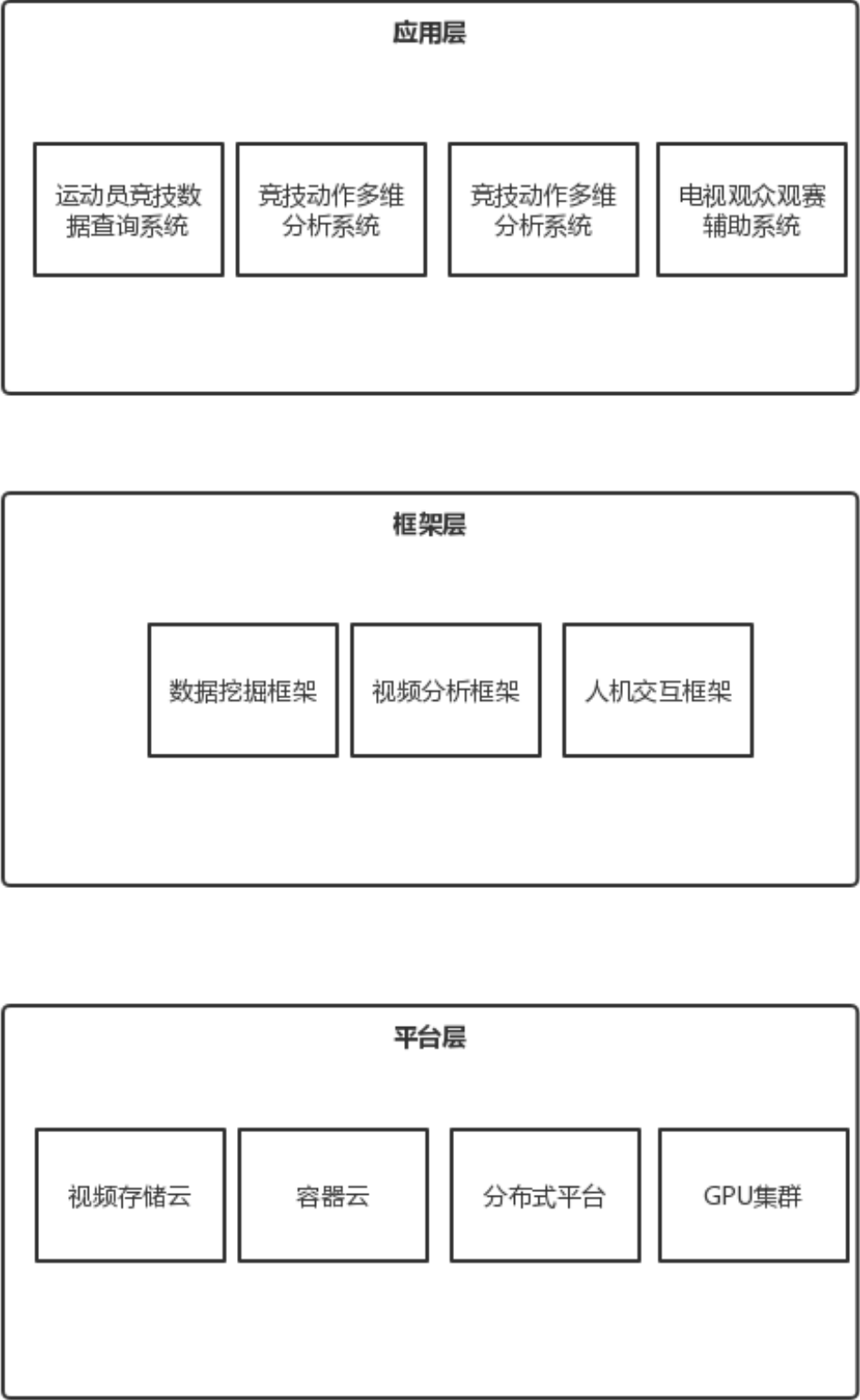


目录

一、	技术架构	2
二、	视频分析与理解技术	3
2.1.	分割 Segmentation	3
2.1.1	基于 CNN 的图像分割	3
2.1.2	整合局部和全局信息	4
2.1.3	基于 RNN 的图像分割	6
2.1.4	基于实例细分的图像分割	7
2.1.5	视频分割网络	8
2.2	检测 Detection	9
2.2.1	Region Based 方法	9
2.2.2	Unified 方法	10
2.3	分类 Classification	11
2.3.1	基于 iDT (improved dense trajectories)特征的方法	11
2.3.2	Two Stream Network 及衍生方法	12
2.3.3	C3D Network 及其相关衍生方法	14
2.3.4	其他方法	15
2.4	跟踪 Tracking	16
2.4.1	Observation Model	16
2.4.2	Tracking Model	17
2.5	预测 Prediction	18
2.5.1	Action Prediction	18
2.5.2	Motion Trajectory Prediction	18
2.6	理解 Comprehension	19
2.6.1	Temporal action localization in untrimmed videos via multi-stage cnns	19
2.6.2	Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos	19
2.6.3	A Pursuit of Temporal Accuracy in General Activity Detection	20
2.6.4	Temporal Unit Regression Network for Temporal Action Proposals	20
2.6.5	UntrimmedNets for Weakly Supervised Action Recognition and Detection ...	20

一、技术架构



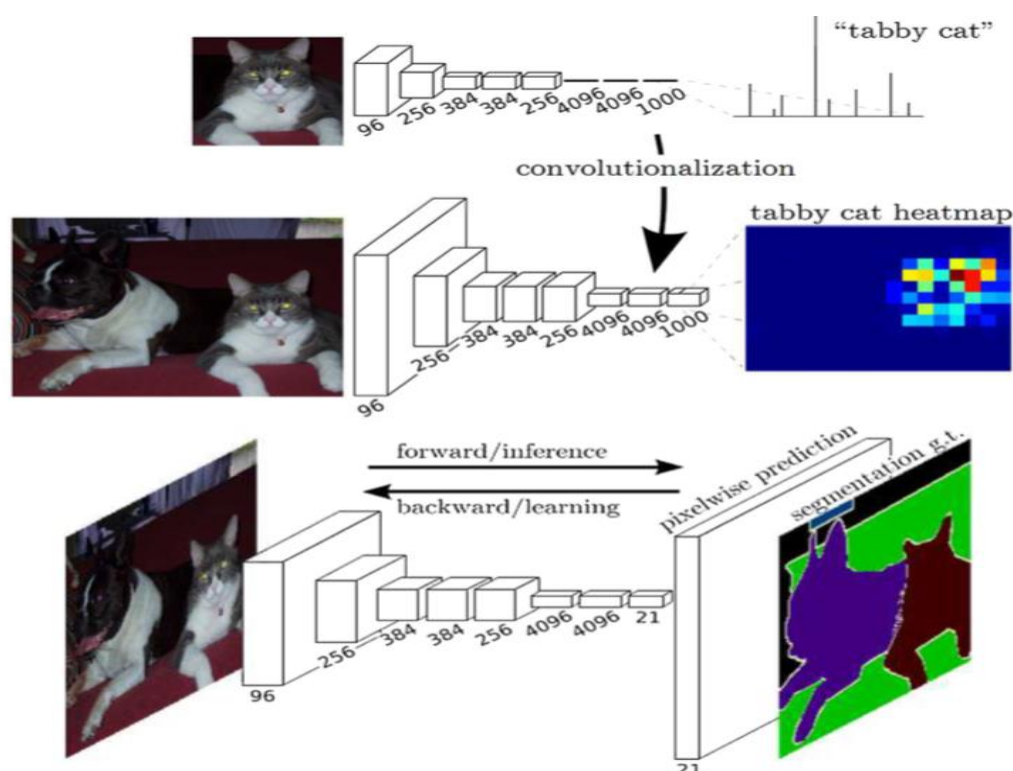
二、视频分析与理解技术

2.1.分割 Segmentation

2.1.1 基于 CNN 的图像分割

2.1.1.1 基于 FCN 的图像分割

FCN[d001]: 用卷积替换完全连接的层来转换用于分类的 CNN 以产生空间热图

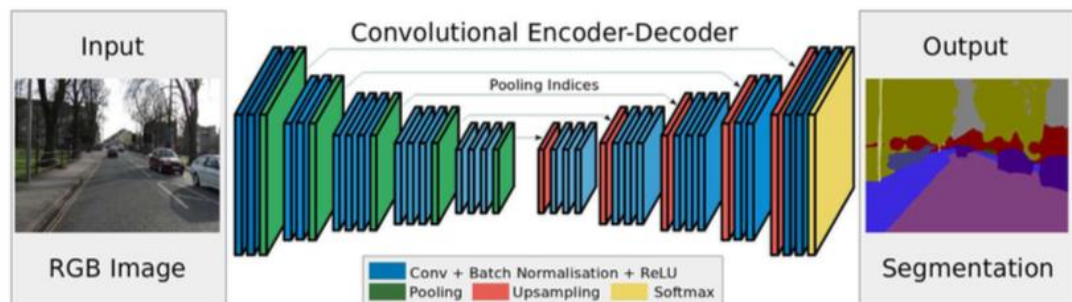


2.1.1.2 基于 U-Net 的图像分割

U-Net[d002]: 包括捕获上下文的收缩路径（减采样）和实现精确定位的对称扩展路径（增采样）。该体系结构还具有跳过连接，允许每个阶段的解码器从签约路径学习相关特征。

