第 59 卷 第 2 期/2022 年 1 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

# 多摄像机视场下基于一种DTN的多人脸实时跟踪系统

任国印[1,2]，吕晓琪[1,2,3*]，李宇豪[2]

[1]内蒙古科技大学机械工程学院，内蒙古 包头 014010；

[2]内蒙古科技大学信息工程学院，内蒙古 包头 014010；

[3]内蒙古工业大学，内蒙古 呼和浩特 010051

**摘要** 为了实现跨摄像机区域的多人脸图像跟踪，提出了一种基于双三分支孪生网络(DTN)的跨摄像机跟踪网络。具体方法是应用Chinese Whisper(CW)人脸聚类算法对同一行人的人脸图像进行聚类，并根据聚类结果通过智能监控确定捕获的目标人脸。通过改进FaceNet的网络结构和训练函数，实现了行人面部的精确跟踪。在LFW数据集上训练DTN后，通过边缘样本挖掘损失(MSML)和焦点损失难样本平衡训练，人脸识别率可以提高到99.51%。实验结果表明：通过比较同一视频监控场内人脸特征的相似性，所提网络可以通过该区域跟踪行人的人脸目标；通过摄像机间人脸特征的实时传输，实现了跨摄像机的人脸跟踪。

**关键词** 图像处理；人脸聚类；双三分支孪生网络；人脸跟踪；跨摄像机；人脸识别

**中图分类号** TP391.4 **文献标志码** **doi**：10.3788/LOP202259.0210004

# Multi Face Real-Time Tracking System Based on DTN in Multi Camera Field of View

Ren Guoyin[1,2], Lü Xiaoqi[1,2,3*], Li Yuhao[2]

[1]*School of Mechanical Engineering, Inner Mongolia University of Science & Technology, Baotou, Inner Mongolia* 014010, *China*;

[2]*School of Information Engineering, Inner Mongolia University of Science & Technology, Baotou, Inner Mongolia* 014010, *China*;

[3]*Inner Mongolia University of Technology, Hohhot, Inner Mongolia* 010051, *China*

**Abstract** In order to realize multi face image tracking across camera regions, a cross camera tracking network based on double three branch twin network (DTN) is proposed. The specific method is to apply Chinese Whisper (CW) face clustering algorithm to cluster the face images of the same pedestrian, and determine the captured target face through intelligent monitoring according to the results of face clustering. By improving the network structure and training function of FaceNet, pedestrian face tracking is realized accurately. After training DTN on LFW data set, the face recognition rate can be improved to 99.51% through margin sample mining loss (MSML) and focus loss difficult sample balance training. Experimental results show that by comparing the similarity of face features in the same video surveillance field, the proposed network can track pedestrian face targets through this area; through the real-time transmission of face features between cameras, cross camera face tracking is realized.

**Key words** image processing; face clustering; double three branch twin network; face tracking; cross camera; face recognition

**收稿日期**：2021-01-08；**修回日期**：2021-02-07；**录用日期**：2021-03-09

**基金项目**：国家自然科学基金（61771266，81571753）、包头市青年创新人才项目（0701011904）

**通信作者**：*635302395@qq.com

## 1 引 言

目前，深度学习主要应用于人脸识别、行人跟踪、行为分析、交通统计和动作识别等领域[1-3]。改进的深度学习策略提高了人脸跟踪的精度和鲁棒性。然而，许多跟踪网络不能跨摄像机使用，单个监控摄像机的视野非常有限。

多智能监控系统的人脸跟踪已经成为一个研究热点[4-8]。多摄像机人脸跟踪的目的是在不同的摄像机视野下识别出同一行人的人脸，并用相同的id标记出这些人脸，根据人脸图像的特征距离计算出人脸图像的特征相似度。跨摄像机人脸跟踪在监控领域中占有重要的地位。例如，从大范围的海量视频中寻找目标行人的轨迹，该技术可以取代效率低下的人工搜索过程，可以用来搜寻失踪人员或追踪嫌疑人。

在视角不同、行人面部过小、光照条件不同和背景混乱的情况下，对同一行人的面部使用交叉摄像机进行跟踪仍然是一项具有挑战性的任务。为扩大监控范围，多台智能监控联合识别已经成为一个研究热点，许多研究人员在这方面做了大量的工作。Tian等[9]提出了一种多摄像机行人计数方法，该方法通过结合位置信息和时间信息的相关性解决了摄像机之间重复统计的问题。Lin等[10]提出了一种多摄像机下的行人跟踪方法，该方法通过识别车道标志线来检测视场边缘线，解决了多条轨道之间的衔接问题，且不需要特征匹配。

Elhamifar等[11-12]提出了一种基于Chinese Whisper(CW)的人脸聚类算法。CW算法可以在没有指定聚类中心的情况下，以较快的速度完成大数据量图像的人脸聚类。Zhang等[13]提出了一种将双三分支孪生网络(DTN)和CW人脸聚类算法相结合的FaceNet网络，该网络将学习直接嵌入欧氏空间进行人脸识别，优点是可以将人脸图像对齐后输入到网络中，事先将除人脸外的图像内容去掉，这样做可大大降低干扰信息的过拟合。该网络仅使用128维人脸特征，人脸识别率高达99.63%。林增敏等[4]和周晨辰[8]分别用CW算法完成了人脸图像去重和基于图聚类的人脸识别。Schroff等[14]提出了一个深度学习的框架，进一步研究了人脸自适应检测与聚类相结合的方法。结合视频中的人脸特征，通过深度学习得到人脸特征向量的聚类，通过实验可以发现对于背景变化频繁的视频，人脸聚类是有效的。在一个固定的场景中，人脸存在多个重复帧，首先要解决的问题是聚类算法在同一视场中共采集到哪些人的人脸。然而就目前的聚类算法而言，事先未给出类数的聚类算法很少，例如K-means算法、GMM算法必须在聚类前指定类数，但实时视频监控中的人脸类别是未知的，虽然MSC算法和DBS算法可以不必预先确定类别，但需要指定类簇窗口的半径。然而视频监控图像的视频帧随着拍摄不断增多，因此很难预测类簇窗口的大小[15-18]。HAC算法是一种自下而上的聚类算法[19-20]，该算法将每个数据点视为一个簇，然后计算所有簇之间的距离，直到所有簇合并成最终的簇。虽然这种聚类算法不需要预先指定类别，但利用比较像素值或直方图的方法来判断两幅图像之间的距离，效率很低。

