51%	74%	86%	93%
		✓		✓	**54%**	74%	**87%**	**94%**

图 3. **不同 "CNN-RNN" 方法的详细结果**在不同训练和测试的设定下。对于 [20] 的结果，我们遵循原始协议，并使用具有原始顺序视频序列的预先计算的光流。对于其余方法，禁用光流输入。

4 "CNN-RNN" 的诊断分析

作为概念上的证明，首先为视频行人重识别提供了对递归网络的诊断分析。我们的分析是基于 [20] 中介绍的方法进行的，该方法已被其许多后续工作广泛采用 [6], [21]–[24], [32]。为了方便起见，我们将省略 X 的上标来代表通用摄像机视图中的视频序列。

观察 1. [20] 中完成的典型 CNN-RNN 方法无法有效地捕获期望的时间依赖性，并隐式导致了无序表示。

我们通过评估（没有重新训练）有序和无序视频序列上的 "CNN-RNN" 体系结构来说明这一点。结果在 Fig. 3中。令人惊讶的是，常用的 "CNN-RNN" 体系结构在有序视频序列上表现更好。这表明视频帧中的顺序依赖假设可能对视频行人重识别无效，因此 RNN 会忘记以前的状态。在这种情况下，在无序情况下会有更多的输入图像。例如，对于一个 T 帧的视频序列，RNN 总共收到 t 帧作为输入，这里在无序的情境中存在 C_T^t 输入方法，而在有序的情境中只有 $T-t+1$ 输入的方法。

我们进一步展示了 [20] 中一个单个 RNN 在各种输入场景（例如有序，无序和堆叠同一帧）中导致的性能差异可忽略不计。更具体地说，我们删除时间池，并在 [20] 中获取 RNN 的最后输出，以进一步了解 RNN 在视频行人重识别中的优点。在最初的两个场景中，我们输入一个有序和一个无序的视频帧（128 视频帧，如 [20] 所做的那样）到 RNN 中。在第三个场景中，我们重复地堆叠选定视频中随机的一帧 128 次来生成一个假的视频。所有实验都重复了 10 次，结果在 Fig. 2中。可以看出，RNN 导致几乎相同的结果，这意味着 RNN 以前的所有输入，无论是有序还是无序的，甚至是完全相同的，都不那么重要。有人可能会说忘记它的先前状态可能是 RNN 的算法缺陷。我们凭经验发现，其他高级学习方法（例如 LSTM）也可以得到类似的结果。有关更详细的结果，请参阅 Fig. 3。

观察 2 . 当在随机打乱的视频序列上训练时，相同的 CNN-RNN 方法的效果与在有序视频序列上训练时的效果相当。为了进一步了解 "CNN-RNN" 的优点并阐明我们的动机，我们对 "CNN-RNN" [20] 系统进行了重新训练，但随机打乱每个 mini-batch 中的输入视频。我们使用开源的实现 [20]。结果总结在了 Fig. 3中。我们还包括以上分析的结果，以进行更详细的比较。在最初的设定中，我们包括了 [20] 的结果。在这个设定中，"CNN-RNN" 在原始的有序视频序列上进行训练，并在有序的视频序列上进行评估。第二种设定是 "常规训练，打乱后的训练" 总结了我们之前的研究。也就是说，我们在原始顺序视频序列上训练 "CNN-RNN"，并在随机打乱的视频序列上对其进行评估，在最后两行中，我们在训练阶段随机打乱每个时期的视频序列。之后，我们在随机打乱的视频和原始顺序序列上测试网络。我们的结果表明，RNN 无法按预期捕获时间依赖性。

有人可能会说 [20] 中使用的光流在这些情况下可能起着至关重要的作用。为了对此进行研究，我们进行了另一组实验，其中完全禁用了光流输入。此外，还研究了更高级的递归结构，例如 GRU 和 LSTM。我们观察到的是：i）光流对于此任务是有益的，尤其是在按序序列训练系统时。但是，当在打乱的序列上训练系统时，此优点变得不那么明显。ii）普遍认为对视频行人重识别的时间信息进行编码的帧顺序对此处研究的所有三个循环结构均无益。有关统计显着性检验的更多观察结果和详细分析，请参阅补充文件。