2.1.1.3 SegNet 和 Bayesian SegNet

SegNet[d003] Bayesian SegNet[d004]: 解码器由一组增采样和卷积层组成，最后接着是 **softmax** 分类器，用于预测与输入图像具有相同分辨率的输出的像素标签。解码器中的每个增采样层对应于编码器部分中的最大池化。这些层使用编码器阶段中相应特征映射的最大池索引对上采样特征映射进行增采样。然后将上传的地图与一组可训练的滤波器组进行卷积，以生成密集的特征映射。当特征贴图恢复到原始分辨率时，它们将被输入 **softmax** 分类器以生成最终分割。

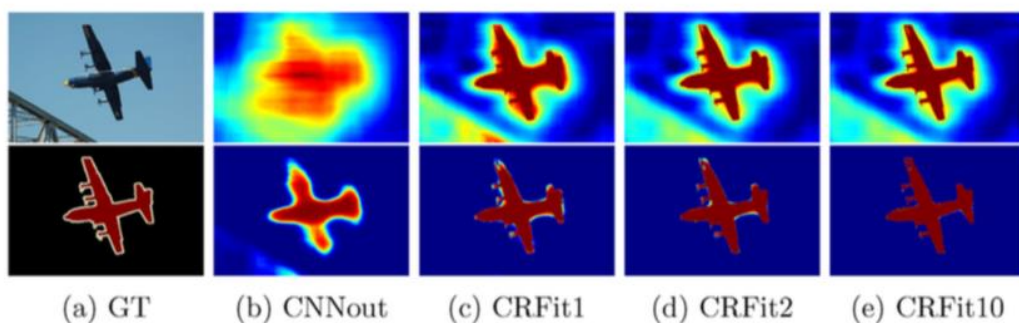


2.1.2 整合局部和全局信息

条件随机场，拓展卷积，多尺度聚合，上下文建模延迟

2.1.2.1 DeepLab

DeepLab[d005,d0066]: 利用 Krähenbühl 和 Koltun 的完全连接的成对 CRF 作为其管道中的分离后处理步骤来细化分割结果。它将每个像素建模为场中的节点，并且对于每对像素使用一个成对项，无论它们位于多远（该模型称为密集或完全连通因子图）。通过使用该模型，考虑短期和长程交互，使得系统能够恢复由于 CNN 的空间不变性而丢失的分段中的详细结构。尽管通常完全连通的模型是无效的，但是可以通过概率推理有效地近似该模型。

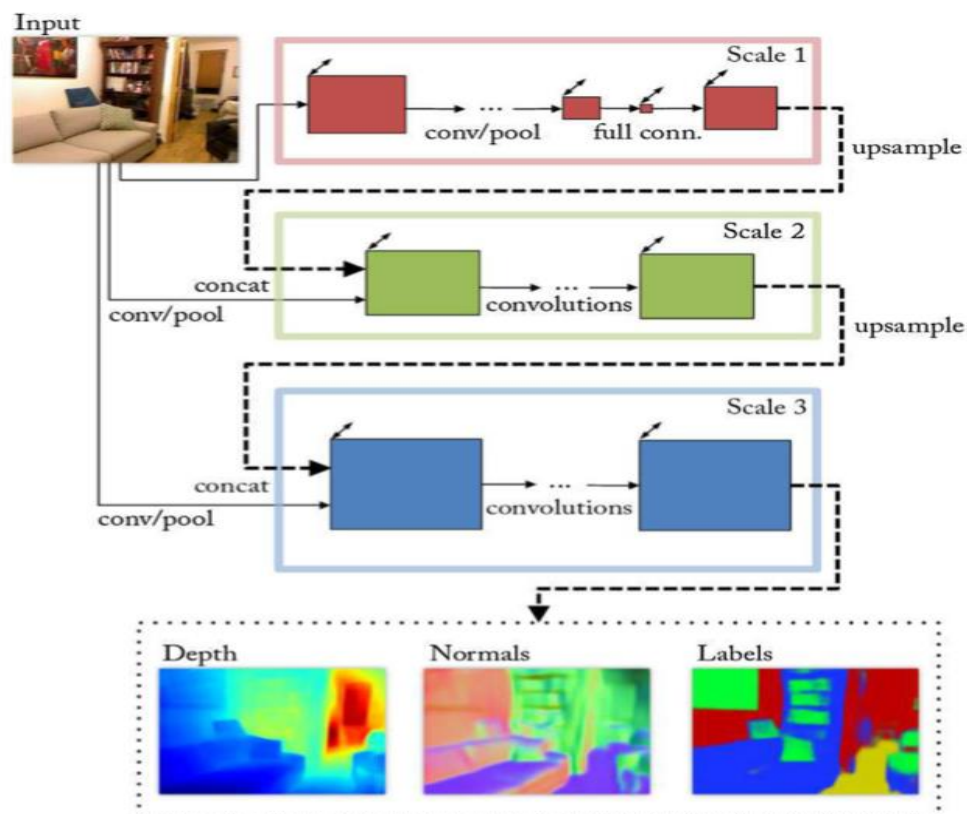


2.1.2.2 CRFasRNN

CRFasRNN[d007]: 通过将平均场推断步骤展开为 RNN，将 CRF 与 FCN 完全集成，并对整个网络进行端到端的训练

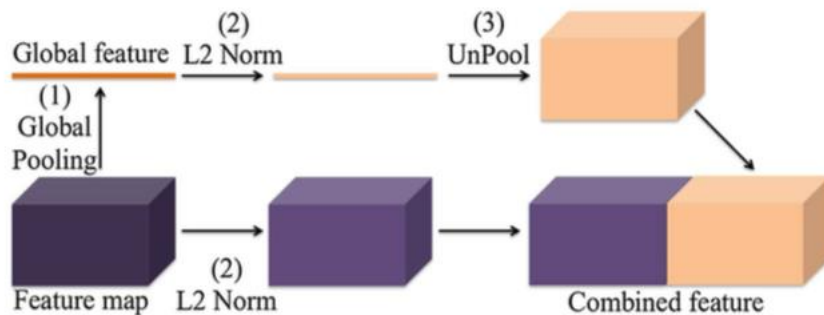
2.1.2.3 Multi-scale-CNN

Multi-scale-CNN[d010,d011,d012,d013]: 处理上下文知识集成的另一种可能方式是使用多尺度预测。几乎 CNN 的每个单个参数都会影响生成的要素图的比例。换句话说，相同的架构将对输入图像的像素数量产生影响，该像素的数量与特征图的像素相对应。这意味着过滤器将隐式学习检测特定尺度的特征（可能具有某种不变度）。此外，这些参数通常与手头的问题紧密耦合，使得模型难以推广到不同的尺度。克服该障碍的一种可能方法是使用多尺度网络，该网络通常利用针对不同尺度的多个网络，然后合并预测以产生单个输出



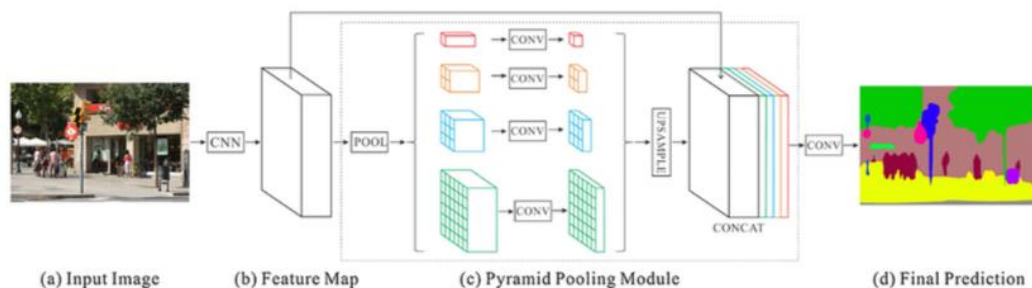
2.1.2.4 ParseNet

ParseNet[d014]: 早期融合，将全局特征解散为与局部特征相同的空间大小，然后将它们连接起来以生成在下一层中使用的组合特征或学习分类器



2.1.2.5 PSPNet

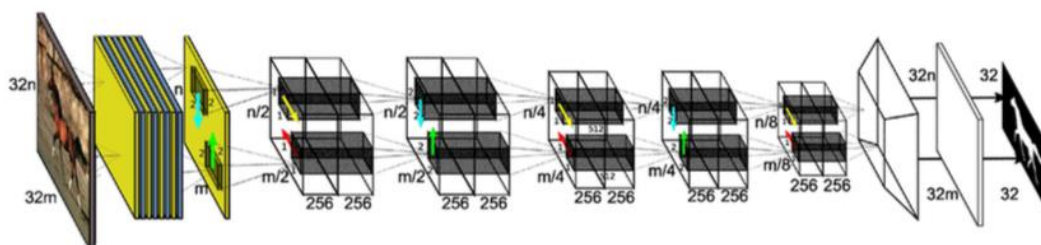
PSPNet[d015]: 提供聚焦在四个不同金字塔尺度的特征融合的金字塔解析模块，以便嵌入来自复杂场景的全局背景



2.1.3 基于 RNN 的图像分割

2.1.3.1 ReSeg

ReSeg[d016]: 输入图像用 VGG-16 网络的第一层处理，为得到的特征映射提供信息。进入一个或多个 ReNet 层进行微调。最后，使用基于转置卷积的上采样层来调整特征映射的大小。在这种方法中，使用了门控循环单元（GRU），因为它们在内存使用和计算能力方面取得了良好的性能平衡。



