基于深度学习的视觉跟踪算法综述

Survey on Visual Tracking Algorithms Based on Deep Learning Technologies

作者：

学号：

专业：

日期：2021-06-17

关键词：

计算机视觉；视觉跟踪；深度学习；

摘要：

视觉跟踪是计算机视觉的重要研究领域之一，传统的视觉跟踪算法难以很好地解决复杂背景中的跟踪问题，如光线变化、目标发生较大的尺寸和姿态变化或目标被遮挡等。而深度学习的引入为视觉跟踪研究开辟了新的途径，本文简要介绍了视觉跟踪和深度学习的研究现状，重点分析了基于深度学习的视觉跟踪算法的相关文献，讨论了各算法的优缺点，最后提出了进一步研究的方向以及对基于深度学习的视觉跟踪算法的展望。

目录

一、研究意义与背景

视觉跟踪也称为目标跟踪，是指对图像（视频）序列中的目标进行检测、提取、识别和跟踪，获得目标的运动参数，如位置、速度、加速度和运动轨迹等，并对其进一步处理与分析，实现对目标的行为理解，完成更高一级的任务。近年来，作为计算机视觉领域的重要研究方向，有关视觉跟踪的一些先进的思想和算法陆续被提出，解决了视觉跟踪上的部分难题，并取得了一定的研究成果。但是，在较复杂的情形下，如目标被长时间遮挡、光线剧烈变化、背景混杂、较大的尺寸和姿态变化等，视觉跟踪面临巨大的挑战，无法满足鲁棒性、实时性和准确性的要求。因此，为了克服这一挑战，深度学习被尝试引入视觉跟踪领域。

深度学习是机器学习领域中较新的研究领域，由Geoffrey Hinton于2006年提出[1]，之后在诸多领域取得了巨大的成功，但其在视觉跟踪领域的应用却是在2010年之后。本文将首先介绍目前比较流行的传统的视觉跟踪算法并将其归类，然后重点研究自2010年以来出现的基于深度学习的视觉跟踪算法，总结各种算法的思想特点以及各自的优缺点。最后针对所做的分析，指出目前仍需解决的问题和未来的研究方向，并对深度学习在视觉跟踪上的应用予以展望。

二、深度学习在视觉跟踪上的应用

2.1 传统的视觉跟踪算法

从目标表征模型的角度，视觉跟踪算法可大致分为生成式和判别式。生成式算法将跟踪视为一个目标匹配的过程：生成一个目标的模板，寻找匹配程度最高或重构误差最小的（即后验概率最大的）位置作为目标位置的预测。目前较流行的生成式算法主要有核方法[2]、稀疏表达[3−4]、密度估计[5]、增量学习[6]、高斯混合模型[7]等，部分算法通过结合卡尔曼滤波、粒子滤波[1]达到了较好的跟踪效果。[2]采用各向同性核函数的均值漂移跟踪算法，融合了特征直方图、相似性度量和卡尔曼滤波等方法，取得了一定的效果。[4]对目标的历史和局部信息进行稀疏编码，采用空间金字塔和均值传递方法使算法具有较好的鲁棒性。[5]基于均值漂移模式查找算法提出一种新的核密度估计技术，即随着时间的推移，密度模式依次传播。实验证明将序列核密度估计运用于目标外观建模时具有一定的准确性和紧凑性。[6]采用增量的方式来学习和适应低维特征空间的表示，从而反映出目标外观的变化，并结合马尔卡夫链蒙特卡洛方法和粒子滤波方法来完成最终的目标跟踪。

判别式算法将目标跟踪转化为一个二分类问题，通过训练分类器，使其能够在图像（视频）序列中区分目标和背景，从而实现跟踪。近年来常用的判别式算法包括支持向量机（SVM）[9−10]、随机森林[11]、多样本学习[12]、Boosting[13−14]等。[11]在简单模板和均值漂移光流的基础上，引入随机森林算法，使用学习到的高置信度的特征来更新模板，以提高跟踪算法的自适应性。[12]将多个样本进行分类学习，以确定目标表征模型，从而解决了样本歧义问题。[14]提出了在线半监督学习强分类器，对目标特征进行选择，在一定程度上缓解了跟踪漂移和遮挡问题。