本文提出了一种基于DTN的实时多人脸跟踪方法，并构建了一个多摄像机视场范围内的人脸跟踪与识别系统。首先，将局域网分为3层：人脸采集层、人脸特征提取层和人脸特征传输层。层与层之间采用push、pull和share的方式共享人脸特征向量，通过打开数据通道完成人脸特征的交换。利用YOLOv3和WIDER FACE人脸数据集采集小目标的人脸，采用CW聚类算法对同一摄像机的人脸数据进行聚类。利用难样本输入的DTN构建Siam16模型，实现多摄像机跟踪。

## 2 关键算法和网络设计

### 2.1 人脸聚类算法

CW算法是Markov聚类算法(MCL)的一个特例，CW算法和MCL算法都是基于图的聚类的。MCL将所有图都当作目的地，而图片的遍历过程可想象成一个漫步者随机到达这些目的地的模拟。该漫步者可以穿越所有目的地，也可以不穿越到较远的场景，而是走到距离起始点较近的场景就结束。MCL通过不断更新到达所有节点的概率来模拟图的转移过程，这里到达所有节点的概率可理解为图与图之间的相似性大小。最终该漫步者在$K$步之后找到了最佳的路线，即最终得到让MCL算法收敛的转移矩阵。该扩张过程是通过相邻矩阵和当前转移矩阵相乘实现的。该算法是一种非线性运算，每次扩张就相当于遍历所有节点一次，因此MCL的算法复杂度为$O(n^2)$。

CW是基于矩阵运算的算法，是在MCL算法的

基础上改进的，漫步者在每次行进时仅保留相邻节点中概率最大的节点作为目的地，这种改进能大大优化算法的性能。具体表现在虽然前几次的迭代和MCL的密集矩阵运算无差别，但在后面的迭代中，矩阵往往是稀疏的，即总是以邻域的最大类作为聚类中心。CW聚类算法的流程如图1所示。

1）初始化：CW构造一个无向图，每个节点是一个类，当相似度超过阈值时，两个节点连接形成关联边，权重为相似度。

2）迭代：初始化之后，每个节点都有属于自己的类别。然后，每个节点将所有相邻节点中权重最大的节点对应的类别作为当前节点的新类别，并完成类别更新。遍历所有节点后，重复迭代，直到满足迭代次数。

因为CW每次遍历都以相邻边的最大权重作为对象，所以算法复杂度取决于相邻边数。如果每个节点每次遍历所有边，那么CW算法的复杂度与MCL一致为$O(n^2)$。如果每个节点只有一条相邻边，那么CW算法的复杂度为$O(n)$。但事实上这两种情况都不太可能发生，因此CW算法的复杂度介于两者之间。

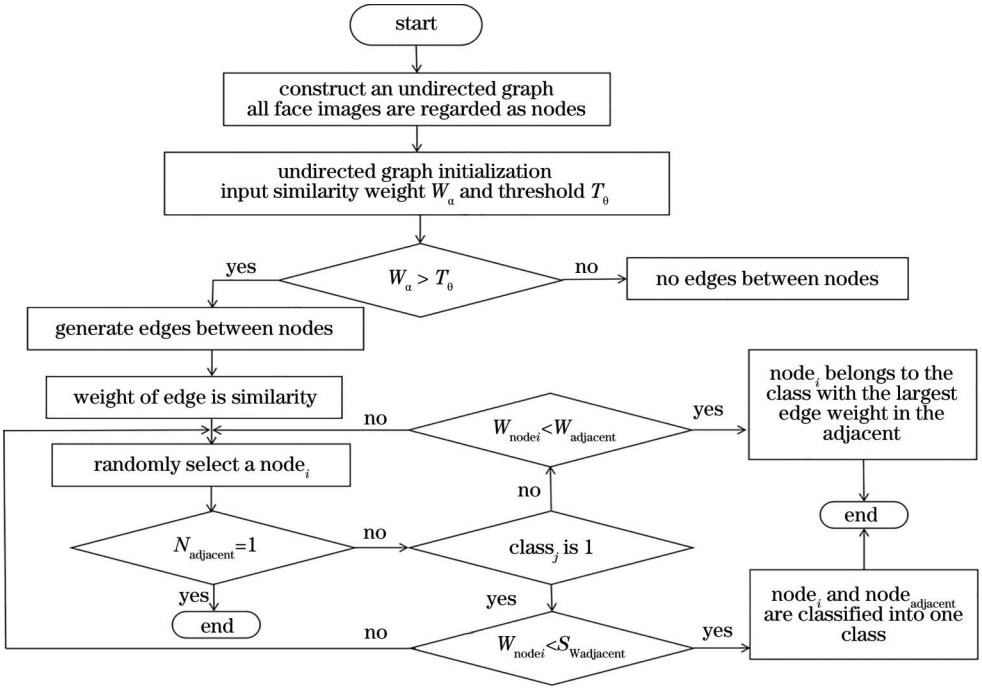


图1　CW算法的流程图

Fig. 1　Flow chart of CW algorithm

现有聚类算法的缺点及有待改进的问题：大多数聚类算法在聚类前需要指定聚类类别或聚类半径，如K-Means和DBSCAN，但这不适合视频监控目标，且对海量图像数据的处理能力不足。虽然HAC分层聚类不要求指定聚类的数量，但该算法对距离度量的选择不敏感，导致算法复杂度为$O(n^3)$，十分低效。表1列出了不同聚类算法的应用场景和算法复杂度。从表中可以看出，CW算法无论从算法复杂度上还是从应用场景上都适用于实时视频处理。CW算法特别适合于无向图聚类；该算法以收敛速度快而著称，且边缘数随迭代次数的增加呈线性增加，特别适用于大型图的聚类。因此针对视频监控无预知聚类中心且数据产生快的特点，CW聚类算法是一种好的选择。