5 提出的解决方案

5.1 集合排序器

人们普遍认为，以帧顺序编码的时间信息对于视频行人重识别至关重要，因此，社区在很大程度上已将使用循环结构视为一种良好做法，这是理所当然的。第一次，我们深入研究了为视频行人重识别任务建模时间信息的效果。我们的分析表明，常用的递归结构可能不如我们期望的那样捕获视频行人重识别任务的顺序依存关系。实际上，它本质上是学习无序表示形式，我们认为这可能与视频行人重识别任务更相关，

通过使用 RNN 输出在不同时间步的时间池中的多个特征表示。此外，在每个时间步长中，RNN 的输出均由当前输入特征控制。

因此，我们认为以无序的方式明确地解决视频行人重识别任务是有益的。这进一步激励我们从不同角度解决视频行人重识别任务。通过将视频行人重识别任务视为无序整体排名的任务，我们提出了一种简单而强大的解决方案，其中，每个基础排名都由具有单个图像帧的人员重新标识来体现。可以通过 Kemeny-Young 方法 [47] 汇总多个排名结果。从整体学习的角度来看 [48]，强度和多样性都是必不可少的，在我们的案例中，它们分别由基于图像的人员重新识别方法的准确性和视频中图像帧的差异来保证。与传统方法中的视频对相比，我们的方法可提供更多种类的图像对。通过在 KemenyYoung 方法下提供更详细的排名，所有这些图像对均有助于最终 re-id 准确性，并且它们在我们的设置中都经过了良好的训练。通过两个例子 Sec. 5.2，我们表明，所提出的解决方案可以轻松地与任何基于图像的人员重新识别方法集成。

基础原理. 更正式地，对于每个在测试集中的输入视频 $X^{p_i} = [\mathbf{x}_1^{p_i}, \mathbf{x}_2^{p_i}, \cdots, \mathbf{x}_{T^{p_i}}^{p_i}]$，我们解决方案返回一个排名列表集合，由 T^{p_i} 基础列表 $R = \{R_t\}_{t=1}^{T^{p_i}}$ 组成，这里排名 R_t 由基础排序器 R 生成。基础排序器最终的目标是:

$$\begin{aligned} &\mathbf{d}(R(\mathbf{x}_t^{p_i}), R(\mathbf{x}_{t'}^{g_i})) < \mathbf{d}(R(\mathbf{x}_t^{p_i}), R(\mathbf{x}_{t''}^{g_j}))), \\ &\forall j \neq i, \forall t \in \{1, 2, \cdots, T^{p_i}\}, \\ &\forall t' \in \{1, 2, \cdots, T^{g_i}\}, \forall t'' \in \{1, 2, \cdots, T^{g_j}\}, \\ &i, j = 1, 2, \cdots, N, \end{aligned} \tag{3}$$

如 Eqn. (1) 中定义的那样，$\mathbf{d}$ 表示距离测量值，例如常用的欧几里得距离。

与采用不同基础模型的常规集成学习方法不同，此处一个视频的每个基础排名均由接收不同帧的相同重新识别网络体现。[2]

定义第 i^{th} 个询问视频 X^{p_i} 的第 t^{th} 张图像 $\mathbf{x}_t^{p_i}$ 和第 j^{th} 个测试集合 X^{g_j} 的距离如下:

$$\begin{aligned} &d_t^{ij} = \min(d(R(\mathbf{x}_t^{p_i}),\ R(\mathbf{x}_{t'}^{g_j})))), \\ &\forall t' \in 1, 2 \cdots T^{g_j}\ and\ i, j = 1, 2 \cdots N, \end{aligned} \tag{4}$$

第 i^{th} 个身份基础排名列表能够通过如下式子得到:

$$R_t^i = \arg sort([d_t^{i1}, d_t^{i2}, \cdots, d_t^{iN}]), \tag{5}$$

这里 $\arg sort$ 是排序运算符。在将每个视频的距离与 i^{th} 视频（$\mathbf{x}_t^{p_i}$）中的 t^{th} 图像进行排名后，返回每个视频的索引。在不失一般性的前提下，我们假设视频的排名随着距离的增加而增加，即，距离最小的视频获得的排名最低。Eqn. (5) 中的每个基础排名列表实际上是对 $\{1, 2, \cdots, N\}$ 的排列。此外，每个基础排名列表 R_t^i 由 Sec. 5.2中介绍的不同基础排序器 R 返回。