但是，上述算法普遍存在着局限性，即大多依赖于低级的手工设计的特征，而无法提取高级的语义信息。手工设计特征需要专家知识，对特定场景的适应度往往很好，但其泛化能力较弱，无法扩展到更普遍的情况。为了弥补这一缺陷，引入了深度学习。深度模型具有强大的学习能力、高效的特征表达能力和获取高级语义信息的能力，这为视觉跟踪带来了新的研究思路。

2.2 基于深度学习的视觉跟踪算法

深度学习因其深层结构而可以从大量数据中主动学习特征，从而避免了传统方法手工设计提取特征的缺点，并成功运用于图像分类[15]、人脸识别[16]、物体检测[17]等领域。现有的深度学习架构主要有3种，分别为深度卷积神经网络（CNN）、深度置信网络（DBN）、堆栈自编码器（SAE）。

CNN这种深层网络结构对平移、比例缩放、倾斜或者共他形式的变形具有高度不变性，因此能够提取对平移、缩放和旋转不变的观测数据的显著特征，已广泛应用于语音识别、图像识别等众多领域。

DBN无论在时间还是算法效率上都有很好的效果，且用连续层的二进制或真值的变量来学习高层表示的分布，因此具有很好的灵活性。目前，DBN已被成功应用于不同的领域中，如文本表示、音频事件分类以及可视化数据分析等。

SAE的训练方法与DBN类似，采用逐层训练的方法，最后得到的SAE与其他深度神经网络一样，具有强大的表达能力。它能够学到输入的多层次表达和“部分-整体分解”结构，第一层能够提取原始输入的一阶特征（比如图片边缘），第二层可以获得二阶特征（比如物体轮廓），更高层会学到更高阶的特征（比如识别人的眼睛）。因此，SAE能够进行快速文件编码、语音特征编码以及基于内容的图像检索。

视觉跟踪的基础是提取目标特征，继而确定其在图像（视频帧）中的位置，完成跟踪任务。一个好的特征表达能够提高系统的整体性能，深度学习恰恰满足了这一要求。然而，基于深度学习的视觉跟踪算法并不像分类、识别和检测那样容易成功。在跟踪时，只对初始帧进行目标标注，在后续的视频序列中根据学到的特征进行目标定位，由于目标的变形、场景变化、遮挡等因素容易产生跟踪漂移，导致目标丢失。即使如此，依然相继出现了基于深度学习的视觉跟踪算法，在跟踪精度和成功率上都取得了比传统算法更好的效果。

目前，基于深度学习的视觉跟踪算法从采用的深度模型来看，可分为基于CNN的跟踪算法[18−24]和基于SAE的跟踪算法[31−32]，其中部分综合了粒子滤波、支持向量机（SVM）和AdaBoost等传统算法。所采用的算法分类如表1所列。

表1 基于深度学习的视觉跟踪算法分类

深度学习模型	SVM	AdaBoost	粒子滤波	相关滤波	N/A
CNN	[19]		[24]	[21]	[18, 20, 22 – 23]
SAE		[25]	[25 – 26]		

其中，N/A表示只采用了深度学习模型而没有结合传统算法。在图像和视觉领域，CNN是应用较为成功的一个深度模型，如著名的深度网络AlexNet[15]，RCNN[27]，Deep2D-Net[28]，GoogleNet[29]和VGG-Net[30]都采用了CNN。另外还有其他算法采用了已训练成熟的CNN不同架构的深度网络，如表2所列。

表2 基于CNN不同架构的视觉跟踪算法分类

	2010	2014	2015	2016
CNN-DIY	[18]	[20]		[24]
RCNN			[19]	
VGG-Net			[21 – 23]	

表2中，2011-2013年没有出现基于CNN的视觉跟踪论文，2013年，[25]提出了基于SAE的视觉跟踪算法。从中可以看出，大部分算法都是基于已训练成熟的CNN不同架构的深度网络提出相应的视觉跟踪算法。

