基于深度学习的行人属性多标签识别

李亚鹏 万遂人#*

(东南大学生物科学与医学工程学院,南京 210096)

摘 要:行人属性通常指的是行人的一些可被观察到的外部特征,如性别、年龄、服饰、携带品等。作为行人外部的软生物特征,行人属性对于行人检测和再识别是非常重要的,并且在智能视频监控场景和基于视频的商业智能应用中显示出巨大的潜力。在目前的行人属性多标签分类识别中,主要有基于手工设计特征的方法和基于深度学习的方法。然而,手工设计特征的方法难以应对复杂的真实视频监控场景,在实际应用中取得的效果并不是很理想。采用深度卷积网络模型,包含 3 个卷积层和 2 个全连接层,使用 Sigmoid 交叉熵损失函数,训练平台为 Caffe 深度学习框架,通过在包含19 000张行人图片的 PETA 数据集上对 10 种行人属性进行训练和测试,得到 85.2% 的平均识别精度。加入正样本比例指数因子改进损失函数后,平均识别精度达到 89.2%,使网络性能有明显的提高。

关键词:深度学习;行人属性;多标签识别

中图分类号:R318 **文献标志码**:A **文章编号**:0258-8021(2018)04-0423-06

Multi-Label Recognition of Pedestrian Attributes Based on Deep Learning

Li Yapeng Wan Suiren#*

(*School of Bioscience and Medical Engineering, Southeast University, Nanjing* 210096, *China*)

Abstract: Pedestrian attributes usually refer to some of the external characteristics of pedestrians that can be observed, such as gender, age, clothing type, carrying objects, etc. As soft biological features of pedestrians, pedestrian attributes are very important for pedestrian detection and re-identification, and show great potential in intelligent video surveillance scenarios and video based business intelligence applications. Among the current multi-label classification methods of pedestrian attributes, two of them are mainly employed, one is based on handcrafted features and the other is based on the deep learning methods. However, the methods of handcrafted features are difficult to deal with complex real video surveillance scenes, results obtained in practical applications are not ideal. In this paper we used a deep convolutional network model with three convolutional layers and two full-connected layers. Using the Sigmoid cross-entropy loss function, the training platform was the Caffe deep learning framework, the dataset used was PETA containing 19,000 pedestrian images. Ten kinds of pedestrian attributes were trained and tested, and an average recognition accuracy of 85.2% was reached. After adding the positive sample proportional exponential factor to improve the loss function, the average recognition accuracy reached 89.2%, which significantly improved the performance of the network.

Key words: deep learning; pedestrian attributes; multi-label recognition

引言

行人属性如性别、年龄、服饰、携带品等,作为行人的外部软生物特征应用在监控领域,已经吸引了大量的关注。例如,行人属性作为有用的线索已经被用来进行人物检索[1-2]、人物识别[3-6]、面部验证[7]和人物再识别[8],并且在智能视频监控场景和基于视频的商业智能应用中显示出巨大的潜力。

doi: 10.3969/j.issn.0258-8021.2018.04.005

收稿日期: 2018-01-18, 录用日期: 2018-04-26

#中国生物医学工程学会会员(Member, Chinese Society of Biomedical Engineering)

* 通信作者(Corresponding author),E-mail: srwan@seu.edu.cn

在许多现实世界的监控场景下，摄像机通常安装在远处以覆盖广泛的区域，因此被捕获的行人图像分辨率较低，难以获得高质量的脸部图像。然而，在这种场景下的行人属性依然有很高的应用潜力，因为相对于传统的生物识别技术，行人属性已经显示出多个优点，比如光照不变性和对比不变性。

行人属性分类面临着 3 个主要的挑战。第一，由于服装外观的多样，照明条件的差异和相机视角的不同，导致严重的类内差异。第二，行人属性具有复杂的局部特征，这意味着一些属性只能在某些确定的或不确定的局部身体区域被识别。例如，长头发和头肩部位最相关，书包可能以不确定的高度出现在图像的左边或者右边。因此，提取行人属性是非常困难的。第三，行人属性分类是多标签分类问题，而不是多类分类问题，因为行人属性不是完全相互排斥的。