2.1.3.2 LSTM-CF

LSTM-CF[d017]: 使用两种不同的数据源：RGB 和深度。RGB 管道依

赖于 DeepLab 体系结构的变体在三个不同尺度上连接特征以丰富特征表示。全局上下文在深度和光度数据源上垂直建模，在这些垂直上下文中在两个方向上进行水平融合。

2.1.3.3 2D-LSTM

2D-LSTM[d018]: 输入图像被分成非重叠窗口，这些窗口被馈送到四个单独的 LSTM 存储块中。这项工作强调其在单核 CPU 上的低计算复杂性和模型简单性。

2.1.3.4 rCNNs

rCNNs[d019]: 其通过使用不同的输入窗口大小考虑先前的预测而反复训练具有不同的输入窗口大小。通过这种方式，预测标签会自动平滑，从而提高性能

2.1.4 基于实例细分的图像分割

2.1.4.1 SDS

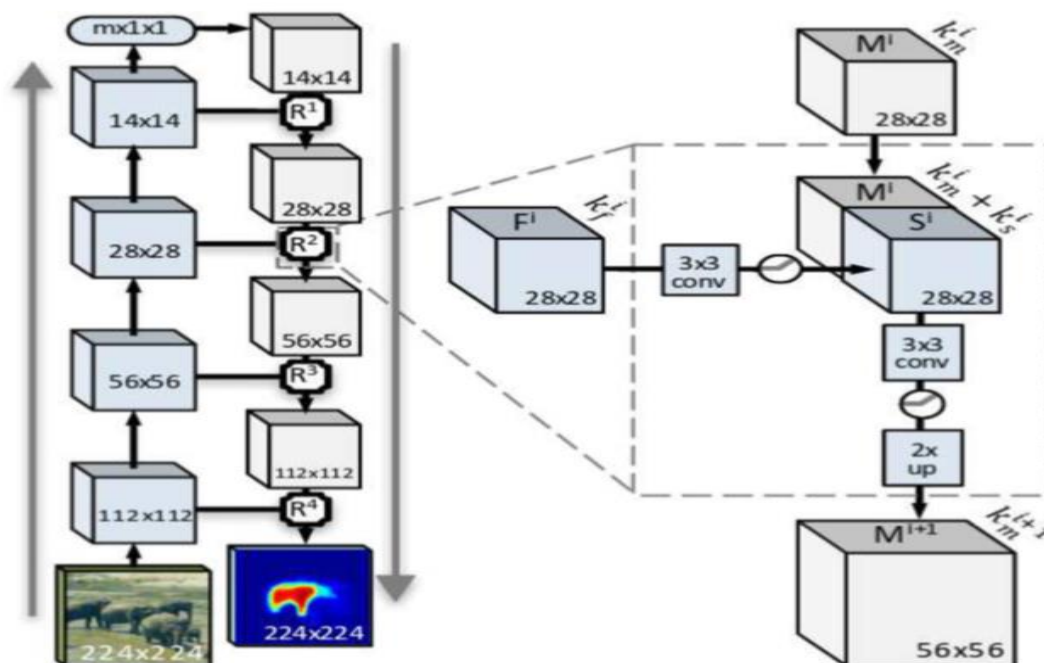
SDS[d020]: 首先使用自下而上的分层图像分割和称为多尺度组合分组 (MCG) [d027]的对象候选生成过程来获得区域提议。对于每个区域，通过使用 Region-CNN (R-CNN) [d028]的改编版本来提取特征，其使用由 MCG 方法提供的边界框而不是选择性搜索以及区域前景特征来微调。然后，通过在 CNN 特征之上使用线性支持向量机 (SVM) 来分类每个区域提议。

2.1.4.2 DeepMask

DeepMask[d021]: 基于单个 ConvNet 的对象提议方法。这个模型预测输入补丁的分段掩码以及此补丁包含对象的可能性。这两个任务是联合学习的，由单个网络计算，共享大多数层，除了最后一个任务特定的层。

2.1.4.3 SharpMask

SharpMask[d022]: 引入了一个渐进式细化模块，在自上而下的架构中融合了前一层到下一层的功能



2.1.5 视频分割网络

2.1.5.1 Clockwork Convnet

Clockwork Convnet [d024]: 该网络是 FCN 的改编，以利用视频中的时间线索来减少推理时间，同时保持准确性。特征速度网络中特征的时间变化率 - 跨帧不同层，因此浅层的特征变化比深层的特征更快。在该假设下，可以将层分组为阶段，根据其深度以不同的更新速率处理它们。通过这样做，由于其语义稳定性，可以在帧上保持深度特征，从而节省推理时间。作者提出了两种更新率：固定和自适应。固定时间表只为重新计算网络每个阶段的功能设置了一个恒定的时间范围。自适应调度以数据驱动的方式触发每个时钟，例如，取决于运动量或语义变化。

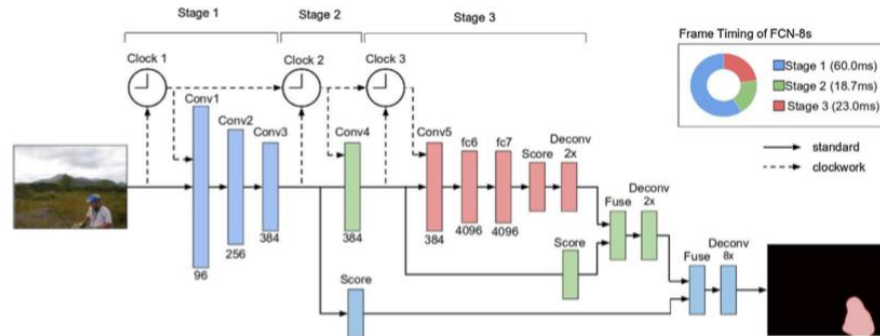
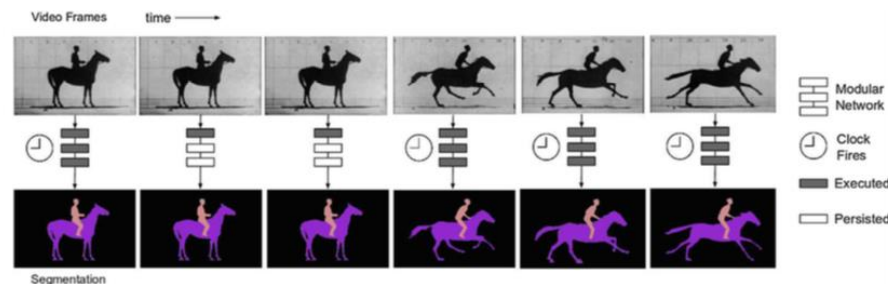


Fig. 21. The clockwork FCN with three stages and their corresponding clock rates.

Figure extracted from [95].



2.1.5.2 End2End Vox2Vox

End2End Vox2Vox [d025]: 引入的 Convolution 3D (C3D) 网络，并通过在最后添加反卷积层来扩展它以进行语义分割。他们的系统通过将输入分成 16 帧的剪辑来工作，分别对每个剪辑进行预测。它的主要贡献是使用 3D 卷积。这些卷积利用三维滤波器，这些滤波器适用于跨多个通道的时空特征学习。

2.1.5.3 SegmPred

SegmPred [d026]: 新颖的方法，能够预测未来尚未观察到的视频帧的语义分割图。该模型包含一个双尺度的体系结构，以对抗和非对抗的方式进行训练，以处理模糊的预测结果。

2.2 检测 Detection

2.2.1 Region Based 方法

2.2.1.1 RCNN

[g001]first to explore CNN for generic object detection.Region Proposals=> warp=>feature extraction =>SVM Classification =>Bounding Box

Regression

2.2.1.2 SPPNet

[g002]introduced the traditional spatial pyramid pooling(SPP), Enable convolutional feature sharing

2.2.1.3 Fast RCNN

[g003]First to enable end to end detector training (when ignoring the process of Region Proposals generation).Design a RoI pooling layer (a special case of SPP layer),

2.2.1.4 Faster RCNN

[g004]Propose RPN for generating nearly cost free and high quality Region Proposals instead of selective search; Unify RPN and Fast RCNN into a single network by sharing CONV layers;

2.2.1.5 RCNN $\ominus$ R

[g005]Replace selective search with static RPs

2.2.1.6 RFCN

[g006]Fully convolutional detection network; Minimize the amount of region wise computation; Design a set of position sensitive score maps using a bank of specialized CONV layers; Faster than Faster RCNN without sacrificing much accuracy.