2. 在我们看来，使用常规集成方法中使用的不同排序器的说服力较弱，因为这会大大增加来自多个不同模型的模型复杂性。

对于每对视频 (X^{p_i}, X^{g_j}) $(i, j = 1, 2 \cdots N)$，用 $M_{i,j}$ 表示，R_t^i 的倍数 i 排在 j 的前面，如下表示:

$$M_{i,j} = \sum_{t=1}^{T^{p_i}} I(R_t^i(i) < R_t^i(j)), \tag{6}$$

这里 I 是一个指标函数 ($I(x) = 1$ 当且仅当 $x == True$)。定义 $\mathbf{Y}_{i,j}$ 为真当且仅当 $M_{i,j} > \frac{T^{p_i}}{2}$。此外，假设 i^{th} 视频的真实排名列表是 $\hat{R}^i$，然后根据 Hoeffding 的不等式 [49]，我们有如下结果:

提议 1. 令 $r = Pr((R_t^i(i) < R_t^i(j)) \neq (\hat{R}^i(i) < \hat{R}^i(j))$ 表示每个基础排序器的错误率 R_t^i。假设 $r < \frac{1}{2} - \varepsilon$, $(0 < \varepsilon < \frac{1}{2})$，在独立同分布假设下，我们有:

$$Pr(\mathbf{Y}_{i,j} = False | (\hat{R}^i(i) < \hat{R}^i(j)) \leq e^{-2\varepsilon^2 T^{p_i}}. \tag{7}$$

尽管对于霍夫丁不等式的独立同分布假设可能在实践中不成立，但它可以对视频行人重识别进行分析研究，并阐明了如何改进它。

一致性讨论. 对于第 i^{th} 训练视频 X^{g_i}，为了汇总每个基础排序器 $R_t^i, t = 1, 2, \cdots, T^{g_i}$ 的多个结果，变量 $\mathbf{Y}_{i,\,j}$ 必须定义传递关系，这就是说 $\mathbf{Y}_{i,\,k}$ 为真如果 $\mathbf{Y}_{i,\,j}$ 和 $\mathbf{Y}_{j,\,k}$ 都为真。当对所有的 i, j, k 都成立的时候，基础排序器 $R_t^i, t = 1, 2, \cdots, T^{g_i}$ 被认为是连续的 [50]。在 [50] 中的理论分析表明在同样的条件下，$R_t^i, t = 1, 2, \cdots, T^{g_i}$ 与概率 1 一致当 T^{g_i} 变向无穷大的时候。这可以保证我们使用 Kemeny–Young 方法 [47] 来汇总最终排名列表。具体来说，Kemeny–Young 方法使用每个基础排序器 $R_t^i, t = 1, 2, \cdots, T^{g_i}$ 优先投票的方法。我们先创建一个 0-1 二进制矩阵 $\mathcal{M} \in \{0, 1\}^{N \times N \times T^{p_i}}$，这里 $\mathcal{M}_{ijt} = 1$ $(i, j = 1, 2 \cdots N$ 和 $t = 1, 2 \cdots T^{g_i})$ 当且仅当 $R_t^i(i) < R_t^i(j)$。否则，$\mathcal{M}_{ijt} = 0$。然后沿着 2nd 和 3th 维度求和，返回向量 $\mathcal{M}_{i\cdot\cdot}$，其中 k^{th} 条目 $\mathcal{M}_{i\cdot\cdot}(k)$ 总结了库集中的 k^{th} 标识在所有基本排名中受到青睐的频率。通过以递减的方式对向量 $\mathcal{M}_{i\cdot\cdot}(k)$ 进行排名，可以获得最终的排名列表。

5.2 基础排序器

通常，我们的解决方案可广泛应用于任何图像行人重识别方法。但是，如建议 2 中的 Eqn. (7) 所示，在独立同分布假设下，最终集合的误差范围随 ε 的平方呈指数下降。因此，在实践中使用性能最佳的基础排序器将是有益的。受这个启发，我们着手进行基于最新图像行人重识别方法。为了说明我们的观点，我们介绍了我们提出的方法的两个实现：*EnsembleRanker-RW* 和 *EnsembleRanker-CRF*。前者在 [41] 中使用基于随机游走的方法作为基础排序器，而后者在 [42] 中使用基于 CRF 的方法。我们选择这两种方法，是因为它们在基于图像的人的身份识别方面具有优越的性能。下面简要介绍了他们的主要技术，感兴趣的读者可以参考原始文章以获取更多详细信息。