当前的行人属性识别方法主要集中在两个应用场景：自然场景和监控场景。许多研究人员非常注意自然场景属性识别，并在目标识别、人脸识别等方面取得巨大成功。例如，自然场景中的属性识别首先由 Ferrari 等提出[9]。他们提出了一种概率生成模型来学习低级视觉属性，如“条纹”和“斑点”。Siddiquie 等对不同查询属性之间的相关性进行明确的建模，并生成了检索列表[10]。Kumar 等探索比较面部特征，并通过二分类器进行面部验证[11]。

监控场景中的属性识别也有一些开创性的研究。Layne 等首先使用支持向量机（SVM）来识别属性（如“性别”，“背包”），并用其促进行人再识别[12-13]。为了解决混合场景中的属性识别问题，朱建清等引入行人数据库（APiS），并使用增强算法来识别属性[14]。邓玉斌等构建了行人属性数据库[15]（PETA），利用 SVM 和马尔可夫随机场识别属性。然而，这些方法都是使用手工特征，并不能有效地代表监视场景中的图像。另外，属性之间的关系被忽略，这对属性识别任务非常重要。例如，长发特征的女性比男性的概率更高，所以头发长度可以帮助识别性别。

受到深度学习在不同传统计算机视觉任务上的突出表现的启发，一些研究人员开始用深度卷积神经网络的方法进行行人属性分类。李党伟等提出了可用于学习不同属性间相关性的卷积神经网络模型（DeepMAR），与传统手工特征方法相比，在行人属性识别精度上取得了更好的结果[16]。朱建清等提出了一个多标签卷积神经网络模型（MLCNN）来进行行人属性识别[17]。Hiroshi 等通过异构学习和稀有率方法提高在数据集不平衡情况下的属性识别率[18]。在行人属性多标签识别任务中，大部分行人属性数据集都存在样本属性分布失衡的问题。受 Levi 等[19]在研究年龄和性别问题所使用模型的启发，本研究使用了一个卷积神经网络模型来实现行人属性多标签分类。该模型以 AlexNet[20]为基础，削减了两个卷积层和一个全连接层，并做了一些改动。该卷积神经网络模型使用 Sigmoid 交叉熵损失函数，并通过加入正样本比例指数因子来应对样本属性分布失衡的问题。通过在 PETA 数据集上进行实验验证，取得了良好的识别效果。

1 材料和方法

1.1 方法

卷积神经网络作为深度学习的一种经典模型，能够从数据中自动学习并提取特征，其泛化能力显著优于传统方法。本课题采用深度卷积神经网络的方法，研究行人属性多标签分类识别。

1.2 Sigmoid 交叉熵损失函数

Sigmoid 函数是一种 S 型函数，可以将神经网络输出端的分类得分转换为相应的输出概率，如式（1）所示，$p_{n,l}$ 为第 n 个样本第 l 个属性的输出概率。对于拥有多个属性的多标签分类，需要综合考虑所有属性的损失，整体的 Sigmoid 交叉熵损失函数如式（2）所示。在数据集中，由于各个属性的分布不平衡比较严重，各个属性正样本在所有样本中所占比例差异也很大。例如，戴帽子属性要比性别属性的正样本比例少很多，因为现实中的行人通常也是不带帽子的居大多数。为了应对属性的严重不平衡分布，提高损失函数对模型的优化能力，在综合考虑每个属性的损失值时引入了正样本比例指数因子 w_l。w_l 表示第 l 个属性损失值的权重，正样本比例越小，该属性损失值越大。此时损失函数值可以由式（3）求出。p_l 是训练集中第 l 个属性正样本所占的比例。实验过程中，式（4）中 σ 参数的值取 1。实验中还测试了 w_l 因子对识别精度的影响。

$$p_{n,l} = 1/(1 + \exp(-x_{n,l})) \tag{1}$$

$$\text{Loss} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L}(y_{n,l}\ln(p_{n,l}) + (1 - y_{n,l})\ln(1 - p_{n,l})) \tag{2}$$

$$\text{Loss} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{L} w_l(y_{n,l}\ln(p_{n,l}) + (1-y_{n,l})\ln(1-p_{n,l})) \quad (3)$$