2.2.1.7 Mask RCNN

[g007]object instance segmentation; Extends Faster RCNN by adding another branch for predicting an object mask in parallel with the existing branch for Bounding Box prediction

2.2.2 Unified 方法

2.2.2.1 OverFeat

Enable convolutional feature sharing; Multiscale image pyramid CNN

feature extraction[g008];

2.2.2.2 YOLO

First efficient unified detector; Drop Region Proposals process completely; divides an image into grids and predicts class probabilities, bounding box locations and confidences scores for those boxes. [g009]

2.2.2.3 YOLOv2

Propose a faster DarkNet19; Use a number of existing strategies to improve both speed and accuracy; Achieve high accuracy and high speed; YOLO9000 can detect over 9000 object categories in real time. [g010]

2.2.2.4 SSD

Effectively combine ideas from RPN and YOLO to perform detection at multiscale CONV layers. [g011]

2.3 分类 Classification

2.3.1 基于 iDT (improved dense trajectories)特征的方法

DT 算法[e001]的基本思路为利用光流场来获得视频序列中的一些轨迹，再沿着轨迹提取 HOF, HOG, MBH, trajectory4 种特征，其中 HOF 基于灰度图计算，另外几个均基于 dense optical flow (密集光流) 计算。最后利用 FV (Fisher Vector) 方法对特征进行编码，再基于编码结果训练 SVM 分类器。而 iDT[e002]改进的地方在于它利用前后两帧视频之间的光流以及 SURF 关键点进行匹配，从而消除/减弱相机运动带来的影响，改进后的光流图像被成为 warp optical flow

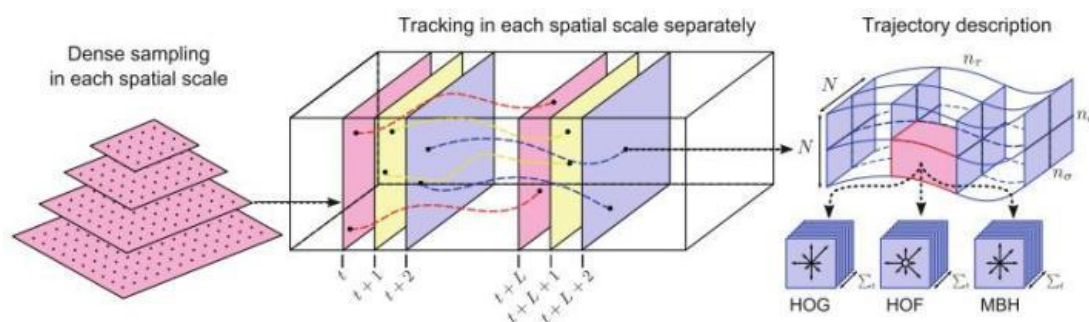


Fig. 2 Illustration of our approach to extract and characterize dense trajectories. *Left* Feature points are densely sampled on a grid for each spatial scale. *Middle* Tracking is carried out in the corresponding spatial scale for L frames by median filtering in a dense optical flow field. *Right*

The trajectory shape is represented by relative point coordinates, and the descriptors (HOG, HOF, MBH) are computed along the trajectory in a $N \times N$ pixels neighborhood, which is divided into $n_x \times n_y \times n_z$ cells

而[e003] 使用了两层的 fv 编码，应该是基于 iDT 方法的改进效果最好的文

章。

2.3.2 Two Stream Network 及衍生方法

2.3.2.1 Two-Stream Convolutional Networks for Action Recognition in Videos

Two Stream[e004]方法最初在这篇文章中被提出，基本原理为对视频序列中每两帧计算密集光流，得到密集光流的序列（即 **temporal** 信息）。然后对于视频图像（**spatial**）和密集光流（**temporal**）分别训练 CNN 模型，两个分支的网络分别对动作的类别进行判断，最后直接对两个网络的 **class score** 进行 **fusion**（包括直接平均和 **svm** 两种方法），得到最终的分类结果。注意，对与两个分支使用了相同的 2D CNN 网络结构。

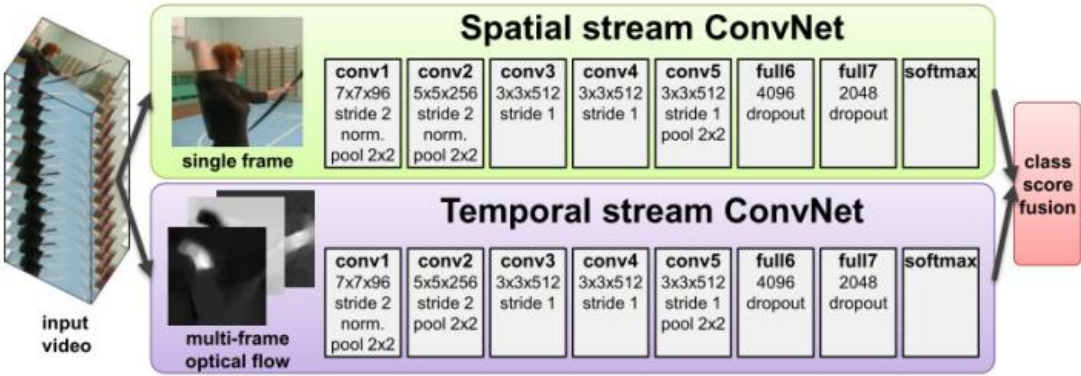
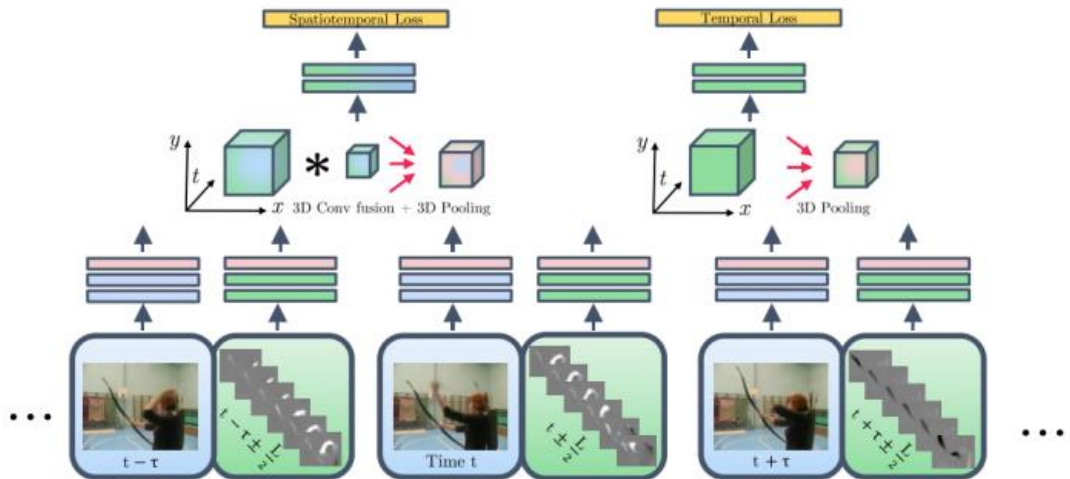


Figure 1: Two-stream architecture for video classification.

2.3.2.2 Convolutional Two-Stream Network Fusion for Video Action Recognition

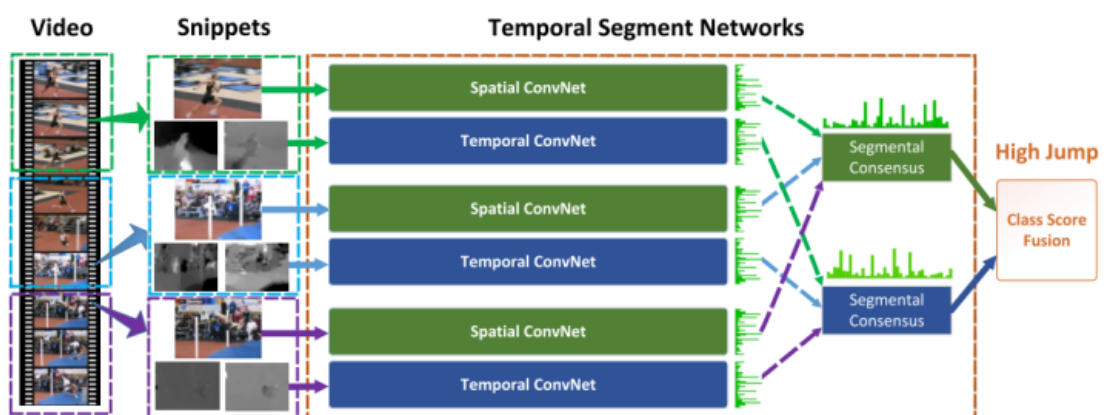
[e005]这篇论文的主要工作为在 **two stream network** 的基础上，利用 CNN 网络进行了 **spatial** 以及 **temporal** 的融合，从而进一步提高了效果。此外，该文章还将基础的 **spatial** 和 **temporal** 网络都换成了 **VGG-16 network**。



2.3.2.3 Temporal Segment Networks: Towards Good Practices for Deep Action Recognition

[e006]这篇文章提出的 TSN 网络也算是 spatial+temporal fusion，结构图见下图。这篇文章对如何进一步提高 two stream 方法进行了详尽的讨论，主要包括几个方面

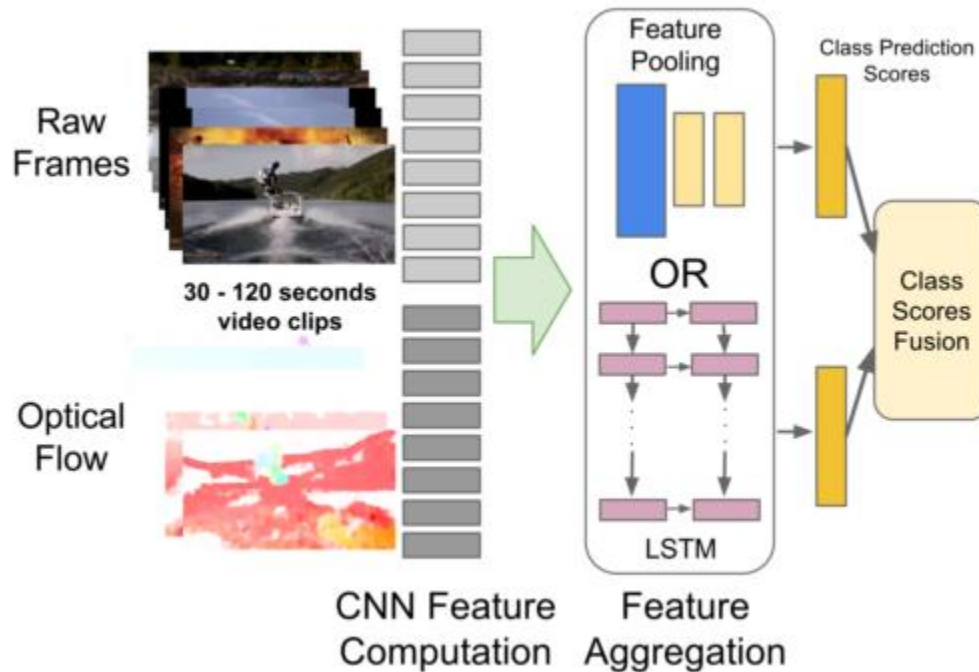
1. 输入数据的类型：除去 two stream 原本的 RGB image 和 optical flow field 这两种输入外，这篇文章中还尝试了 RGB difference 及 warped optical flow field 两种输入。最终结果是 RGB+optical flow+warped optical flow 的组合效果最好。
2. 网络结构：尝试了 GoogLeNet, VGGNet-16 及 BN-Inception 三种网络结构，其中 BN-Inception 的效果最好。
3. 训练策略：包括 跨模态预训练，正则化，数据增强等



2.3.2.4 Beyond Short Snippets: Deep Networks for Video Classification

Joe

[e007]这篇文章主要是用 LSTM 来做 two-stream network 的 temporal 融合



2.3.3 C3D Network 及其相关衍生方法

C3D[e008]是 facebook 的一个工作，采用 3D 卷积和 3D Pooling 构建了网络.