Ensemble-RW. 该方法使用 [41] 中的技术训练行人重识别模型。更具体地说，它采用新颖的分组改组随机游走操作，以

充分利用训练图像之间的亲和力来细化测试图像与训练图像之间的亲和力。它将随机游走操作集成到深度神经网络的训练过程中。除了真实的标识标签外，还可以通过随机游走操作对特征进行分组和改组，从而提供更丰富的监管。在 CUHK03，Market-1501 和 DukeMTMC 上，它的 top-1 准确性分别为 91.5，92.7 和 80.7 [41]。

Ensemble-CRF. 此方法采用 [42] 中的技术。它通过将 CRF 模型与深度神经网络结合使用新颖的相似性学习方法进行人的重新识别。它通过统一的图形模型对组中图像之间的相似度进行建模，并借助组相似度以多尺度的方式学习局部相似度。考虑到更多的图像间关系，据报告，在挑战性场景下，学习到的相似性度量更加健壮并与图像一致。该方法在 CUHK03，Market-1501 和 DukeMTMC 上分别达到 90.2，93.5 和 84.9 的 top-1 精度，[42]。

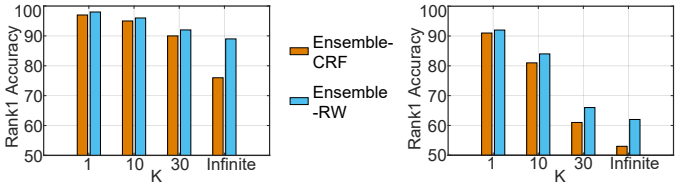
5.3 与传统集成学习的区别

现存的集成行人重识别方法 [51], [52] 主要来自于图像行人重识别的角度。[51] 中的工作训练了集成的 RankSVM，而 [52] 中的方法则使用多个功能来获取多个指标并通过加权平均值进行组合。相比之下，我们研究了在深度学习的框架下针对视频行人重识别进行集成学习的可行性，其中分别通过共享的基本排名和 KemenyYoung 方法保证了效率和鲁棒性。更重要的是，我们从经验和理论上证明了该方法的有效性。其他传统的集成方法 [48]，通常使用不同的基本模型并产生更大的模型复杂度，因此在这里进行研究的兴趣不大。

6 实验结果

在本节中，我们将用三种具有挑战性的基准数据集评估我们提出的视频行人重识别方法：iLIDS-VID [7], PRID-2011 [4], and MARS [8]。iLIDS-VID 数据集包含 300 人，其中每个人由不重叠摄像机捕获的两个视频序列表示。PRID-2011 数据集包含 749 人，由两个不重叠的相机捕获。按照在 [7] 中使用的协议，选择具有超过 21 帧的序列，从而获得 178 身份。对于 iLIDS-VID 和 PRID-2011，每个数据集被随机分为 50% 的人员进行训练和 50% 的人员进行测试。MARS 数据集是最大的基于视频的人员重新识别基准，具有 1,261 身份和 DPM 检测器 [53] 和 GMMCP tracker [54] 生成的约 20,000 视频序列。2 相机，平均有 13.2 序列。数据集中有 3,248 个干扰物序列。

对于 Ensemble-RW，使用随机梯度下降法以 40 epoch 训练网络，学习率为 $1e-6$，batch 为 32。对于 Ensemble-CRF，使用随机梯度下降法以 100epoch 训练网络，学习率为 $1e-2$，batch 为 132。其他超参数设置为与原始参数相同的值。我们使用两个 Nvidia TiTan GPU 训练网络。对于前两个数据集，在测试过程中，我们将第一个摄像机视为测试，将第二个摄像机视为训练 [7]。对于 MARS，我们在 [8] 中遵循完全相同的评估协议。测试序列通过网络获得特征向量，并使用欧氏距离