$$w_l = \exp(-p_l/\sigma^2) \quad (4)$$

1.3　网络结构

本研究使用的模型具有 3 个卷积层和 2 个全连接层组成,模型的网络结构如图 1 所示。

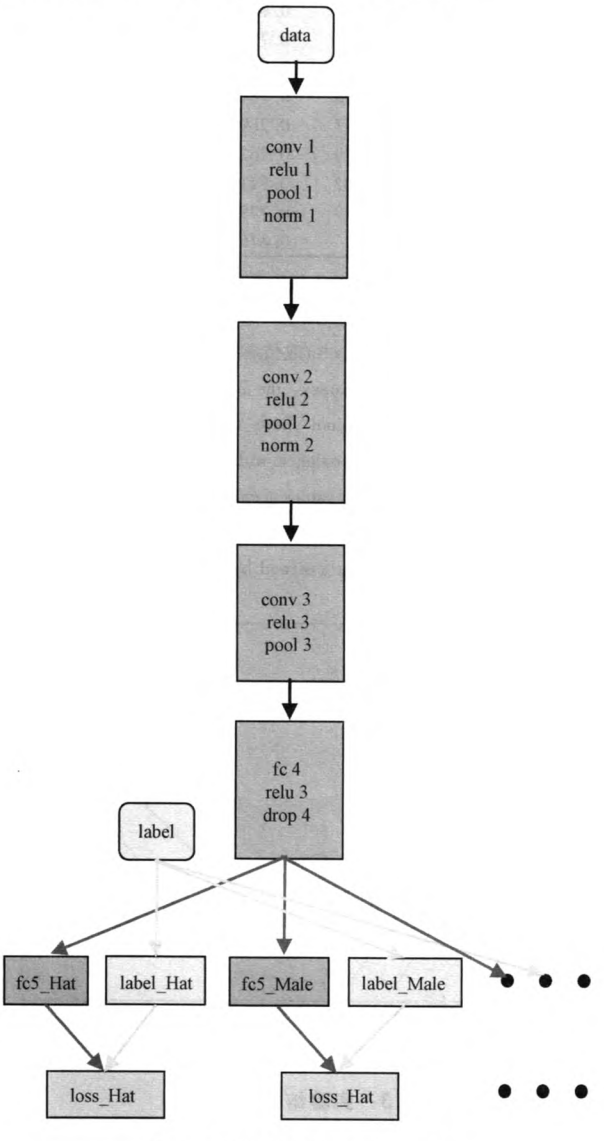

图 1　卷积网络模型结构流程

Fig. 1　The structure chart of the CNN net model

首先,将图像大小调整为 256 × 256,在训练时采用随机剪裁策略以扩充数据集,剪裁尺寸为 227 × 227,剪裁后的图片大小与剪裁前的图片相差不大,一般不会造成图片信息的损失。网络参数采用高斯分布初始化,标准差为 0.01。3 个卷积层和 2 个全连接层的详细定义如下:

1)conv1 包含 96 个滤波器,核尺寸为 7 × 7,步长为 3,填充数为 1。通过卷积层 conv1 得到 96 个大小为 75 × 75 的特征图。然后通过 ReLU 激活函数,再通过 pooling 层降采样,pooling 层的核尺寸大小为 3 × 3,步长为 2,得到输出为 96 × 37 × 37。

2)conv2 包含 256 个滤波器,核尺寸为 5 × 5,步长为 1,填充数为 2。通过卷积层 conv1 得到 256 个大小为 37 × 37 的特征图。然后通过 ReLU 激活函数,再通过 pooling 层降采样,pooling 层的核尺寸大小为 3 × 3,步长为 2,得到输出为 256 × 18 × 18。

3)conv3 包含 384 个滤波器,核尺寸为 3 × 3,步长为 1,填充数为 1。通过卷积层 conv1 得到 384 个大小为 18 × 18 的特征图。然后通过 ReLU 激活函数,再通过 pooling 层降采样,pooling 层的核尺寸大小为 2 × 2,步长为 2,得到输出为 384 × 9 × 9。

4)全连接层 fc4 将卷积层 conv3 得到的 384 × 9 × 9的输出特征进行全连接,神经元个数为 512。通过 dropout 层 drop4 控制训练时工作的神经元个数,以抑制过拟合。