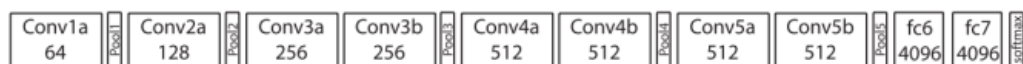
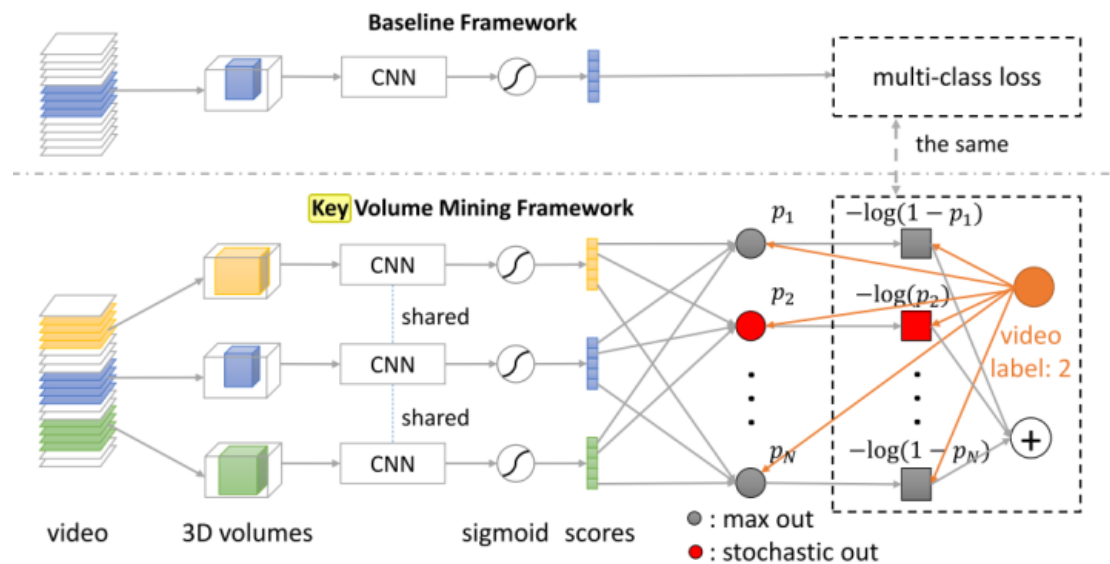


Figure 3. **C3D architecture.** C3D net has 8 convolution, 5 max-pooling, and 2 fully connected layers, followed by a softmax output layer. All 3D convolution kernels are $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Number of filters are denoted in each box. The 3D pooling layers are denoted from pool1 to pool5. All pooling kernels are $2 \times 2 \times 2$, except for pool1 is $1 \times 2 \times 2$. Each fully connected layer has 4096 output units.

相关论文还有[e009][e010]

2.3.4 其他方法

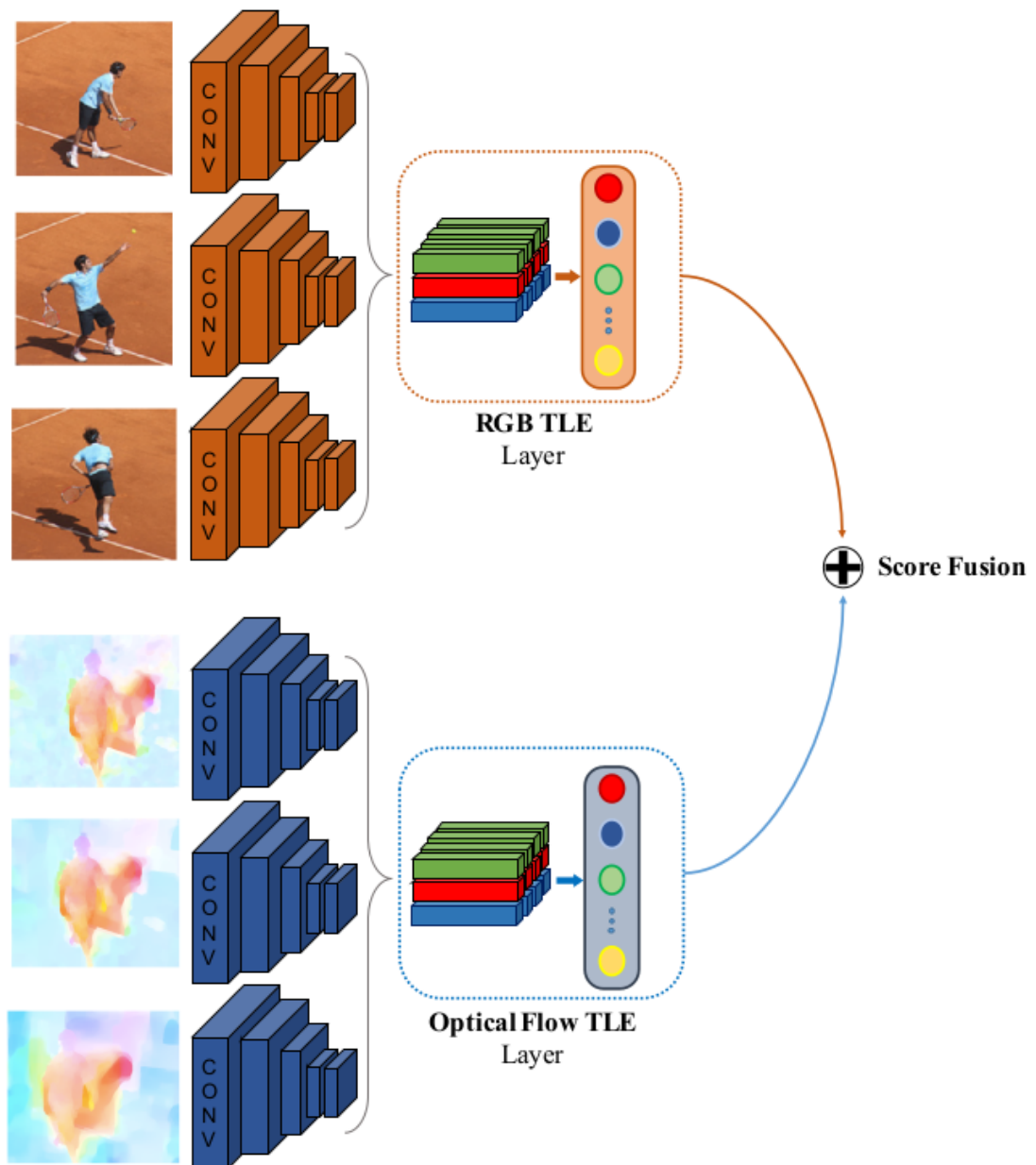
2.3.4.1 A Key Volume Mining Deep Framework for Action Recognition



本文[e011]主要做的是 key volume 的自动识别。通常都是将一整段动作视频进行学习，而事实上这段视频中有一些帧与动作的关系并不大。因此进行关键帧的学习，再在关键帧上进行 CNN 模型的建立有助于提高模型效果

2.3.4.2 Deep Temporal Linear Encoding Networks

本文[e012]主要提出了“Temporal Linear Encoding Layer”时序线性编码层，主要对视频中不同位置的特征进行融合编码。至于特征提取则可以使用各种方法，文中实验了 two stream 以及 C3D 两种网络来提取特征



2.4 跟踪 Tracking

2.4.1 Observation Model

2.4.1.1 Appearance Model

Visual Representation: Color histogram[b001], Gaussian mixture models, sparse representation, manifold learning ,hidden Markov random field model, Robust Principle Component Analysis (RPCA)[b002].

Statistical Measuring:

Strategy	Employed Cue	Ref
Boosting	Color,HOG,shapes,covariance matrix, etc	[b003],[b004],[b005]
Concatenating	Color, HOG, optical flow, etc	[b006]
Summation	Color, depth, correlogram, LBP, etc	[b007][b008][b009]
Product	Color, shapes, bags of local features, etc	[b010][b011][b012][b013]
Cascading	Depth, shape, texture, etc	[b014][b015]

2.4.1.2 Motion Model(mutual motion model)

Linear Motion Model :[b016], [b017], [b018].