(a) PRID-2011 (b) iLIDS-VID

图 4. 不同 K 时，Ensemble-RW 和 Ensemble-CRF 的 CMC Rank 1 准确性。我们展示了采样率如何影响解决方案的准确性，我们展示了 $K = 1, 10, 30$ 和 $Infinite$ 的结果，其中 $K = Infinite$ 表示仅对视频中的第一张图像进行采样的情况。实验结果表明，精度随着 K 的减小而增加。

和来自所有训练序列的预先计算的特征向量进行排序。然后，我们使用 $KemenyYoung$ 方法将结果汇总为最终排名列表。

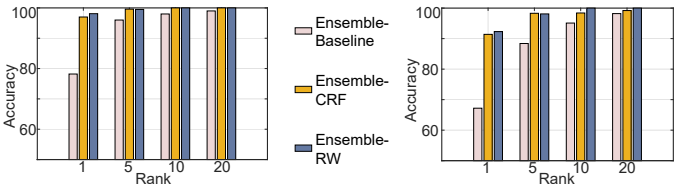
6.1 消融研究

我们首先提供一些消融研究，以研究 Eqn. (7) 中的两个关键方面。我们在 iLIDS-VID 和 PRID-2011 上进行了这些实验。

T 的影响：. 我们的理论分析表明，在一定条件下，最终系统的误差范围随图像帧数呈指数下降。但是，实际上，由于存储，传输和计算资源的限制，我们无法利用视频序列的无限长度。为了研究视频长度的影响，我们展示 Ensemble-EW 和 Ensemble-CRF 在 PRID-2011 和 iLIDS-VID 数据集上的结果。更具体地说，我们从每个视频序列中的每个 K 图像中采样一个图像。我们设置 $K = Infinite$ 来确保 $K \geq T$，在这种情况下，我们使用第一帧来进行图像行人重识别实验，就像在 [8] 中所做的那样。当 $K = 1$ 时，我们使用所有可能的帧。结果如 Fig. 4所示。显然，K 对最终性能有反作用，即，精度随 K 的减少而增加。

关于“从测试到训练”的模式，行人重识别主要可以分为两种策略：图像到图像和视频到视频。第一种模式是在文献中主要被研究的，而直到最近，我们已经研究过视频到视频的行人重识别，其他情况（例如，图像到视频和视频到图像）并不常见，可以将其视为视频到视频的特例，我们的发现显然证明了这一点。人们普遍认为视频到视频模式（也是本文的重点）会更有利，因为测试单元和训练单元都包含比单个图像更丰富的视觉信息。在下面的章节会进一步说明其它普遍实际的情况（例如视频中的动态时序建模）可能无助于视频行人重识别提高性能。

ε 的影响:. 从另一个角度来看，Eqn. (7) 中的错误界限显示出提高了最终准确性，即减少了每个基础排序器的错误。为了说明这一点，我们在 [41] 中训练了 baseline 模型。此方法不使用打乱洗随机游走操作，因此无法完善测试图像和训练图像之间的亲和力。在 [41] 中使用的所有数据集上，此版本的性能都较差。我们称此方法为“Ensemble-Baseline”，其结果总结在 Fig. 5中，很明显，在所有情况下，使用更强的基础排序器可以带来更好的总体准确性。



(a) PRID-2011 (b) iLIDS-VID

图 5. 对比不同的基础排序器我们对比了三个不同排序器的结果。[41] 中的 baseline 方法禁用了细化测试图像和训练图像之间的亲和力的功能，已集成到我们的系统中，称为 "Ensemble-Basline"。我们在两个数据集上的结果表明，基础排序器更强表现更好。

	Rank 1	Rank 5	Rank 20	MAP
Liu *et al.* [55]	68.3	81.4	90.6	52.9
Zheng *et al.* [8]	68.3	82.6	89.4	49.3
Zhou *et al.* [24]	70.6	90.0	97.6	50.7
Chen *et al.* [22]	71.0	89.0	96.0	-
Zhang *et al.* [44]	71.2	85.7	94.3	-
Li *et al.* [56]	82.3	-	-	65.8
Wu *et al.* [57]	80.8	92.1	96.1	67.4
Ensemble-CRF	85.3	**95.0**	97.1	79.3
Ensemble-RW	**86.8**	94.6	**97.5**	**80.1**
Ensemble-CRF (MC)	85.1	94.8	97.0	79.2
Ensemble-RW (MC)	86.3	94.4	97.4	**80.1**