5)全连接层 fc5 将全连接层 fc4 得到的 512 的输出进行全连接,神经元个数为 1。

最后,loss 层将 fc5 得到的结果通过 Sigmoid 函数进行概率计算,得到预测标记,与真实标记相比计算损失,并对网络进行优化。

1.4　算法验证

本实验所使用的数据集为 PETA dataset[15]。PETA 中的所有图像都在当前流行的行人再识别数据库中收集,PETA 数据集包含19 000张图片,分辨率最小为 17 × 39,最大为 169 × 365。19 000 张图片中共包含有8 705个行人,每个行人用 61 个二分类属性标签和 4 个多分类属性标签进行标。PETA 中的图像在背景、照明和视角上具有很大的差异。PETA 中的一些图像已经在图 2 中显示。广泛采用的实验方案是将数据集随机分为 3 个部分:训练集包含9 500张图像,验证集包含1 900张图像,测试集包含7 600张图像。

如果一张图片中出现了某个属性,那么这张图片对于该属性为正样本,否则为负样本。例如,一张图片上的行人戴了帽子而没有戴眼镜,则对于帽子属性,该图片为正样本,对于眼镜属性,该图片为负样本。本研究从 65 类属性标签中选取 10 类属性标签进行实验。

训练网络时,采用随机梯度下降法(SGD)优化网络,初始学习率为 0.001,参数 weight decay 设置为 0.005,batch 大小为 100,训练20 000个 epoch,momentum 为 0.9,训练结束时的最小学习率

图 2 PETA 数据集中的行人图像

Fig. 2 The pedestrian images in PETA

为0.000 001。

为了验证本研究网络结构的性能，实验时使用了 PETA 数据集上经常使用的行人属性分类方法 ikSVM[13] 做对比。训练时，通过增加正样本数目或负样本数目使正负样本比例平衡来为每个属性训练一个 ikSVM 分类器。

2 结果

本研究所用的深度卷积网络模型所得到的实验结果是通过使用 Caffe[21] 深度学习框架获得的。Caffe 由伯克利 AI 研究所（BAIR）和社区贡献者开发，是一个以表达、速度和模块化为基础的深度学习框架。实验中，在对 10 类属性标签识别时，将多标签多类分类转化为对每个标签中单个类别的二分类任务。PETA 的基本评估标准是计算每个属性的平均识别精度。ikSVM 算法取得的结果是在 Matlab 上实验得到的。ikSVM 算法和本研究使用网络的实验结果如表 1 所示。

为了更好地分析实验结果，将表 1 的结果图像化显示，如图 3 所示，横轴表示属性（按正样本比例排序），纵轴表示识别精度。从实验中可以看出，使用传统手工设计特征算法 ikSVM 在行人属性多标签分类识别任务中，10 类属性标签平均识别精度达到 0.819。本研究所用的卷积网络结构在行人属性多标签分类识别任务中，识别结果有了明显的提高，10 类属性标签平均识别精度达到 0.852，在损失函数中加入正样本比例权重因子后，识别精度有了更进一步的提高，平均识别精度达到 0.892。

表 1 10 类属性标签识别精度

Tab. 1 The recognition accuracy of 10 classes of attributes

属性标签	比例	ikSVM	accu	accu_ratio
accessoryHat	0.102	0.928	0.976	0.983
accessoryNothing	0.749	0.796	0.872	0.906
accessorySunglasses	0.029	0.946	0.950	0.971
carryingBackpack	0.197	0.759	0.799	0.837
lowerBodyCasual	0.861	0.855	0.893	0.921
lowerBodyJeans	0.306	0.787	0.752	0.820
personalLess30	0.497	0.769	0.759	0.829
personalLess45	0.329	0.762	0.791	0.849
personalLess60	0.102	0.811	0.947	0.968
personalMale	0.549	0.779	0.787	0.840
mA	*	0.819	0.852	0.892

注：表中 mA 代表平均精度，第 2 列为属性正样本在数据集中所占比例，第 3 列为 ikSVM 算法取得的精度，第 4 列和第 5 列分别表示在训练时参数 w_l 均取 0.1 和按式（3）取值时获得的精度。

Note: mA represents mean accuracy, the 2nd column shows the ratio of positive samples, the 3rd column shows the accuracy obtained by the ikSVM algorithm when all w_l assigned with 0.1, the 4th column shows accuracy when w_l assigned with values according to formula (3). And the 4th and 5th columns indicate that the accuracy obtained when the parameter w_l is taken as 0.1 or assigned by formula (3), respectively.