Non-linear Motion Model:[b019],[b020]

2.4.1.3 Interaction Model

Social Force Models:

Ref	employed Forces
[b021]	repulsion, constancy, fodelity
[b022]	replulsion, constancy, attraction, coherence
[b023]	coherence
[b024]	relulsion, coherence
[b025]	constancy, fidelity, repulsion
[b026]	constancy, coherence

Crowd Motion Pattern Models:[b027],[b028],[b029],[b030],[b031],[b032]

2.4.1.4 Exclusion Model

Occlusion Handlin

2.4.2 Tracking Model

2.4.2.1 Probabilistic inference[b028],[b012],[b033],[b034]

Kalman filter [b035], [b036]

Extended Kalman filter [b037]

Particle filter [b016], [b010], [b038], [b008], [b039], [b040], [b041], [b042]

2.4.2.2 Deterministic Optimization

Bipartite graph matching: [b016], [b043], [b044], [b045], [b046], [b023], [b047], [b048]

Dynamic Programming: [b049], [b050], [b051], [b052], [b053], [b054], [b055], [b056], [b057], [b058]

Min-cost max-flow network flow: [b054], [b059], [b060], [b061], [b062], [b063]

Conditional random field: [b064], [b020], [b065], [b066]

MWIS:[b006],[b017]

2.5 预测 Prediction

2.5.1 Action Prediction

2.5.1.1 Early Action Classification

integral bag-of-words , dynamic bag-of-words [c001]

sparse coding [c002]

motion velocity peaks detection [c003]

information divergence [c013][c004]

deep learning:[c016][c014][c017][c015]

2.5.1.2 Intention Prediction

Stochastic Context Sensitive Grammar [c019]

hierarchical event[c018]

Probabilistic Suffix Tree [c019]

Predictive Accumulative Function [c020]

conditional random field [c018]

2.5.2 Motion Trajectory Prediction

Clustering to learn pattern[c021][c005][c006][c007]

inverse optimal control; inverse reinforcement learning[c011][c012]

dynamic Bayesian network[c010]

Gaussian Process

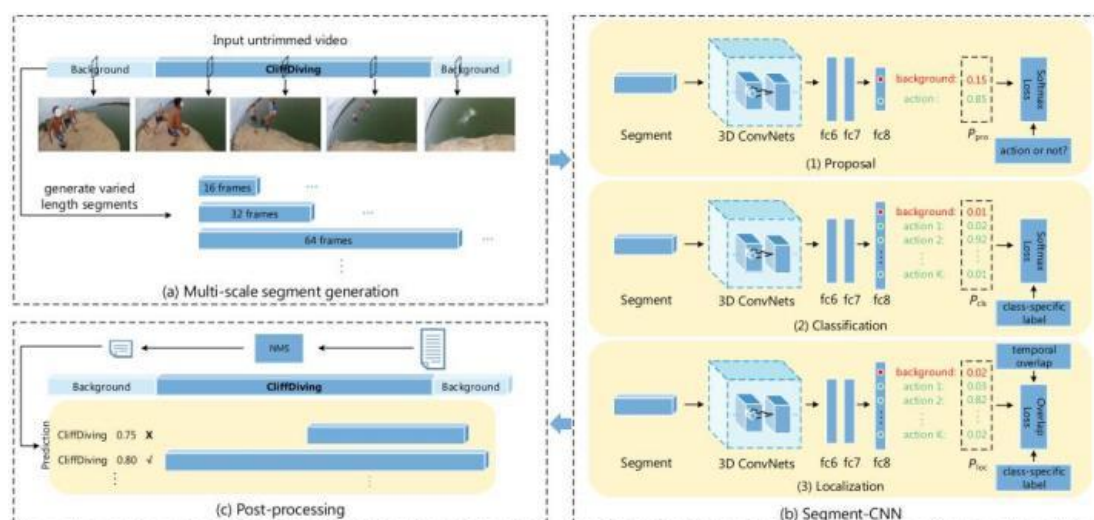
Hidden Markov Model

deep learning (RNN/LSTM/GAN):[c008][c009][c022][c023][c024][c025]

2.6 理解 Comprehension

2.6.1 Temporal action localization in untrimmed videos via multi-stage cnns

该方法首先使用滑窗的方法生成多种尺寸的视频片段(segment), 再使用多阶段的网络(Segment-CNN)来处理。SCNN 主要包括三个子网络, 均使用了 C3D network。第一个是 proposal network, 用来判断当前输入的视频片段是一个动作的概率; 第二个为 classification network, 该网络用于给视频片段分类, 但该网络不用于测试环节, 而只是用作初始化 localization network; 第三个子网络为 localization network, 该网络的输出形式依旧为类别的概率, 但在训练时加入了重叠度相关的损失函数, 使得网络能更好的估计一个视频片段的类别和重叠度。最后采用了非极大化抑制 (NMS) 来去除重叠的片段, 完成预测。[k001]



2.6.2 Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos

基于 C3D (3D CNN 网络) 设计了一个卷积逆卷积网络, 输入一小段视频, 输出 frame-level 的动作类别概率。该网络主要是用来对 temporal action detection 中的动作边界进行微调, 使得动作边界更加准确, 从而提高 mAP。[k002]

2.6.3 A Pursuit of Temporal Accuracy in General Activity Detection

用 TSN(Temporal Segment Network, 用于 Action Recognition 的 state-of-the-art 方法,也是他们组的工作)在视频序列上每隔几帧判断该时刻属于动作的概率(是动作/不是动作),基于这个概率的曲线提出了一种叫 Temporal Actionness Grouping(TAG)的 proposal 方法。之后再进行类别的判断。[k003]

2.6.4 Temporal Unit Regression Network for Temporal Action Proposals

采用了类似 Faster-RCNN 的思路,主要是作为 proposal 的方法。pipeline 比较简单[k004]

2.6.5 UntrimmedNets for Weakly Supervised Action Recognition and Detection

提出了一种可以用于 Action Recognition 和 Action Detection 的弱监督方法。[k005]

三、相关技术的可能应用

- 手部脚部轨迹跟踪,利用显示出来的轨迹,回顾分析目标的动作是否到位
连贯 跟踪
- 模式识别:识别出目标的一些习惯动作
- 标准动作解析:在目标做出动作后,分析其标准动作,向观众说明
- 招式之间的克制关系:进攻和防守等不同的动作之间的胜率
- 运动员&教练:分析对手技术特点;比赛时分析自己和对手状态
- 转播:自动剪辑得分镜头和精彩镜头

Ref:

[d001] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015) 3431–3440.

[d002] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI), vol. 9351 of LNCS, Springer, 2015, pp. 234–241 (available on 1505.04597), <http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a>.

- [d003] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for scene segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (12) (2015) 2481–2495.
- [d004] A. Kendall, V. Badrinarayanan, R. Cipolla, Bayesian Segnet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding, 2015 arXiv:1511.02680.
- [d005] C. Liang-Chieh, G. Papandreou, I. Kokkinos, K. Murphy, A. Yuille, Semantic image segmentation with deep convolutional nets and fully connected CRFs, *International Conference on Learning Representations* (2015).
- [d006] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs, *CoRR* abs/1606.00915, 2016 <http://arxiv.org/abs/1606.00915>.
- [d007] S. Bell, P. Upchurch, N. Snavely, K. Bala, Material recognition in the wild with the materials in context database, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015) 3479–3487.
- [d008] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P.H. Torr, Conditional random fields as recurrent neural networks, *Proceedings of the IEEE International Conference on Computer Vision* (2015) 1529–1537.
- [d009] A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation, 2016 arXiv:1606.02147.
- [d010] A. Raj, D. Maturana, S. Scherer, Multi-scale Convolutional Architecture for Semantic Segmentation, 2015.
- [d011] D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, *Proceedings of the IEEE International Conference on Computer Vision* (2015) 2650–2658.
- [d012] A. Roy, S. Todorovic, A multi-scale CNN for affordance segmentation in RGB images, in: *European Conference on Computer Vision*, Springer, 2016, pp. 186–201.
- [d013] X. Bian, S.N. Lim, N. Zhou, Multiscale fully convolutional network with application to industrial inspection, in: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2016, pp. 1–8.
- [d014] W. Liu, A. Rabinovich, A.C. Berg, Parsenet: Looking Wider to See Better, 2015 arXiv:1506.04579.
- [d015] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid Scene Parsing Network, *CoRR* abs/1612.01105, 2016 <http://arxiv.org/abs/1612.01105>.
- [d016] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, A. Courville, ReSeg: a recurrent neural network-based model for semantic segmentation, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2016).
- [d017] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, L. Lin, LSTM-CF: Unifying Context Modeling and Fusion with LSTMs for RGB-D Scene Labeling, *Springer International Publishing, Cham*, 2016, pp. 541–557.
- [d018] W. Byeon, T.M. Breuel, F. Raue, M. Liwicki, Scene labeling with LSTM recurrent neural networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015) 3547–3555.