2.2.1 基于CNN的视觉跟踪算法

从表1可以看出，当前视觉跟踪算法中所采用的深度模型大多为CNN，这与CNN的研究发展和自身结构特点有关。因为CNN采用局部感受野和权值共享，所以其具有平移不变性、光照不变性以及对遮挡的鲁棒性等重要特征。

基于单纯CNN的跟踪算法没有结合传统算法而完成了跟踪任务。CNN深层网络结构在提取目标特征的同时，可以将目标和背景分类，从而从背景中判别出目标。[18, 20, 22 – 23]只采用了CNN模型，通过不同的深度结构、选择策略、训练方法和模型更新等，提出了具有不同特点的跟踪算法。

[18]基于3个卷积层和若干降采样层的CNN对当前帧和上一帧进行采样，获取目标和背景的空间、时间特征。在首个降采样层后，将输出的特征图分别送入两个通道，一个提取全局信息，另一个提取局部信息，共同构成目标的概率图。采用了两个采样对，分别输入两个CNN网络，得到4张关键点的概率图，从而提高了跟踪的精确度。该算法于2010年被提出，具有一定的开创性，但只给出了以人作为跟踪目标的示例。

[20]采用的CNN虽然只包含了两个卷积层、两个降采样层以及最后的全连接层，但其输入为目标图像的4个通道的信息，分别进入各自的网络，即共有4个并列的CNN，并在全连接层进行综合，得到输出向量。最后根据该向量来判别样本是目标还是背景，从而估计目标位置。在训练网络时，加入截断损失函数来保持尽可能多的训练样本并降低跟踪误差累积的风险。采用迭代随机梯度下降，引入时间序列选择机制，以确保跟踪目标时不产生漂移。该算法对目标和背景信息进行较详细的提取，具有很好的判别能力，但因缺乏目标全局的高级语义信息，在目标发生急速运动变化、长时间遮挡时无法达到较好的鲁棒性。

[22]所用的深度模型分为共享层和特定层，其中共享层采用VGG-Net，截取3个卷积层和2个全连接层；特定层由若干域组成，包含了目标正样本和负样本。为了估计当前帧的目标状态，在上一帧目标位置的周围取N个候选目标区域，分别计算出每个候选区域属于正样本和负样本的得分，将属于正样本的最高得分的候选区域作为当前帧的目标区域，用邻接矩形回归模型调整目标状态。同时使用长短时间更新策略，使跟踪趋于鲁棒性和适应性。

[23]也将VGG-Net网络应用到所提算法中，另外添加了一般性网络（GNet）和特殊性网络（SNet），两者具有相同的结构，都含有2个卷积层。将VGG的Conv4-3（第10卷积层）的输出作为SNet的输入，Conv5-3（第13卷积层）的输出作为GNet的输入。论文中通过分析大量实验发现，Conv4-3的特征图保存了更多中间层次的信息，能够更加精确地将属于同一类别的不同图像区分开来；Conv5-3的特征图保存了更多高级语义信息，能够将人脸和非人脸区分开来。GNet和SNet分别生成各自的前景热图，而最终目标定位是通过一个干扰项检测机制来决定使用哪一个热图。该算法的亮点在于没有将VGG看成一个黑箱，而是分析不同层提取的信息特点，进而利用了深度网络的一个重要优点：不同层编码不同类型的特征，底层编码更适合有区分性的细节特征，高层更适合提取高级语义的目标类别特征。同时，在GNet和SNet前，通过选择网络（sel-Net）对输入的特征图进行选择，除去了许多不相干的噪声特征图，这些使得该算法具有较好的判别力和鲁棒性。

混合CNN是在CNN深度模型的基础上使用了传统的视觉跟踪算法，如粒子滤波、SVM等。过深的网络结构不利于参数训练，且对定位精度处理得不好，不能精确地跟踪目标。由于随着网络的加深，空间信息会被稀释，因此这些算法尝试在CNN的架构上引入传统经典算法来进一步提高算法的性能。从实验结果看，都取得了不错的效果。