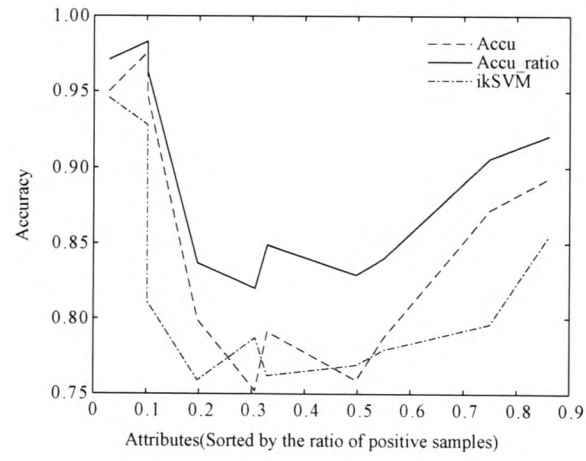

图 3 实验结果的图形化

Fig. 3 The image of experimental results

3 讨论

本研究分别使用深度卷积神经网络模型和 ikSVM[13] 算法对行人属性进行了分类识别。ikSVM 是一种基于 SVM 和传统手工设计特征的方法。用来训练 ikSVM 分类器的特征向量具有2 784个维度，包括 8 个颜色通道（RGB、HSV 和 YCbCr），以及在亮度通道上使用 Gabor 滤波器和 Schmid 滤波器获得的 21 个纹理通道。深度卷积网络模型直接从原始图像像素中提取抽象特征，通过逐层卷积，在高

层提取出表征能力很强的特征，并将高层特征用于分类。

从识别精度上可以看出，与传统手工设计特征训练的分类器 ikSVM 相比，本研究模型在大部分属性上取得了更高的识别精度。但在某些属性上 ikSVM 算法识别精度超过了本研究没有改进损失函数的网络，这是由于两者的学习机制不同所导致的，ikSVM 是针对每个属性训练的单独分类器，而本研究模型是对 10 类属性联合训练的分类器。同时可以看出，训练过程中在损失函数中增加正样本比例指数因子 w_l，以增大正样本比例较少属性的损失值，从而增大对于样本比例失衡的惩罚，使网络模型的属性识别精度有了明显的提高，并且都超过了 ikSVM，在应用中可以考虑将这一因子作为提高网络性能的一项重要因素。

由图 3 可见，由于数据集中的某些属性（如 accessoryHat、accessorySunglasses、personalLess60）正负样本比例严重不平衡，会对识别精度造成影响，比如对于眼镜这一属性，正样本比例只有 0.029，一个没有经过学习直接把样本划分为负类的分类器就能获得 0.971 的准确度。所以，造成了在属性正负样本比相对均衡处，识别精度比属性正负样本比例失衡处明显降低，降低范围在 0.163 ~ 0.224 之间，这说明加入 w_l 来平衡属性样本比例失衡所造成影响的能力有限，后续研究可以考虑同时加入正负样本比例指数因子。在正负样本分布相对比较平衡的属性上（如 personalMale、personalLess30、personalLess45），改进损失函数后网络模型仍取得了更高的识别精度，这说明改进后深度卷积神经网络有更强的特征提取和表征能力，能明显提高行人属性分类精度。

4　结论

本研究主要探讨了深度学习的方法在行人属性多标签分类识别中的应用。使用了一个拥有 3 个卷积层和 2 个全连接层的深度卷积神经网络模型，训练所用的数据集为标注好的 PETA 行人属性数据集，通过在 caffe 深度学习框架中进行学习训练，与传统手工设计特征训练的分类器相比取得了更好的分类识别精度。为了平衡属性样本比例失衡对模型性能造成的影响，在损失函数中增加了正样本比例指数因子，使网络模型的性能有了明显的提高。未来如果继续增加网络的深度，扩充训练数据集，优化损失函数，预计网络的性能会有进一步的提升。

参考文献

[1] Jaha ES, Nixon MS. Analysing soft clothing biometrics for retrieval [C]// Biometric Authentication. Cham: Springer International Publishing, 2014: 234-245.