表1　CW算法与几种聚类算法性能比较

Table 1　Performance comparison between CW algorithm and several clustering algorithms

| Parameter | K-means | DBSCAN | HAC | MCL | CW |
|---|---|---|---|---|---|
| Algorithm complexity | $O(n)$ | $O(n^2)$ | $O(n^3)$ | $O(n^2)$ | $O(n)$-$O(n^2)$ |
| Unknown number of clusters | × | × | √ | √ | √ |
| Real-time monitoring | × | × | × | × | √ |

### 2.2　DTN 设计

在特征向量相似性比较中，单卷积神经网络可以提取特征向量，但不适合预测图像的相似性或相异性问题。CW 算法最早应用在 FaceNet 网络中来比较人脸的特征相似度[14]。FaceNet 网络由三分支孪生卷积网络组成。FaceNet 网络包含了人脸聚类模块，但本实验组做了一个新的尝试，用改进的四分支 VGG-16 孪生网络进行特征提取，并完成人脸图片相似度比较。

FaceNet 是 Google 工程师 Schroff 等[14]提出的人脸识别模型。FaceNet 用到了两个异构分支网络，分别是 ZFNet 和 GoogleNet v1。利用 Triplet 损失训练模型输出 128 维的特征向量，Triplets 由来自同一人的两张人脸图像和来自另一人的第三张图像组成，训练的目的是使来自同一人的人脸对之间的欧氏距离要远小于来自不同人的人脸对之间的欧氏距离[21]。三元组的样本选择至关重要，作者设计了在线的难例挖掘策略 Triphard 损失来保证网络训练过程中持续增加三元组的训练难度。FaceNet 在 LFW 数据集上训练，准确率为 99.63%。

DTN 是在 FaceNet 的网络模型基础上改进的，DTN 的网络模型如图 2 所示。

改进 1：DTN 的 3 个分支网络是 VGG-16，VGG-16 网络有 16 层的深度，ZFNet 只有 8 层，但 GoogleNet v1 有 22 层的深度结构，因此尝试用 3 个 16 层分支的 VGG-16 孪生网络与一个 8 层加 22 层的异构孪生网络进行识别精度对比。

改进 2：DTN 使用样本挖掘损失（MSML）[22]作为损失函数完成训练，其中 FOCAL LOSS[23]主要负责样本均衡控制。孪生网络输出 4096 维的特征向量。

改进 3：FaceNet 训练用 3 样本输入构成 2 对难样本并用 Triphard 损失训练[24]。而 DTN 训练用 4 样本输入构成 3 对难样本，增加了负样本对的选取难度。前 3 层（1～3 层）构成 2 个正样本和 1 个负样本对的三分支孪生网络 $T_1$，后 3 层（2～4 层）构成 2 个负样本和 1 个正样本对的三分支孪生网络 $T_2$，即 2 个平行互相传递参数的三分支孪生网络。$T_1$ 的作用是缩短最远的正样本（最不相似的正样本对）之间的三重态损耗距离，$T_2$ 的作用是推开最近的负样本（最不相似的负样本对）之间的三重态损耗距离。这将使负样本与正样本完全分离，相当于同时使用 Triphard 损失 2 次，使不同类型样本之间的模糊区域消失。

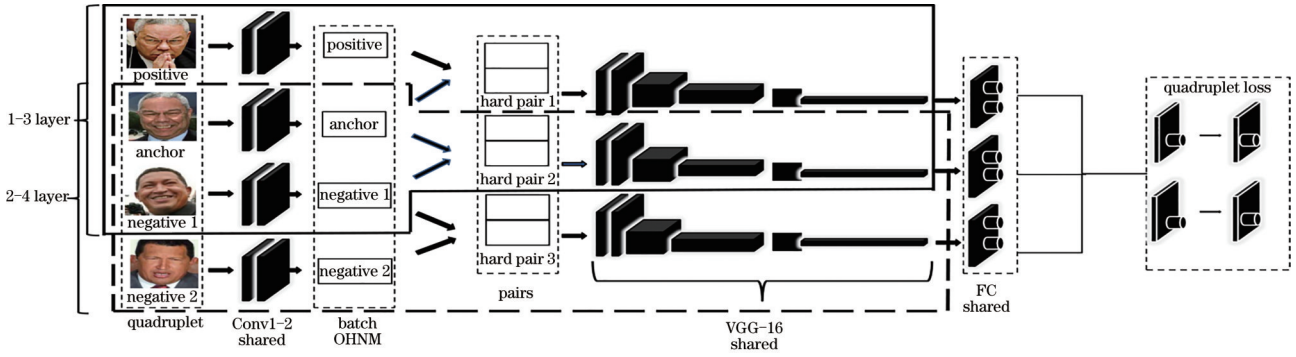


图 2　改进的包括两个三孪生网络的四孪生网络

Fig. 2　Improved quadruplet network including two triplet networks

## 3　实验与评价部分

### 3.1　人脸检测与结果分析

图 3 为 YOLOv3 损失函数的训练过程。在针对 WIDER FACE 人脸数据集的 YOLOv3 训练中，使用 Batch 8，subdivision 8，因此在训练输出中，训练迭代包含八组在 Region 为 82 和 Region 为 106 两个不同尺度上预测到的参数。每个组还包含一个图像，该图像与批处理和细分的设置值一致。即每批随机抽取 8 个样本，将这些样本分为 8 次批量参与训练，减少所有样本的内存消耗。本实验所用 GPU 是 NVIDIA 1080Ti 显卡，训练耗时大约 45 min，损失曲线平均成绩在 4.216 左右，测试帧速平均为 40 frame/s。侧脸和小脸图片测试结果很好，如图 4 和图 5 所示。

### 3.2　人脸聚类结果分析

每个监控摄像头的监控范围内都有许多行人，每个人都会被检测到大量属于同一个人的冗余人脸数据。如果将这些人脸图像的特征值都传递给其他摄像机，导致不必要的计算资源浪费和信道占用。在每个监控摄像机单元中，可以通过 CW 聚类算法对摄像头捕捉到的人脸图像进行去重，然后在每个视频区域用去重后的特征向量通过 push、