图 6. 在 MARS 上对比其它先进方法。

6.2 对比其它先进方法

改进结果的重要工作如下：1) McLaughlin *et al.* [20] 使用时间平均池在每个时间点汇总 RNN 输出；2) Chen *et al.* [22] 采用了 [20] 和一个类似的网络结构，但是融合了 CNN 和 RNN 特征；3) Xu *et al.* [23] 使用了在 CNN 中空间金字塔池化和产生更有判别性特征的注意力模型；4) zhang *et al.* [32] 聚合了 CNN 和双向 RNN；5) Zhou *et al.* [24] 既使用了时序注意力模型，循环单元，又使用了更复杂的三元损失函数；6) Liu *et al.* 的使用 AMOC 网络的方法使用两流卷积体系结构从相邻帧的集合中共同学习外观表示和运动上下文。7) Li *et al.* [56] 使用空间注意力集合来发现相同的身体部位。8) Zhang *et al.* [57] 该方法采用可解释的强化学习训练代理来每次验证一对图像。从 Fig. 7中，我们可以看到，所提出的方法均优于 "CNN-RNN" 及其几个先进的扩展。Fig. 6将提出的方法与 MARS 上的其他最新方法进行了比较。遵循标准评估协议，我们将 CMC 和 MAP 都用作评估指标 [8]。根据排名结果计算平均精度（AP），并在所有查询中计算平均平均精度（mAP），这被视为最终的重新识别精度。为了便于演示，我们忽略了 CMC 排名 15 的结果。同样，我们提出的解决方案在 CMC 排名和 MAP 得分方面优于所有这些方法。此外，我们还包括该方法的其他变体，它们使用交叉熵蒙特卡罗算法 [64] 来聚合多个排序器。我们观察到，使用其他聚合方法，我们提出的解决方案在 CMC 排名和 MAP 得分方面仍然优于所有这些方法。

这里我们也提供了一些公开的讨论。 Eqn. (7) 提出了一种提高行人重识别性能的好方法：在所有可用的视频帧上使用最佳的图像行人重识别方法。但是，这是不平凡的，值得就以下问题进行未来的研究工作：1) 如何满足独立同分布假设？并且 2) 给定有限的计算预算，如何在 T 和 ε 之间实现良好的折衷？这些可以在各个方面进行研究：(1) 增强的数据集处理能力，其中可以减少各个帧之间的依赖性，例如随机帧跳过或关键帧选择 [65]。但是，这也减少了视频长度。(2) 使用更多计算密集型方法有利于提升 ε。

7 结论

我们首先对视频行人重识别的常用 "CNN-RNN" 方法进行诊断分析。我们证明文献中使用的 RNN 通常可能无效来捕获时间依存关系并隐式学习无序表示，我们认为这与视频行人重识别更为相关，然后基于此观察结果，提出了一种简单而又功能强大的视频行人重识集成学习方法；此外，我们从理论上证明，这两种方法都采用了图像行人重识别方法和较长的视频序列，我们通过两个示例来证明这一想法，并且在多个广泛使用的数据集上，我们提出的解决方案明显优于现有的最新技术方法。

参考文献

[1] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *IEEE CVPR*, 2006.

[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE CVPR*, 2010.

[3] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1622–1634, 2013.

[4] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person reidentification by descriptive and discriminative classification," in *SCIA*, 2011.

[5] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *Eur. Conf. Comput. Vis.*, 2012.

[6] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person reidentification via recurrent feature aggregation," in *Eur. Conf. Comput. Vis.*, 2016.

[7] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person reidentification by video ranking," in *Eur. Conf. Comput. Vis.*, 2014.

[8] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person reidentification," in *Eur. Conf. Comput. Vis.*, 2016.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE CVPR*, 2015.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2014.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.