- [d019] P.H. Pinheiro, R. Collobert, Recurrent convolutional neural networks for scene labeling, ICML (2014) 82–90.
- [d020] B. Shuai, Z. Zuo, B. Wang, G. Wang, DAG-recurrent neural networks for scene labeling, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016) 3620–3629.
- [d021] P.O. Pinheiro, R. Collobert, P. Dollar, Learning to segment object candidates, in: Advances in Neural Information Processing Systems, 2015, pp.1990–1998.
- [d022] P.O. Pinheiro, T.-Y. Lin, R. Collobert, P. Dollár, Learning to refine object segments, in: European Conference on Computer Vision, Springer, 2016, pp.75–91.
- [d023] S. Zagoruyko, A. Lerer, T. Lin, P.O. Pinheiro, S. Gross, S. Chintala, P. Dollár, A multipath network for object detection, in: Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19–22, 2016, 2016 <http://www.bmva.org/bmvc/2016/papers/paper015/index.html>
- [d024] E. Shelhamer, K. Rakelly, J. Hoffman, T. Darrell, Clockwork convnets for video semantic segmentation, in: Computer Vision – ECCV 2016 Workshops, Springer, 2016, pp. 852–868.
- [d025] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Deep end2end voxel2voxel prediction, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2016) 17–24.
- [d026] N. Neverova, P. Luc, C. Couprie, J.J. Verbeek, Y. LeCun, Predicting Deeper into the Future of Semantic Segmentation, CoRR abs/1703.07684, 2017.
- [d027] P. Arbeláez, J. Pont-Tuset, J.T. Barron, F. Marques, J. Malik, Multiscale combinatorial grouping, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014) 328–335.
- [d028] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
- [g001]. Girshick R., Donahue J., Darrell T., Malik J. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, pp. 580–587
- [g002]. He K., Zhang X., Ren S., Sun J. (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. In: ECCV, pp. 346–361
- [g003]. Girshick R. (2015) Fast R-CNN. In: ICCV, pp. 1440–1448
- [g004]. Ren S., He K., Girshick R., Sun J. (2015) Faster R-CNN: Towards real time object detection with region proposal networks. In: NIPS, pp. 91–99
- [g005]. Lenc K., Vedaldi A. (2015) R-CNN minus R. In: BMVC15
- [g006]. Dai J., Li Y., He K., Sun J. (2016) RFCN: object detection via region based fully convolutional networks. In: NIPS, pp. 379–387
- [g007]. He K., Gkioxari G., Dollar P., Girshick R. (2017) Mask RCNN. In: ICCV
- [g008]. Sermanet P., Eigen D., Zhang X., Mathieu M., Fergus R., LeCun Y. (2014) OverFeat: Integrated recognition, localization and detection using convolutional networks. In: ICLR
- [g009]. Redmon J., Divvala S., Girshick R., Farhadi A. (2016) You only look once: Unified, real-time object detection. In: CVPR, pp. 779–788

[g010]. Redmon J., Farhadi A. (2017) YOLO9000: Better, faster, stronger. In: CVPR
 [g011]. Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C., Berg A. (2016) SSD: single shot multibox detector. In: ECCV, pp. 21–37

[e001] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *IJCV*,103(1):60–79, 2013.
 [e002] H. Wang and C. Schmid. Action recognition with improved trajectories. *InProc.ICCV*, pages 3551–3558,2013
 [e003] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked fisher vectors. In *Proc. ECCV*,pages 581–595, 2014.
 [e004]Two-Stream Convolutional Networks for Action Recognition in Videos
 [e005]Convolutional Two-Stream Network Fusion for Video Action Recognition
 [e006]Temporal Segment Networks: Towards Good Practices for Deep Action Recognition
 [e007]Beyond Short Snippets: Deep Networks for Video Classification Joe
 [e008]D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, Learning Spatiotemporal Features with 3D Convolutional Networks, *ICCV 2015*, PDF.
 [e009]Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, Caffe: Convolutional Architecture for Fast Feature Embedding, *arXiv* 2014.
 [e010]A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, Large-scale Video Classification with Convolutional Neural Networks, *CVPR 2014*.
 [e011] A Key Volume Mining Deep Framework for Action Recognition
 [e012]Deep Temporal Linear Encoding Networks

[b001] C. Huang, Y. Li, and R. Nevatia, “Multiple target tracking by learning based hierarchical association of detection responses,” *Pattern Analysis & Machine Intelligence IEEE Transactions on*,vol. 35, no. 4, pp. 898- 910, 2013
 [b002]X. Wang, Q.Wang, “Coupled data association and l1 minimization for multiple object tracking under occlusion,”*SPIE/COS Photonics Asia. International Society for Optics and Photonics*, vol. 9273, 2014.
 [b003]Y. Li, C. Huang, and R. Nevatia, “Learning to associate: Hybridboosted multi-target tracker for crowded scene,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 2953– 2960.
 [b004] B. Yang and R. Nevatia, “Online learned discriminative partbased appearance models for multi-human tracking,” in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 484– 498.
 [b005] C.-H. Kuo, C. Huang, and R. Nevatia, “Multi-target tracking by on-line learned discriminative appearance models,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 685– 692.
 [b006] W. Brendel, M. Amer, and S. Todorovic, “Multiobject tracking as maximum weight independent set,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern*

Recognit., 2011, pp. 1273–1280.

[b007] D. Mitzel, E. Horbert, A. Ess, and B. Leibe, “Multi-person tracking with sparse detection and continuous segmentation,” in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 397–410.

[b008] Y. Liu, H. Li, and Y. Q. Chen, “Automatic tracking of a large number of moving targets in 3d,” in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 730–742.

[b009] V. Takala and M. Pietikainen, “Multi-object tracking using color, texture and motion,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–7.

[b010] M. Yang, F. Lv, W. Xu, and Y. Gong, “Detection driven adaptive multi-cue integration for multiple human tracking,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1554–1561.

[b011] B. Song, T.-Y. Jeng, E. Staudt, and A. K. Roy-Chowdhury, “A stochastic graph evolution framework for robust multi-target tracking,” in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 605–619.

[b012] J. Giebel, D. M. Gavrila, and C. Schnorr, “A bayesian framework ” for multi-cue 3d object tracking,” in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 241–252.

[b013] J. Berclaz, F. Fleuret, and P. Fua, “Robust people tracking with global trajectory optimization,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 744–750.

[b014] H. Izadinia, I. Saleemi, W. Li, and M. Shah, “(MP)2T: Multiple people multiple parts tracker,” in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 100–114.

[b015] D. M. Gavrila and S. Munder, “Multi-cue pedestrian detection and tracking from a moving vehicle,” *Int. J. Comput. Vis.*, vol. 73, no. 1, pp. 41–59, Jan. 2007.

[b016] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, “Robust tracking-by-detection using a detector confidence particle filter,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 1515–1522.

[b017] K. Shafique, M. W. Lee, and N. Haering, “A rank constrained continuous formulation of multi-frame multi-target tracking problem,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[b018] Q. Yu, G. Medioni, and I. Cohen, “Multiple target tracking using spatio-temporal markov chain monte carlo data association,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[b019] B. Yang and R. Nevatia, “Multi-target tracking by online learning of non-linear motion patterns and robust appearance models,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1918–1925.

[b020] B. Yang and R. Nevatia, “An online learned CRF model for multitarget tracking,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2034–2041.

[b021] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, “You’ll never walk alone: Modeling social behavior for multi-target tracking,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2009, pp. 261–268.

[b022] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, “Who are you with and where are you going?” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern*

Recognit., 2011, pp. 1345–1352.

[b023] Z. Qin and C. R. Shelton, “Improving multi-target tracking via social grouping,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2012, pp. 1972–1978.

[b024] W. Choi and S. Savarese, “Multiple target tracking in world coordinate with single, minimally calibrated camera,” in Proc. Eur. Conf. Comput. Vis., 2010, pp. 553–567.

[b025] P. Scovanner and M. F. Tappen, “Learning pedestrian dynamics from the real world,” in Proc. IEEE Int. Conf. Comput. Vis., 2009, pp. 381–388.

[b026] S. Pellegrini, A. Ess, and L. Van Gool, “Improving data association by joint modeling of pedestrian trajectories and groupings,” in Proc. Eur. Conf. Comput. Vis., 2010, pp. 452–465.

[b027] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, “Crowd analysis: A survey,” *Mach. Vis. Appl.*, vol. 19, no. 5, pp. 345–357, May 2008.

[b028] L. Kratz and K. Nishino, “Tracking with local spatio-temporal motion patterns in extremely crowded scenes,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2010, pp. 693–700.

[b029] X. Zhao, D. Gong, and G. Medioni, “Tracking using motion patterns for very crowded scenes,” in Proc. Eur. Conf. Comput. Vis., 2012, pp. 315–328.

[b030] L. Kratz and K. Nishino, “Tracking pedestrians using local spatiotemporal motion patterns in extremely crowded scenes,” *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 34, no. 5, pp. 987–1002, May 2012.

[b031] M. Rodriguez, S. Ali, and T. Kanade, “Tracking in unstructured crowded scenes,” in Proc. IEEE Int. Conf. Comput. Vis., 2009, pp. 1389–1396.

[b032] S. Ali and M. Shah, “Floor fields for tracking in high density crowd scenes,” in Proc. Eur. Conf. Comput. Vis., 2008, pp. 1–14.

[b033] T. E. Fortmann, Y. Bar-Shalom, and M. Scheffe, “Sonar tracking of multiple targets using joint probabilistic data association,” *IEEE J. Ocean. Eng.*, vol. 8, no. 3, pp. 173–184, Mar. 1983.

[b034] Y. Ban, S. Ba, X. Alameda-Pineda, and R. Horaud, “Tracking multiple persons based on a variational Bayesian model,” in Proc. Eur. Conf. Comput. Vis. Workshops, 2016, pp. 52–67.

[b035] M. Rodriguez, J. Sivic, I. Laptev, and J.-Y. Audibert, “Data-driven crowd analysis in videos,” in Proc. IEEE Int. Conf. Comput. Vis., 2011, pp. 1235–1242.

[b036] D. Reid, “An algorithm for tracking multiple targets,” *IEEE Trans. Autom. Control*, vol. 24, no. 6, pp. 843–854, Jun. 1979.