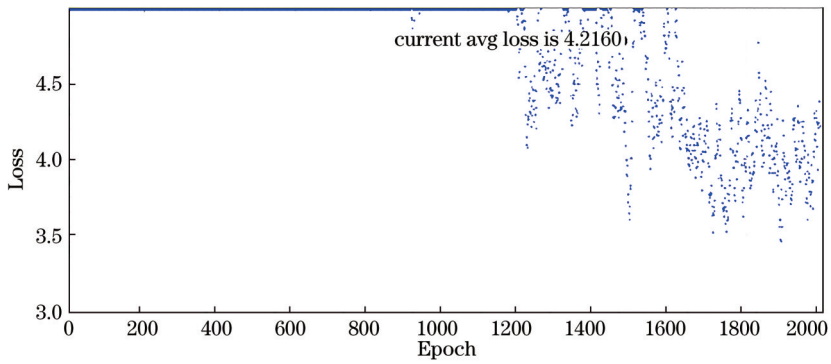


图 3　在 WIDER FACE 数据集上训练损失函数

Fig. 3　Training the loss function of WIDER FACE data set

图 4　侧脸检测

Fig. 4　Side face detection

图 5　小脸检测

Fig. 5　Small face detection

share、pull 的方式将特征共享给各摄像头用于跟踪多目标。File transfer protocol(FTP)有 3 种实现数据同步的通信方式：pull、push 和 share。push 将本地数据同步到远程服务器，pull 将远程服务器数据同步到本地，share 是服务器之间的数据共享和同步，如图 6 所示。

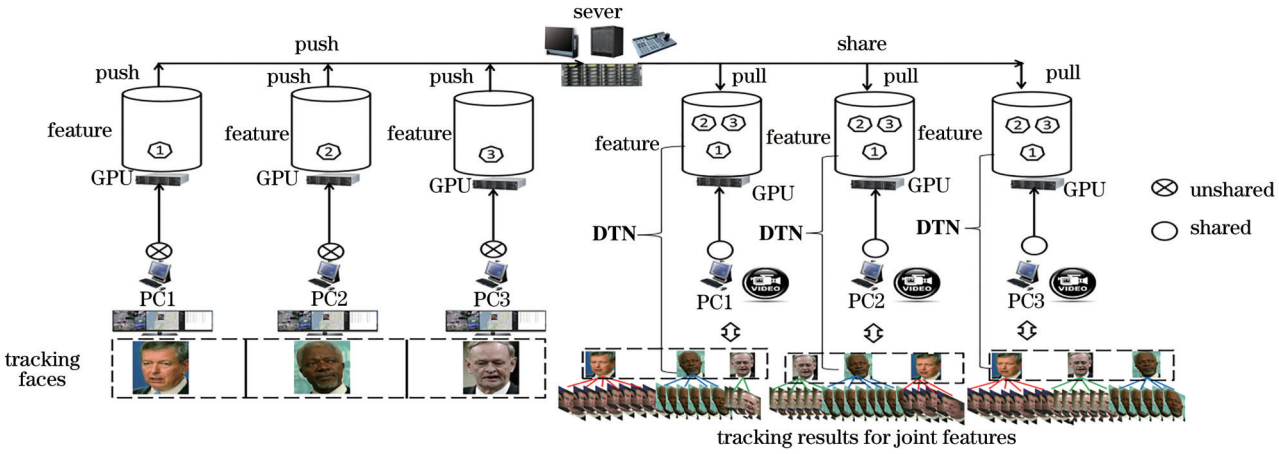


图 6　CW 算法聚类相似人脸节点

Fig. 6　CW algorithm clustering similar face nodes

人脸聚类的结果用 CW 三维节点表示，如图 7 所示。每个人脸图像相当于 CW 算法中的一个节点，在聚类之前，这些节点分散在整个聚类空间中[25]。当根据 DTN-VGG-16 的相似度权重开始比较两幅图像所有人脸节点的相似度时，根据相似度距离确定这些节点在聚类簇中的分布。相似度最高的节点位于聚类中心，相似度高的节点距离聚类中心较近，相似度低的节点距离聚类中心较远。在图 7 中，三维坐标表示节点的空间距离即特征向量的欧氏距离，$z$ 轴表示集群节点的相似性。相似性高的节点在簇峰附近聚集，相似度低的节点在聚类峰的底部。从图 7 中可以清楚地看到，CW 聚类算法可以找到相似的人脸簇，该算法在用摄像机分析行人类型中起着至关重要的作用，图 8 为图 7 中簇节点的可视化数据。

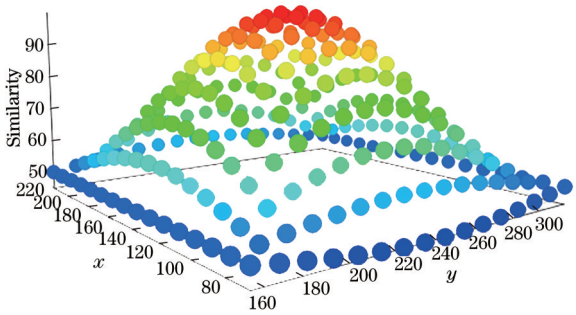


图 7　节点聚类簇

Fig. 7　Clustering with nodes

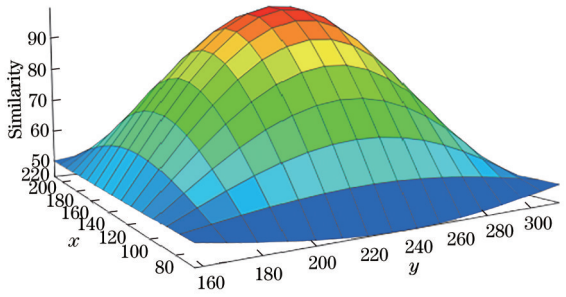


图 8　等值面显示图

Fig. 8　Isosurface display map

传统的聚类算法很容易放弃小样本或孤立节点。孤立的节点很难聚焦，但是 CW 算法使用较少的迭代来允许这些孤立的节点找到自己的聚类簇，这样小样本人脸就不会丢失。即使行人在镜头中停留的时间很短，他们的脸也会被捕捉和聚集。CW 算法的这一特性可以解决摄像机跟踪中行人人脸丢失的问题。