	PRID-2011				iLIDS-VID			
	Rank 1	Rank 5	Rank 10	Rank 20	Rank 1	Rank 5	Rank 10	Rank 20
Karanam *et al.* [58]	35.1	59.4	69.8	79.7	24.9	44.5	55.6	66.2
Karanam *et al.* [59]	40.6	69.7	77.8	85.6	25.9	48.2	57.3	68.9
Wang *et al.* [7]	41.7	64.5	77.5	88.8	34.5	56.7	67.5	77.5
Li *et al.* [60]	43.0	72.7	84.6	91.9	37.5	62.7	73.0	81.8
Wu *et al.* [21]	49.8	77.4	90.7	94.6	42.6	70.2	86.4	92.3
Li *et al.* [33]	51.6	83.1	91.0	95.5	34.5	63.3	74.5	84.4
Yan *et al.* [6]	58.2	85.8	93.4	97.9	49.3	76.8	85.3	90.0
Liu *et al.* [61]	64.1	87.3	89.9	92.0	44.3	71.7	83.7	91.7
McLaughlin *et al.* [20]	70.0	90.0	95.0	97.0	58.0	84.0	91.0	96.0
Zhang *et al.* [32]	72.8	92.0	95.1	97.6	55.3	85.0	91.7	95.1
Chen *et al.* [22]	77.0	93.0	95.0	98.0	61.0	85.0	94.0	97.0
Xu *et al.* [23]	77.0	95.0	99.0	99.0	62.0	86.0	94.0	98.0
Zhou *et al.* [24]	79.4	94.4	-	99.3	55.2	86.5	-	97.0
Liu *et al.* [55]	83.7	98.3	99.4	**100.0**	68.7	94.3	98.3	99.3
Zhang *et al.* [44]	85.2	97.1	98.9	99.6	60.2	84.7	91.7	95.2
Liu *et al.* [62]	90.3	98.2	99.3	100.0	68.0	86.8	95.4	97.4
Khan *et al.* [63]	92.5	99.3	**100.0**	**100.0**	79.5	95.1	97.6	99.1
Li *et al.* [56]	93.2	-	-	-	80.2	-	-	-
Ensemble-CRF	95.5	**99.6**	**100.0**	**100.0**	90.4	**98.3**	98.4	99.2
Ensemble-RW	**97.1**	99.5	**100.0**	**100.0**	**91.3**	98.1	**100.0**	**100.0**
Ensemble-CRF (MC)	94.7	98.9	**100.0**	**100.0**	89.7	97.7	97.6	99.1
Ensemble-RW (MC)	96.8	99.3	**100.0**	**100.0**	**91.3**	97.9	**100.0**	**100.0**

图 7. 在 PRID-2011 和 iLIDS-VID 上对比其它先进方法。

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE CVPR*, 2015.

[14] Y. Liu, P.-T. Jiang, V. Petrosyan, S.-J. Li, J. Bian, L. Zhang, and M.-M. Cheng, "DEL: Deep embedding learning for efficient image segmentation," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 864–870.

[15] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," in *Adv. Neural Inform. Process. Syst.*, 2019, pp. 35–45.

[16] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *IEEE CVPR*, 2015.

[17] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A uni-fied embedding for face recognition and clustering," in *IEEE CVPR*, 2015.

[18] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M.-M. Cheng, and G. Zheng, "Crowd counting with deep negative correlation learning," in *IEEE CVPR*, 2018.

[19] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *IEEE CVPR*, 2014.

[20] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *IEEE CVPR*, 2016.

[21] L. Wu, C. Shen, and A. v. d. Hengel, "Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach," *arXiv preprint arXiv:1606.01609*, 2016.

[22] L. Chen, H. Yang, J. Zhu, Q. Zhou, S. Wu, and Z. Gao, "Deep spatial-temporal fusion network for video-based person re-identification," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2017.

[23] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Int. Conf. Comput. Vis.*, 2017.

[24] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification„" in *IEEE CVPR*, 2017.

[25] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in *IEEE CVPR*, 2018.

[26] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *IEEE CVPR*, 2015.

[27] M. A. Goodale and A. D. Milner, "Separate visual pathways for perception and action," *Trends in neurosciences*, vol. 15, no. 1, pp. 20–25, 1992.

[28] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Eur. Conf. Comput. Vis.*, 2016.

[29] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "Actionvlad: Learning spatio-temporal aggregation for action classification," *IEEE CVPR*, 2017.

[30] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao, "A key volume mining deep framework for action recognition," in *IEEE CVPR*, 2016.