[b037] D. Mitzel and B. Leibe, “Real-time multi-person tracking with detector assisted structure propagation,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops, 2011, pp. 974–981.

[b038] W. Hu, X. Li, W. Luo, X. Zhang, S. Maybank, and Z. Zhang, “Single and multiple object tracking using log-euclidean riemannian subspace and block-division appearance model,” *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 34, no. 12, pp. 2420–2440, Dec. 2012.

[b039] Y. Jin and F. Mokhtarian, “Variational particle filter for multiobject tracking,”

- in Proc. IEEE Int. Conf. Comput. Vis., 2007, pp. 1–8.
- [b040] C. Yang, R. Duraiswami, and L. Davis, “Fast multiple object tracking via a hierarchical particle filter,” in Proc. IEEE Int. Conf. Comput. Vis., 2005, pp. 212–219.
- [b041] R. Hess and A. Fern, “Discriminatively trained particle filters for complex multi-object tracking,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2009, pp. 240–247.
- [b042] B. Han, S.-W. Joo, and L. S. Davis, “Probabilistic fusion tracking using mixture kernel-based bayesian filtering,” in Proc. IEEE Int. Conf. Comput. Vis., 2007, pp. 1–8.
- [b043] G. Shu, A. Dehghan, O. Oreifej, E. Hand, and M. Shah, “Partbased multiple-person tracking with partial occlusion handling,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2012, pp. 1815–1821.
- [b044] B. Wu and R. Nevatia, “Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors,” *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, Feb. 2007.
- [b045] J. Xing, H. Ai, and S. Lao, “Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2009, pp. 1200–1207.
- [b046] C. Huang, B. Wu, and R. Nevatia, “Robust object tracking by hierarchical association of detection responses,” in Proc. Eur. Conf. Comput. Vis., 2008, pp. 788–801.
- [b047] V. Reilly, H. Idrees, and M. Shah, “Detection and tracking of large number of targets in wide area surveillance,” in Proc. Eur. Conf. Comput. Vis., 2010, pp. 186–199.
- [b048] A. A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu, “Multi-object tracking through simultaneous long occlusions and split-merge conditions,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2006, pp. 666–673.
- [b049] J. K. Wolf, A. M. Viterbi, and G. S. Dixon, “Finding the best set of k paths through a trellis with application to multitarget tracking,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 25, no. 2, pp. 287–296, Feb. 1989.
- [b050] H. Jiang, S. Fels, and J. J. Little, “A linear programming approach for multiple object tracking,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., 2007, pp. 1–8.
- [b051] J. Berclaz, F. Fleuret, and P. Fua, “Multiple object tracking using flow linear programming,” in Proc. IEEE Int. Workshop Perform. Eval. Track. Surveillance, 2009, pp. 1–8.
- [b052] A. Andriyenko and K. Schindler, “Globally optimal multi-target tracking on a hexagonal lattice,” in Proc. Eur. Conf. Comput. Vis., 2010, pp. 466–479.
- [b053] B. Leibe, K. Schindler, and L. Van Gool, “Coupled detection and trajectory estimation for multi-object tracking,” in Proc. IEEE Int. Conf. Comput. Vis., 2007, pp. 1–8.
- [b054] W. Choi and S. Savarese, “A unified framework for multi-target tracking and collective activity recognition,” in Proc. Eur. Conf. Comput. Vis., 2012, pp. 215–230.
- [b055] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, “Multiple object tracking using

k-shortest paths optimization,” *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 33, no. 9, pp. 1806–1819, Sep. 2011.

[b056] Z. Wu, T. H. Kunz, and M. Betke, “Efficient track linking methods for track graphs using network-flow and set-cover techniques,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1185–1192.

[b057] S. Tang, B. Andres, M. Andriluka, and B. Schiele, “Subgraph decomposition for multi-target tracking,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5033–5041.

[b058] —, “Multi-person tracking by multicut and deep matching,” in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016.

[b059] L. Zhang, Y. Li, and R. Nevatia, “Global data association for multi-object tracking using network flows,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[b060] H. Pirsaviash, D. Ramanan, and C. C. Fowlkes, “Globally-optimal greedy algorithms for tracking a variable number of objects,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1201–1208.

[b061] A. Butt and R. Collins, “Multi-target tracking by lagrangian relaxation to min-cost network flow,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1846–1853.

[b062] Z. Wu, A. Thangali, S. Sclaroff, and M. Betke, “Coupling detection and data association for multiple object tracking,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1948–1955.

[b063] P. Lenz, A. Geiger, and R. Urtasun, “FollowMe: Efficient online min-cost flow tracking with bounded memory and computation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4364–4372.

[b064] B. Yang, C. Huang, and R. Nevatia, “Learning affinities and dependencies for multi-target tracking using a CRF model,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1233–1240.

[b065] A. Milan, K. Schindler, and S. Roth, “Detection- and trajectory-level exclusion in multiple object tracking,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3682–3689.

[b066] N. Le, A. Heili, and J.-M. Odobez, “Long-term time-sensitive costs for crf-based tracking by detection,” in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2016.

[c001] M. S. Ryoo, “Human activity prediction: Early recognition of ongoing activities from streaming videos,” in *ICCV*, 2011.

[c002] Y. Cao, D. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. Dickinson, J. Siskind, and S. Wang, “Recognizing human activities from partially observed videos,” in *CVPR*, 2013.

[c003] K. Li, J. Hu, and Y. Fu, “Modeling complex temporal composition of actionlets for activity prediction,” in *ECCV*, 2012.

[c004] M. Hoai and F. D. la Torre, “Max-margin early event detectors,” in *CVPR*, 2012.

[c005] B. Zhou, X. Wang, and X. Tang, “Random field topic model for semantic

- region analysis in crowded scenes from tracklets,” in CVPR, 2011.
- [c006] B. Morrisand and M. Trivedi, “Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 11, pp. 2287–2301, 2011
- [c007] K. Kim, D. Lee, and I. Essa, “Gaussian process regression flow for analysis of motion trajectories,” in ICCV, 2011.
- [c008] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, “Social gan: Socially acceptable trajectories with generative adversarial networks,” in CVPR, 2018.
- [c009] H. Su, J. Zhu, Y. Dong, and B. Zhang, “Forecast the plausible paths in crowd scenes,” in IJCAI, 2017.
- [c010] J. F. P. Kooij, N. Schneider, F. Flohr, and D. M. Gavrila, “Contextbased pedestrian path prediction,” in *European Conference on Computer Vision*. Springer, 2014, pp. 618–633.
- [c011] P. Abbeel and A. Ng, “Apprenticeship learning via inverse reinforcement learning,” in ICML, 2004.
- [c012] B. Ziebart, A. Maas, J. Bagnell, and A. Dey, “Maximum entropy inverse reinforcement learning,” in AAAI, 2008.
- [c013] Y. Kong, D. Kit, and Y. Fu, “A discriminative model with multiple temporal scales for action prediction,” in ECCV, 2014.
- [c014] Y. Kong, Z. Tao, and Y. Fu, “Deep sequential context networks for action prediction,” in CVPR, 2017.
- [c015] Y. Kong, S. Gao, B. Sun, and Y. Fu, “Action prediction from videos via memorizing hard-to-predict samples,” in AAAI, 2018.
- [c016] S. Ma, L. Sigal, and S. Sclaroff, “Learning activity progression in lstms for activity detection and early detection,” in CVPR, 2016.
- [c017] K. Simonyan and A. Zisserman, “two-stream convolutional networks for action recognition in videos,” in NIPS, 2014.
- [c018] H. S. Koppula and A. Saxena, “Anticipating human activities using object affordances for reactive robotic response,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2016.
- [c019] M. Pei, Y. Jia, and S.-C. Zhu, “Parsing video events with goal inference and intent prediction,” in ICCV. IEEE, 2011, pp. 487–494.
- [c020] K. Li and Y. Fu, “Prediction of human activity by discovering temporal sequence patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1644–1657, Aug 2014.
- [c021] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, “Semantic-based surveillance video retrieval,” *Image Processing, IEEE Transactions on*, vol. 16, no. 4, pp. 1168–1181, 2007.
- [c022] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, “Social lstm: Human trajectory prediction in crowded spaces,” in CVPR, 2016.
- [c023] N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker, “Desire: Distant future prediction in dynamic scenes with interacting agents,” in

CVPR, 2017.

[c024] C. Finn, S. Levine, and P. Abbeel, “Guided cost learning: deep inverse optimal control via policy optimization,” in arXiv preprint arXiv:1603.00448, 2016.

[c025] M. Wulfmeier, D. Wang, and I. Posner, “Watch this: scalable cost function learning for path planning in urban environment,” in arXiv preprint arXiv:1607.02329, 2016.

[k001] Shou Z, Wang D, Chang S F. Temporal action localization in untrimmed videos via multi-stage cnns[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1049-1058.

[k002] Shou Z, Chan J, Zareian A, et al. CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos[J]. arXiv preprint arXiv:1703.01515, 2017.

[k003] Xiong Y, Zhao Y, Wang L, et al. A Pursuit of Temporal Accuracy in General Activity Detection[J]. arXiv preprint arXiv:1703.02716, 2017.

[k004] Gao J, Yang Z, Sun C, et al. TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals[J]. arXiv preprint arXiv:1703.06189, 2017.

[k005] Wang L, Xiong Y, Lin D, et al. UntrimmedNets for Weakly Supervised Action Recognition and Detection[J]. arXiv preprint arXiv:1703.03329, 2017.