### 3.3　DTN 训练与特征提取

从 3 个方面对所提网络性能进行评估：网络训练过程中的参数相关性分析、DTN 在不同损失函数下的特征精度、与先进网络的性能比较。

DTN 的 3 个分支都使用 VGG-16 网络，这里称 DTN 为 DTNCNN16。VGG-16 网络由卷积层及相应的滤波器权值组成，这些权值不变，但卷积的最后一层 max-pooling 和 ReLU（激活函数）层被删除。以这种方式调整 VGG-16 的目的是使所提网络输出 4096 维的特征向量。DTNCNN16 使用 LFW 人脸数据集进行训练并建立模型。对比损失的 margin 参数设置为 0.6，学习率设置为 0.05，如图 9 所示。权重衰减设置为 0.005。最多 10 万次迭代的训练，相当于 1000 个 epoch。训练模型中的损失如图 10 所示。从图 11 中可以看出，在前 20000 次迭代中，损失显著降低，在随后的迭代中，损失函数进一步减小。80000 次迭代时接近 0.38，最终精度接近 99.2%，如图 11 所示。总时间如图 12 所示。

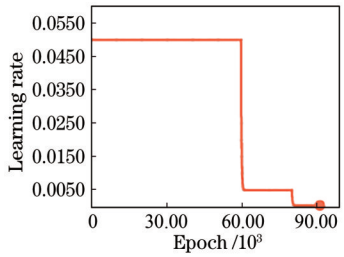


图 9　学习率随迭代次数的变化

Fig. 9　Learning rate varying with epoch

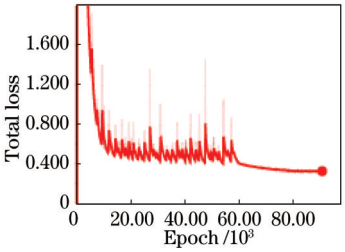


图 10　总损失随迭代次数的变化

Fig. 10　Total loss varying with epoch

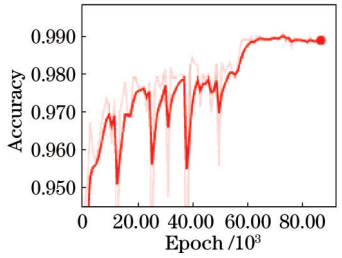


图 11　精确度随迭代次数的变化

Fig. 11　Accuracy varying with epoch

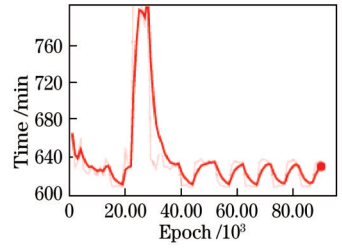


图 12　时间随迭代次数的变化

Fig. 12　Time varying with epoch

表 2 为 LFW 数据集上多种检测网络识别率对比。对于同一 DTNCNN16 网络，采用两个不同尺度的特征向量作为输出向量进行训练，分别为 1024 维和 4096 维。随着特征向量尺度的增大，对比度损失函数的精度也有所提高，4096 维特征向量的精度最高。另一方面，在 LFW 数据集上用不同的网络训练 4096 维特征向量，DTNCNN16 网络性能最好，达到 99.51%。进而得到结论：对孪生网络增加网

表 2　LFW 数据集上多种检测网络识别率对比

Table 2　Comparison of recognition rates of multiple detection networks on LFW data set

| Model | Feature size | Accuracy /% |
|---|---|---|
| 1-DTNCNN16 | 1024 | 99.47 |
| 2-DTNCNN16 | 4096 | 99.51 |
| 3-MSMLCNN16 | 1024 | 99.09 |
| 4-MSMLCNN16 | 4096 | 99.21 |
| 5-TrHardCNN16 | 1024 | 99.02 |
| 6-TrHardCNN16 | 4096 | 99.11 |
| 7-QuadrupletCNN16 | 1024 | 98.45 |
| 8-QuadrupletCNN16 | 4096 | 98.85 |
| 9-TripletCNN16 | 1024 | 98.25 |
| 10-TripletCNN16 | 4096 | 98.78 |
| 11-SiamCNN16 | 1024 | 95.99 |
| 12-SiamCNN16 | 4096 | 96.21 |
| 13-ResNet50 | 4096 | 95.87 |
| 14-VGG-19 | 4096 | 94.14 |
| 15-VGG-16 | 4096 | 92.46 |
| 16-AlexNet | 4096 | 89.04 |

络分支的改进方法可提高度量学习的精度。通过比较 VGG-16、SiamCNN16 和 QuadrupletCNN16 可以看出：增加层数可以提高网络精度；在同一网络模型中，网络特征向量维数越大，图像精度越高。

在通道数相同的情况下，最重要的是如何设计有效的损失函数。对 16 种损失函数进行了比较。各组分别建立真阳性率（TPR）和假阳性率（FPR）的 receiver operating characteristic（ROC）曲线。每个模型都有相似阈值的调整。曲线越靠近左上角，分类器性能就越好，如图 13～16 所示。每张图都有 6 条 ROC 曲线，在 FPR 都取 0.1 的情况下，ROC 越