[19]采用RCNN深度模型，并添加一个分类层SVM，它能够利用从CNN模型学到的特定目标的特征来从背景中区分出跟踪对象。通过SVM选出目标相关的正样本，并将其包含的目标特征（由SVM权重系数确定）沿着CNN模型反向传播到输入层，从而得到每一个正样本的显著图。这些图汇集成目标的显著图，最后使用它来构建观测模型，并采用贝叶斯滤波器进行跟踪。其综合了判别式和生成式算法，对目标定位的精确度和鲁棒性进行了较大的改进。同时，显著图可以有效地可视化目标的空间结构，所以在一定程度上提高了目标定位的精度，并且能够实现像素级的目标分割。

[21]基于VGG-Net深度模型，将输入的目标图片按不同的层次提取特征，即把网络中Conv3-4，Conv4-4，Conv5-4层输出的特征图作为候选，用来估计目标的位置。论文指出，不同层次的输出携带着不同的信息：越靠前的层含有越多的空间细节信息，越往后的层包含越多的语义信息。分别采用相关滤波器来处理3个层次（通道）的输出，计算得到各自的相关

响应图。最后使用粗细转换估计策略，即用最后一层（Conv5-4）相关响应图的最大值位置去逐层搜索最前一层（Conv3-4）的最大值位置，从而以最佳空间分辨率估计目标位置。此算法利用多层次的特征输出，综合相关滤波器，得到了较精确的目标位置，并对目标外观变化具有鲁棒性。

[24]在初始帧对目标域进行采样，经变形和归一化处理后，采用K-means算法提取得到一系列目标域的局部样本，将其作为固定的滤波器集。跟踪时，得到当前帧的采样结果，同时对上一帧中预测目标位置周围的背景进行采样。之前得到的滤波器集分别对处理后的背景样本做差，用差值分别对当前帧的样片进行卷积，提取局部选择特征，将这些特征组合在一起形成目标的全局外观表达。该算法采用粒子滤波方法，利用CNN得到的全局特征图对当前目标位置进行预测。提出的滤波器集可以保留目标的每一局部细节，从而避免了跟踪漂移，但同时增加了大量的样片预处理计算。

2.2.2 基于SAE的视觉跟踪算法

从表1中可以看出，当前采用SAE深度模型的跟踪算法不是很多，而且都与传统的经典算法进行综合使用。这主要是因为SAE通过隐层来学习一个数据的表示或对原始数据进行编码，采用一种不利用分类标签的非线性特征提取方法，目的在于保留和获取更好、更有效的信息表示，而不是对信息分类。视觉跟踪本身就需要将目标从背景中区分开来，因此，目标跟踪并不是SAE的强项。但[25 – 26]尝试利用SAE的高效的信息表示能力，综合传统方法提出自己的算法，取得了比传统跟踪算法更好的效果。

[25]采用SAE深度模型，包含了4个隐藏层和额外的分类层，为了增强其对输入数据编码的鲁棒性，在输入中引入高斯噪声干扰，此时SAE称为堆栈降噪自编码器（SDAE）。第一层使用过完备滤波器，使其能更好地捕获图像结构。预训练SDAE的第一层时，为了加快训练速度，将输入图像分割成5张图片，然后对应训练5个DAE，用这5个DAE的权值来初始化SDAE的第一层，其余各层用普通方法训练。跟踪时，采用粒子滤波器框架，SDAE提取目标特征并将其作为粒子滤波器的状态变量，同时计算每个粒子的置信度，从而确定目标位置。该算法与传统方法相比使用了多种非线性变换，因此得到的图像表达比之前基于PCA的方法更具有表现性，且不需要像基于稀疏编码的方法那样进行解优化，提高了算法的实时性。

[26]也是基于SDAE深度架构，采用的隐层数与[25]相同，同时也在第一层使用了过完备滤波器。该算法的不同之处在于在每个隐层的顶部添加一个sigmoid分类层，即DNN分类器，这些分类器组合成具有强大分类能力的AdaBoost。通过AdaBoost，计算得到包含候选目标区域的置信图，其中，置信度最大的区域为目标的位置。该算法综合了粒子滤波方法，从实验结果分析，在不同的场景下，该算法的跟踪效果比[25]的更好。