[31] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *IEEE CVPR*, 2015.

[32] W. Zhang, X. Yu, and X. He, "Learning bidirectional temporal cues for video-based person re-identification," *IEEE TCSVT*, vol. 28, no. 10, pp. 2768–2776, 2017.

[33] Y. Li, L. Zhuo, J. Li, J. Zhang, X. Liang, and Q. Tian, "Video-based person re-identification by deep feature guided pooling," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2017.

[34] X. Su, Y. Zou, Y. Cheng, S. Xu, M. Yu, and P. Zhou, "Spatial-temporal synergic residual learning for video person re-identification," *arXiv preprint arXiv:1807.05799*, 2018.

[35] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *IEEE CVPR*, 2015.

[36] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *IEEE CVPR*, 2016, pp. 1249–1258.

[37] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[38] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *IEEE CVPR*, 2016.

[39] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification." in *Int. Conf. Comput. Vis.*, 2017.

[40] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *IEEE CVPR*, 2017.

[41] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang, "Deep group-shuffling random walk for person re-identification," in *IEEE CVPR*, 2018.

[42] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep crf for person re-identification," in *IEEE CVPR*, 2018.

[43] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *IEEE CVPR*, 2018, pp. 1179–1188.

[44] J. Zhang, N. Wang, and L. Zhang, "Multi-shot pedestrian re-identification via sequential decision making," in *IEEE CVPR*, 2018, pp. 6781–6789.

[45] J.-B. Boin, A. Araujo, and B. Girod, "Recurrent neural networks for person re-identification revisited," in *ICMIPR*. IEEE, 2019, pp. 147–152.

[46] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *IEEE CVPR*, 2018, pp. 5363–5372.

[47] J. Levin and B. Nalebuff, "An introduction to vote-counting schemes," *Journal of Economic Perspectives*, vol. 9, no. 1, pp. 3–26, 1995.

[48] Y. Ren, L. Zhang, and P. N. Suganthan, "Ensemble classification and regression-recent developments, applications and future directions," *IEEE CIM*, vol. 11, no. 1, pp. 41–53, 2016.

[49] R. J. Serfling, "Probability inequalities for the sum in sampling without replacement," *The Annals of Statistics*, pp. 39–48, 1974.

[50] K. Jong, J. Mary, A. Cornuéjols, E. Marchiori, and M. Sebag, "Ensemble feature ranking," in *PKDD*. Springer, 2004, pp. 267–278.

[51] B. J. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking." in *Brit. Mach. Vis. Conf.*, vol. 2, no. 5, 2010, p. 6.

[52] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel, "Learning to rank in person re-identification with metric ensembles," in *CVR*, 2015, pp. 1846–1855.

[53] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[54] A. Dehghan, S. Modiri Assari, and M. Shah, "Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *IEEE CVPR*, 2015.

[55] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng, "Video-based person re-identification with accumulative motion context," *IEEE TCSVT*, vol. 28, no. 10, pp. 2788–2802, 2018.

[56] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *IEEE CVPR*, 2018.

[57] 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 *IEEE CVPR*, 2018, pp. 5177–5186.

[58] S. Karanam, Y. Li, and R. J. Radke, "Sparse re-id: Block sparsity for person re-identification," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2015.

[59] ——, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *Int. Conf. Comput. Vis.*, 2015.

[60] Y. Li, Z. Wu, S. Karanam, and R. J. Radke, "Multi-shot human re-identification using adaptive fisher discriminant analysis." in *Brit. Mach. Vis. Conf.*, 2015.

[61] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for viceo-based pedestrian re-identification," in *Int. Conf. Comput. Vis.*, 2015.

[62] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *IEEE CVPR*, 2017.

[63] F. M. Khan and F. Brèmond, "Multi-shot person re-identification using part appearance mixture," in *WACV*, 2017.

[64] V. Pihur, S. Datta, and S. Datta, "Rankaggreg, an r package for weighted rank aggregation," *BMC bioinformatics*, vol. 10, no. 1, p. 62, 2009.

[65] J. Meng, H. Wang, J. Yuan, and Y.-P. Tan, "From keyframes to key objects: Video summarization by representative object proposal selection," in *IEEE CVPR*, 2016, pp. 1039–1048.