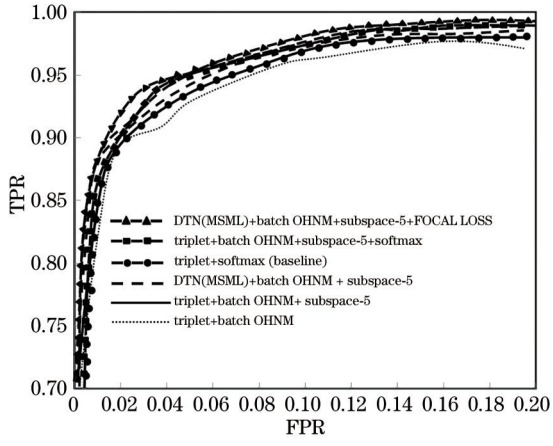


图 13　子空间数目为 5 时不同网络的 ROC 曲线

Fig. 13　ROC for different networks when subspace number is 5

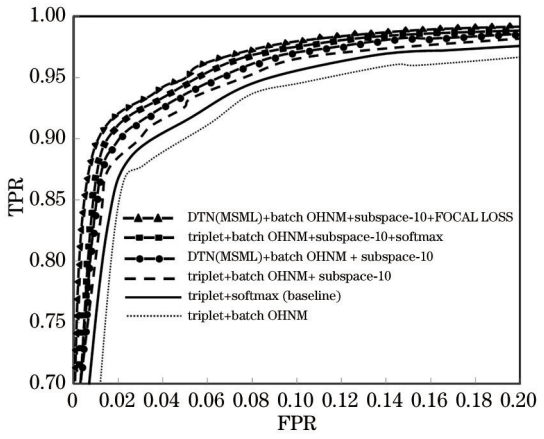


图 14　子空间数目为 10 时不同网络的 ROC 曲线

Fig. 14　ROC for different networks when subspace number is 10

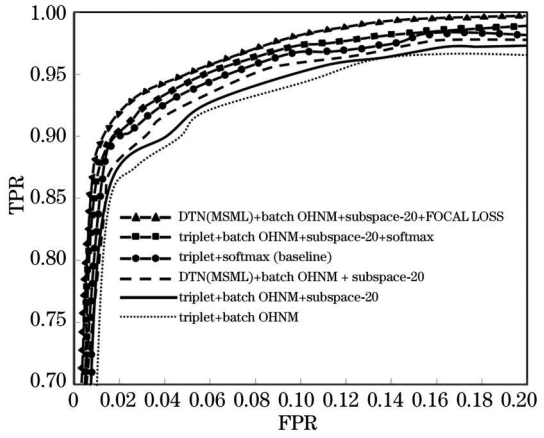


图 15　子空间数目为 20 时不同网络的 ROC 曲线

Fig. 15　ROC for different networks when subspace number is 20

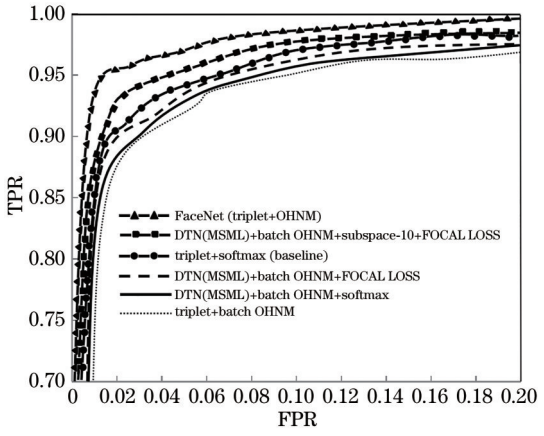


图 16　不同网络与 FaceNet 的 ROC 曲线比较

Fig. 16　ROC comparison between different networks and FaceNet

靠上的曲线分类器产生越高的 TPR，表明 ROC 越高，分类器性能越好。

图 13、图 14 和图 15 使用了两种方法，DTN+batch OHNM+subspace+FOCAL LOSS 和 DTN+batch OHNM+subspace+softmax。图 13、图 14 和图 15 分别是子空间数为 5，10，20 时不同网络的 ROC 曲线。通过对比子空间数目发现，CW 聚类子空间数目的选择是至关重要的。在这两种方法的基础上，通过 FOCAL LOSS 和 softmax 两种方法进行比较，并对其进行训练，原则上解决了样本平衡问题，困难样本的分类效果更好。在此基础上，分别去除子空间聚类进行训练，得出子空间的设置可以提高 FOCAL LOSS 分类的样本搜索速度。通过对几种不同损失函数模型的比较可见，DTN+batch-OHNM+subspace-10+FOCAL LOSS 是效果最好的模型和损失函数。从图 16 中可以看出，虽然 DTN 的精度比 FaceNet 低，但与其他模型相比，DTN 的训练效果有了明显的提高。

FaceNet 由两个异构分支网络组成，分别是 ZFNet 和 GoogleNet v1。由于 GoogleNet v1 在卷积层后增加了多层感知层的深层机制网络，因而提取特征的能力高于 VGG-16。实验结果表明，该改进虽然没有超过 FaceNet 的 99.63%，但由于增加了网络分支的平均深度，并加入了样本均衡和四输入的边缘样本挖掘机制，从而保持了对人脸跟踪和识别的精度。

### 3.4 网络 DTN 的跟踪效果评价

为了表明所提网络在多摄像机数据集上的鲁棒性和泛化能力，利用 Siam16 对中国科学院 GOT-10k 数据集上分割的掩模图像的评价结果进行了验证。对 Siam16 与 LWL(*)[26]、LWL[27]、PrDiMP-50[28]、DiMP-50[29]最新掩模跟踪网络进行比较。

Siam16 的输入样本首先经过盒掩模转换网络，得到初始分割掩模[26]。通过样本预处理生成 3 个样本对。3 个样本对分别是行人目标掩模和完整图像样本对、背景掩模和行人目标掩模样本对、背景掩模和完整图像样本对。将这 3 对样本作为 Siam16 模型的输入，完成训练。在后续的每一帧中，只需要利用分割掩模的极值来预测目标帧，不需要任何后处理。DiMP-50 完成了对未标记无监督模型的训练。PrDiMP-50 在 DiMP-50 模型中增加概率回归公式。这两种模式已经超越了许多先进的网络。然而，仅从分割图像中学习行人特征并不能突出背景杂波的干扰特征。Siam16 通过对分割的人体掩模区域与背景掩模图像进行比较，突出人体的局部特征。

实验结果表明，所提模型在多摄像机跟踪方面是有效的。首先对表 3 中各参数作解释，其中 AO 表示所有帧的跟踪结果和本地真实标注之间的平均重叠率，SR 表示重叠成功率，$SR_{0.5}$ 表示重叠率高于阈值 0.5 的成功跟踪帧的百分比，$SR_{0.75}$ 表示重叠率高于阈值 0.75 的成功跟踪帧的百分比。从表 3 可以看出，Siam16 的 AO 评分为 79.6%，已超过 PrDiMP-50 和 DiMP-50，比 DIMP-50 模型高 4.3 个百分点，比 PrDiMP-50 模型高 1.8 个百分点。当阈值为 0.5 时，Siam16 的重叠成功率为 90.4%，阈值为 0.75 时，重叠成功率为 74.9%。LWL 采用标记图像作为训练样本，采用紧凑的参数模型捕捉当前目标的信息，具有优化目标模型与真实性标注误差的能力，因此 LWL 模型优于 Siam16 模型。而 LWL(*)模型不仅具有较好的效果，而且通过使用标注图像作为训练样本，同时使用虚假标记进行再监督，可以得到最好的结果。当然，Siam16 模型在几个先进网络中表现也是很令人满意的。