2.3 算法比较

跟踪时，各算法具有不同的更新策略：适应更新和判别更新。适应更新指经过一定时间或帧数后根据当前目标状态对模型进行更新或每帧都更新，如[19, 21, 24]。判别更新即当检测到目标发生较大变化时更新模型，否则不更新，如[20, 24 – 25]。也有算法同时采用两种更新策略，如[18, 26]。适应更新容易产生目标漂移，但实时性好；判别更新不易丢失目标，但跟踪精度和实时性不高。因此，不同的更新策略在不同的情形下有着不同的优势和缺陷。

基于深度学习的视觉跟踪算法虽然是近几年才被提出，但其跟踪效果已超过了传统算法。部分比较结果如表3所示：

表3 传统跟踪算法与深度学习跟踪算法比较

	VTD	**CXT**	**ASLA**	**SCM**	**DL**
精确率	0.576	0.575	0.532	0.649	0.852
成功率	0.416	0.426	0.434	0.499	0.597

其中，DL为深度学习跟踪算法，其他为传统跟踪算法。显然，深度学习跟踪算法比传统跟踪算法具有更好的跟踪性能。该数据综合了多种复杂场景下的跟踪效果，具有一定的普适性。

三、未来发展展望

本文研究了目前基于深度学习的视觉跟踪算法，对其进行了概述和总结，并发现基于深度学习的视觉跟踪算法比传统的跟踪算法有着更好的鲁棒性，而且深度学习与传统跟踪算法的结合往往能获得意想不到的效果。

与图像分类和语音识别相比，深度学习在视觉跟踪领域的应用还未成熟，虽然目前取得了不错的效果，但还有很大的改进空间。我认为基于深度学习的跟踪技术至少还有两个问题需要解决：一是如何根据视觉跟踪的特点，利用领域知识研究并建立更适合于视觉跟踪任务的深度结构模型，在减少计算量的同时保证跟踪精度；二是视觉跟踪具有空间和时间的相关性，如何使用深度学习模型来获取这种相关性。

四、参考文献

[1] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 7 (July 2006), 1527–1554. DOI:https://doi.org/10.1162/neco.2006.18.7.1527

[2] D. Comaniciu, V. Ramesh and P. Meer, "Kernel-based object tracking," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 5, pp. 564-577, May 2003, doi: 10.1109/TPAMI.2003.1195991.

[3] X. Jia, H. Lu and M. Yang, "Visual tracking via adaptive structural local sparse appearance model," 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1822-1829, doi: 10.1109/CVPR.2012.6247880.

[4] Z. Zhang and K. H. Wong, "Pyramid-Based Visual Tracking Using Sparsity Represented Mean Transform," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1226-1233, doi: 10.1109/CVPR.2014.160.

[5] B. Han, D. Comaniciu, Y. Zhu and L. S. Davis, "Sequential Kernel Density Approximation and Its Application to Real-Time Visual Tracking," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 7, pp. 1186-1197, July 2008, doi: 10.1109/TPAMI.2007.70771.

[6] Ross, David A. et al. “Incremental Learning for Robust Visual Tracking.” International Journal of Computer Vision 77 (2007): 125-141.

[7] A. D. Jepson, D. J. Fleet and T. F. El-Maraghi, "Robust online appearance models for visual tracking," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 10, pp. 1296-1311, Oct. 2003, doi: 10.1109/TPAMI.2003.1233903.

[8] D. Varas and F. Marques, "Region-Based Particle Filter for Video Object Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3470-3477, doi: 10.1109/CVPR.2014.444.

[9] S. Avidan, "Support vector tracking," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 26, no. 8, pp. 1064-1072, Aug. 2004, doi: 10.1109/TPAMI.2004.53.

[10] Y. Bai and M. Tang, "Robust tracking via weakly supervised ranking SVM," 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1854-1861, doi: 10.1109/CVPR.2012.6247884.