[2] Dantcheva A, Singh A, Elia P, et al. Search pruning in video surveillance systems: Efficiency-reliability tradeoff [C]// 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). Barcelona: IEEE, 2011: 1356-1363.

[3] Reid DA, Nixon MS, Stevenage SV. Soft biometrics; human identification using comparative descriptions [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 36 (6): 1216-1228.

[4] An Le, Chen Xiaojing, Kafai M, et al. Improving person re-identification by soft biometrics based reranking [C]// 2013 Seventh International Conference on Distributed Smart Cameras (ICDSC). Palm Springs: IEEE, 2013: 1-6.

[5] Dantcheva A, Dugelay JL, Elia P. Person recognition using a bag of facial soft biometrics (BoFSB) [C]//2010 IEEE International Workshop on Multimedia Signal Processing (MMSP). Saint Malo: IEEE, 2010: 511-516.

[6] Martinson E, Lawson E, Trafton G. Identifying people with soft-biometrics at fleet week[C]//Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction. Tokyo: IEEE, 2013: 49-56.

[7] Kumar N, Berg AC, Belhumeur PN, et al. Attribute and simile classifiers for face verification [C]//2009 IEEE 12th International Conference on Computer Vision. Kyoto: IEEE, 2009: 365-372.

[8] Farenzena M, Bazzani L, Perina A, et al. Person re-identification by symmetry-driven accumulation of local features [C]//2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco: IEEE, 2010: 2360-2367.

[9] Ferrari V, Zisserman A. Learning visual attributes [C]// Advances in Neural Information Processing Systems. Vancouver: ACM, 2007: 433-440.

[10] Siddiquie B, Feris RS, Davis LS. Image ranking and retrievalbased on multi-attribute queries [C]//2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Colorado Springs: IEEE, 2011: 801-808.

[11] Kumar N, Berg AC, Belhumeur PN, et al. Describable visual attributes for face verification and image search [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(10): 1962-1977.

[12] Layne R, Hospedales TM, Gong S, et al. Person reidentification by attributes[J]. BMVC, 2012, 2(3): 8-17.

[13] Layne R, Hospedales TM, Gong S. Attributes-based re-identification[C]// Person Re-Identification. London: Springer, 2014: 93-117.

[14] Zhu Jianqing, Liao Shengcai, Lei Zhen, et al. Pedestrian

attribute classification in surveillance: Database and evaluation [C]//2013 IEEE International Conference on Computer Vision Workshops (ICCVW). Sydney: IEEE, 2013: 331-338.

[15] Deng Yubin, Luo Ping, Loy CC, et al. Pedestrian attribute recognition at far distance [C]//Proceedings of the 22nd ACM International Conference on Multimedia. Orlando: ACM, 2014: 789-792.

[16] Li Dangwei, Chen Xiaotang, Huang Kaiqi. Multi-attribute Learning for Pedestrian Attribute Recognition in Surveillance Scenarios [C]//2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR). Kuala Lumpur: IEEE, 2015: 111-115.

[17] Zhu Jianqing, Liao Shengcai, Lei Zhen, et al. Multi-label convolutional neural network based pedestrian attribute classification [J]. Image and Vision Computing, 2017, 58: 224-229.

[18] Fukui H, Yamashita T, Yamauchi Y, et al. Robust pedestrian attribute recognition for an unbalanced dataset using mini-batch training with rarity rate [C]// 2016 IEEE Intelligent Vehicles Symposium (IV). Gothenburg: IEEE, 2016: 322-327.

[19] Levi G, Hassner T. Age and gender classification using convolutional neural networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Boston: IEEE, 2015: 34-42.

[20] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012, 25(2): 1097-1105.

[21] Jia Yangqing, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding [C]// Proceedings of the 22nd ACM International Conference on Multimedia. Orlando: ACM, 2014: 675-678.