表 3 在 GOT-10k 验证集上的性能比较

Table 3 Performance comparison on GOT-10k validation set

| Tracker/Year | $SR_{0.5}$ /% | $SR_{0.75}$ /% | AO /% |
|---|---|---|---|
| LWL(*)[26]/2021 | 95.1 | 85.2 | 86.7 |
| LWL[27]/2020 | 92.4 | 82.2 | 84.6 |
| PrDiMP-50[28]/2020 | 89.6 | 72.8 | 77.8 |
| DiMP-50[29]/2019 | 88.7 | 68.8 | 75.3 |
| Siam16/2021 | 90.4 | 74.9 | 79.6 |

CW 算法借助 Siam16 的训练权重计算两幅图像的相似度从而完成聚类，如图 17 所示。利用改进的 CW 算法，每个摄像机中的人脸集合可以聚类成一个单独的人脸簇，这些人脸簇中心的特征值保存在各自的人脸识别数据库中，并传递给服务器。该服务器被分配给其他摄像机，并存储在自己的网络摄像机的人脸识别库中，在该库中，摄像机使用 Siam16 再次开始比较面部特征，以识别和跟踪本地和跨摄像机的人脸，完成本地和跨摄像机人脸跟踪。

在不同场景中同一个人的脸将由 Siam16 跟踪并由 CW 算法聚类，该 Siam16 将同一个人脸标记成相同的 id，如图 18 所示。每一排都是同一个摄像头拍摄的视频场景，每个场景都有同一个人经过。通过观察图 18 发现，在每个摄像头中，都有同一行人经过，即人脸 id 值为 20 的行人。结果表明，人脸识别和跟踪可以跨摄像机实现。

图 18 的第 1 排分别用摄像机对用 id：5 和 id：20

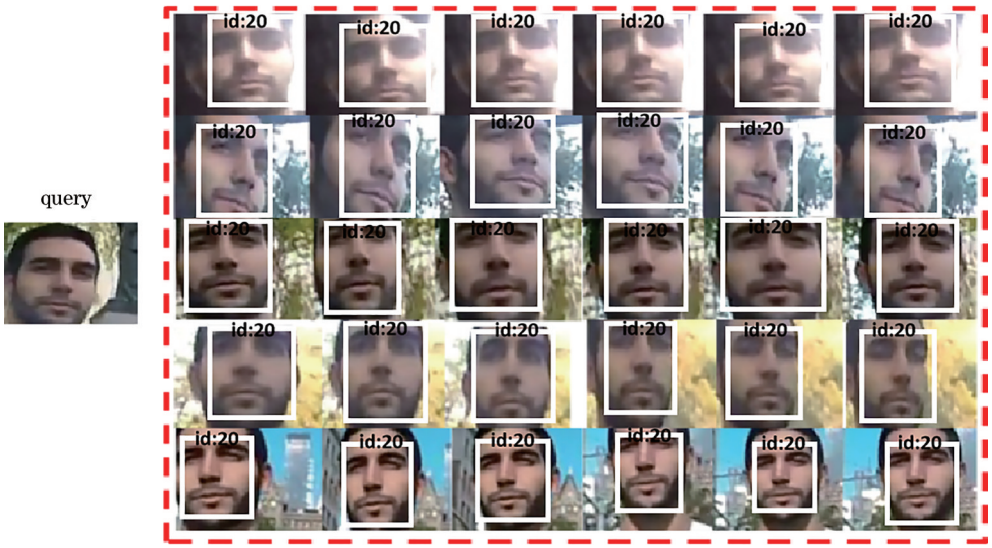


图 17　基于 Siam16 的多摄像机人脸聚类

Fig. 17　Multi camera face clustering based on Siam16


图 18　基于 Siam16 的多摄像机街景人脸跟踪结果

标记的人脸进行了追踪。随着时间的推移，两个人脸跟踪框保持稳定，没有丢失。从第 2 个场景（图 18 的第 2 排）来看，3 个行人的脸分别用 id：11、id：20 和 id：41 的人脸框标记。3 个人脸跟踪框保持稳定，没有在同一场景中丢失。即使 id：41 对应的人脸有被短时间遮挡的情况，后来它还是会被跟踪，而且 41 号人脸的小脸、侧脸的跟踪效果非常好。从第 3 个场景（图 18 的第 3 排）来看，两个人脸分别用 id：20 和 id：15 人脸框标记，并且在同一摄像机场景中跟踪稳定，侧面人脸跟踪效果良好。从第 4 个场景（图 18 的第 4 排）来看，3 个人的人脸分别用 id：20、id：56 和 id：28 人脸框标记，并在同一个摄像

机场景中进行稳定跟踪。小目标面和侧面的跟踪效果也表现很好,同一摄像机的跟踪效果稳定,且跨摄像机的跟踪效果也很稳定。实验结果表明,所提系统达到了多摄像机视场下的人脸跟踪目标,同时解决了小目标、遮挡、侧人脸识别的准确性问题。

## 4 结 论

设计并实现了一个多摄像机视场范围内的人脸跟踪系统。首先,将局域网分为3层:人脸采集层、人脸聚类特征提取层和人脸特征传输层。层与层之间采用push、pull和share的方式共享人脸特征向量。利用YOLOv3和WIDER FACE人脸数据集训练小目标人脸,采用CW聚类算法对同一摄像机的人脸数据进行去重。利用难样本输入的DTN构建Siam16来实现多摄像机跟踪,并通过在GOT-10k多摄像机验证数据集上与几种先进的方法进行性能比较,表明Siam16在多摄像机跟踪方面的优势。通过比较特征向量维数可知,4096维特征是网络精度的最佳输出。在服务器端,所提网络可以实时跟踪多个摄像头的人脸目标。在LFW数据集上训练DTN,通过对比最终确定DTN+batch-OHNM+subspace-10+FOCAL LOSS是效果最好的模型和损失函数。虽然DTN的精度比FaceNet低,但与其他模型相比,DTN的训练效果有了明显的提高,人脸识别率可达到99.51%。

## 参 考 文 献


[1] Li Z D, Zhong Y, Chen M, et al. Fast face image retrieval based on depth feature[J]. Acta Optica Sinica, 2018, 38(10): 1010004.
李振东, 钟勇, 陈蔓, 等. 基于深度特征的快速人脸图像检索方法[J]. 光学学报, 2018, 38(10): 1010004.