[11] J. Santner, C. Leistner, A. Saffari, T. Pock and H. Bischof, "PROST: Parallel robust online simple tracking," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 723-730, doi: 10.1109/CVPR.2010.5540145.

[12] B. Babenko, M. Yang and S. Belongie, "Robust Object Tracking with Online Multiple Instance Learning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 8, pp. 1619-1632, Aug. 2011, doi: 10.1109/TPAMI.2010.226.

[13] H. Grabner and H. Bischof, "On-line Boosting and Vision," 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006, pp. 260-267, doi: 10.1109/CVPR.2006.215.

[14] Grabner H., Leistner C., Bischof H. (2008) Semi-supervised On-Line Boosting for Robust Tracking. In: Forsyth D., Torr P., Zisserman A. (eds) Computer Vision – ECCV 2008. ECCV 2008. Lecture Notes in Computer Science, vol 5302. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-88682-2_19

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (June 2017), 84–90. DOI:https://doi.org/10.1145/3065386

[16] Sun, Yi, et al. "Deeply Learned Face Representations Are Sparse, Selective, and Robust." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2892–2900.

[17] Girshick, Ross & Donahue, Jeff & Darrell, Trevor & Malik, Jitendra. (2013). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 10.1109/CVPR.2014.81.

[18] Fan J, Xu W, Wu Y, Gong Y. Human tracking using convolutional neural networks. IEEE Trans Neural Netw. 2010 Oct;21(10):1610-23. doi: 10.1109/TNN.2010.2066286. Epub 2010 Aug 30. PMID: 20805052.

[19] Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. 2015. Online tracking by learning discriminative saliency map with convolutional neural network. In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15). JMLR.org, 597–606.

[20] Li, Hanxi & Li, Yi & Porikli, Fatih. (2014). Robust Online Visual Tracking with a Single Convolutional Neural Network. 194-209. 10.1007/978-3-319-16814-2_13.

[21] Ma, Chao, et al. "Hierarchical Convolutional Features for Visual Tracking." 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3074–3082.

[22] H. Nam and B. Han, "Learning Multi-domain Convolutional Neural Networks for Visual Tracking," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4293-4302, doi: 10.1109/CVPR.2016.465.

[23] L. Wang, W. Ouyang, X. Wang and H. Lu, "Visual Tracking with Fully Convolutional Networks," 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 3119-3127, doi: 10.1109/ICCV.2015.357.

[24] K. Zhang, Q. Liu, Y. Wu and M. Yang, "Robust Visual Tracking via Convolutional Networks Without Training," in IEEE Transactions on Image Processing, vol. 25, no. 4, pp. 1779-1792, April 2016, doi: 10.1109/TIP.2016.2531283.

[25] Naiyan Wang and Dit-Yan Yeung. 2013. Learning a deep compact image representation for visual tracking. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1(*NIPS'13*). Curran Associates Inc., Red Hook, NY, USA, 809–817.

[26] X. Zhou, L. Xie, P. Zhang and Y. Zhang, "An ensemble of deep neural networks for object tracking," 2014 IEEE International Conference on Image Processing (ICIP), 2014, pp. 843-847, doi: 10.1109/ICIP.2014.7025169.

[27] Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. The 3rd International Conference on Learning Representations (ICLR2015). https://arxiv.org/abs/1409.1556

[28] Ouyang, Wanli & Luo, Ping & Zeng, Xingyu & Qiu, Shi & Tian, Yonglong & Li, Hongsheng & Yang, Shuo & Wang, Zhe & Xiong, Yuanjun & Qian, Chen & Zhu, Zhenyao & Wang, Ruohui & Loy, Chen Change & Wang, Xiaogang & Tang, Xiaoou. (2014). DeepID-Net: multi-stage and deformable deep convolutional neural networks for object detection. CoRR. abs/1409.3505.

[29] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[30] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the Devil in the Details: Delving Deep into Convolutional Nets. Proceedings of the British Machine Vision Conference. BMVA Press, September 2014.