[2] Liu Y Z, Jiang Z Q, Ma F, et al. Hyperspectral image classification based on hypergraph and convolutional neural network[J]. Laser & Optoelectronics Progress, 2019, 56(11): 111007.
刘玉珍, 蒋政权, 马飞, 等. 基于超图和卷积神经网络的高光谱图像分类[J]. 激光与光电子学进展, 2019, 56(11): 111007.

[3] Wang L L, Liu J H, Fu X M. Facial expression recognition based on fusion of local features and deep belief network[J]. Laser & Optoelectronics Progress, 2018, 55(1): 011002.
王琳琳, 刘敬浩, 付晓梅. 融合局部特征与深度置信网络的人脸表情识别[J]. 激光与光电子学进展, 2018, 55(1): 011002.

[4] Lin Z M, Hong C Q, Zhuang W W. Face image deduplication based on fusion of face tracking and clustering[J]. Computer Science, 2020, 47(S2): 615-619.
林增敏, 洪朝群, 庄蔚蔚. 融合人脸跟踪和聚类的人脸图像去重方法[J]. 计算机科学, 2020, 47(S2): 615-619.

[5] Li P L, Zou J C, Li W. Face detection and tracking based on HOG and feature descriptor[J]. Journal of Zhejiang University of Technology, 2020, 48(2): 133-140.
李澎林, 邹嘉程, 李伟. 基于HOG和特征描述子的人脸检测与跟踪[J]. 浙江工业大学学报, 2020, 48(2): 133-140.

[6] Liu G Q, Li T. Research on multi-camera multi-target tracking method based on hierarchical relational model of trajectory tree[J]. Journal of Frontiers of Computer Science and Technology, 2020, 14(6): 1036-1044.
刘冠群, 李婷. 轨迹树层次关系模型多摄像机多目标跟踪研究[J]. 计算机科学与探索, 2020, 14(6): 1036-1044.

[7] Liu B C. Research on multi-source video sequence target tracking algorithm based on machine learning [D]. Changchun: Changchun University of Science and Technology, 2020.
刘保成. 基于机器学习的多源视频序列目标跟踪算法研究[D]. 长春: 长春理工大学, 2020.

[8] Zhou C C. Research of semi-supervised face recognition by convolutional neural networks based on graph clustering[D]. Kunming: Yunnan University, 2019.
周晨辰. 基于图聚类的卷积神经网络半监督人脸识别研究[D]. 昆明: 云南大学, 2019.

[9] Tian M Z, Ni L X, Xu L, et al. Multi-face real-time tracking based on dual panoramic camera for full-parallax light-field display[J]. Optics Communications, 2019, 442: 19-26.

[10] Lin C C, Hung Y. A prior-less method for multi-face tracking in unconstrained videos[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 538-547.

[11] Elhamifar E, Vidal R. Sparse subspace clustering: algorithm, theory, and applications[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(11): 2765-2781.

[12] Pratama M E, Kemas R W, Anisa H. Digital news graph clustering using Chinese whispers algorithm[J].

Journal of Physics: Conference Series, 2017, 801: 012062.

[13] Zhang Z P, Luo P, Loy C C, et al. Joint face representation adaptation and clustering in videos [M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9907: 236-251.

[14] Schroff F, Kalenichenko D, Philbin J. FaceNet: a unified embedding for face recognition and clustering [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 815-823.

[15] Chakraborty S, Paul D, Das S, et al. Entropy weighted power K-means clustering[C]//Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, August 26-28, 2020, Palermo, Sicily, Italy. Cambridge: PMLR, 2020, 108: 691-701.

[16] Shi J Y, He Q J, Wang Z L. GMM clustering-based decision trees considering fault rate and cluster validity for analog circuit fault diagnosis[J]. IEEE Access, 2019, 7: 140637-140650.

[17] Zhong Y, Li J M, Zhu S Z. Clustering geospatial data for multiple reference points[J]. IEEE Access, 2019, 7: 132423-132429.

[18] Ni L N, Li C, Wang X, et al. DP-MCDBSCAN: differential privacy preserving multi-core DBSCAN clustering for network user data[J]. IEEE Access, 2018, 6: 21053-21063.

[19] Chen H P, Yang X W, Lyu Y D. Copy-move forgery detection based on keypoint clustering and similar neighborhood search algorithm[J]. IEEE Access, 2020, 8: 36863-36875.

[20] Kennedy S M, Williamson W, Roth J D, et al. Cluster-based spectral-spatial segmentation of hyperspectral imagery[J]. IEEE Access, 2020, 8: 140361-140391.

[21] Chen W H, Chen X T, Zhang J G, et al. Beyond triplet loss: a deep quadruplet network for person re-identification[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 1320-1329.

[22] Yao H P, Fu D Y, Zhang P Y, et al. MSML: a novel multilevel semi-supervised machine learning framework for intrusion detection system[J]. IEEE Internet of Things Journal, 2019, 6(2): 1949-1959.

[23] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2999-3007.

[24] Lü X, Zhao C R, Chen W. A novel hard mining center-triplet loss for person re-identification[M]// Lin Z C, Wang L, Yang J, et al. Pattern recognition and computer vision. Lecture notes in computer science. Cham: Springer, 2019, 11859: 199-210.

[25] Biemann C. Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems[C]//Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, June 9, 2006, New York City. Morristown: Association for Computational Linguistics, 2006: 73-80.

[26] Zhao B, Bhat G, Danelljan M, et al. Generating masks from boxes by mining spatio-temporal consistencies in videos[EB/OL]. (2021-01-06) [2021-01-07]. https://arxiv.org/abs/2101.02196.

[27] Bhat G, Lawin F J, Danelljan M, et al. Learning what to learn for video object segmentation[M]// Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12347: 777-794.

[28] Bhat G, Danelljan M, van Gool L, et al. Learning discriminative model prediction for tracking[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 6181-6190.

[29] Danelljan M, van Gool L, Timofte R. Probabilistic regression for visual tracking[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 7181-7190.