监控场景中基于图像描述的人物检索

李亚栋[1]，刘永超[1]，庚琦川[1]，周　忠[1]，吴　威[1]
（[1]北京航空航天大学 虚拟现实技术与系统国家重点实验室，北京 100191）

摘　要：监控场景中特定人物的检索是安防领域重要且迫切的需求。近年来，图像检索领域的方法主要基于图像内容，但是由于该类方法需要待检索图像作为输入，无法满足监控安防的实际需求。因此本文提出一种基于图像描述的人物检索方法，并提供一个标注了人物描述的监控数据集 SPCD。在此数据集上验证本文方法，性别预测的准确率达到 86.5%，服饰颜色匹配的准确率达到 93.5%，行为分类的准确率达到 65.5%，本文为监控场景中的人物检索提供了一种新的有效方式。

关键词：监控场景；图像检索；图像描述；多属性标签；监控数据集

文章编号：1004-731X (2002) xx-xxxx-x　　**中图分类号**：　　**文献标识码**：

Surveillance Scene Person Retrieval based on Image Captioning

Li Ya-dong[1]，*Liu Yong-chao*[1]，*Geng Qi-chuan*[1]，*Zhou Zhong*[1]，*Wu Wei*[1]
([1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China)

Abstract：Retrieving specific person in the surveillance scene is an important and urgent demand on the security field. In recent years, the method of image retrieval is mainly based on image content, but this method can't meet the actual needs of surveillance and security because that requires the image to be retrieved as input. In this paper, we propose a method of person retrieval based on image captioning and present a new surveillance dataset SPCD which contains person caption labels. We verify this method in new dataset, the accuracy of gender, dress color and action respectively reaches 86.5%, 93.5% and 65.5%. This paper provides an effective way for retrieving person in the surveillance scene.

Keywords：surveillance scene; image retrieval; image caption; multi-attribute labels; surveillance dataset

0　引　言[1]

随着人们对安防系统的需求不断提高，监控摄像头在社会各种场所被大规模应用。其中查找走失人群，侦查嫌疑目标，预防偷窃斗殴等有碍社会公共秩序行为的发生是几类主要的需求，然而当前多数单位，检索特定人物依靠人工完成，投入大，耗时长，信息水平低下。因此，智能化监控（如图 1 所示）具有重要的现实意义。

人物检索的研究方向主要是基于人物图像内容，这种方法的前提是获得与被检索图像特征相同或相近的图像作为输入，但走失人群、危险人群的监控图像往往难以获得，在实际应用中通常只能依靠文字来检索目标人物。另一方面，当前判别人物性别、行为和服饰等属性的方法相对独立，缺少有效的方法同时获得人物的多种属性特征。现有方法并不能满足监控的实际需求。

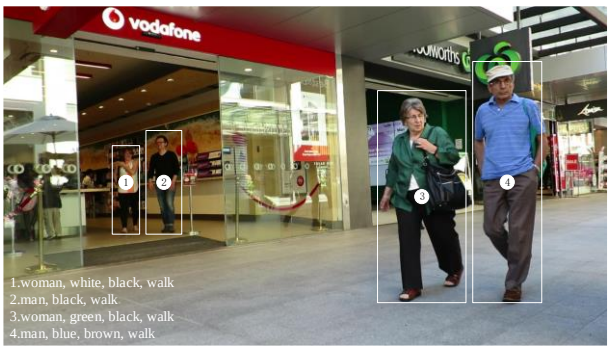

图 1 智能化监控的示例

本文提出了一种监控场景中基于图像描述的人物检索方法，实现检索监控视频中每帧图像的特定人物对象。图像描述问题是用一句话来描述图像的内容，完成了图像到文本的模态转换。而图像检索是用文本检索图像，两者属于互为逆向的问题，因此可以利用图像描述领域的技术自动生成人物图像的文字属性标签，通过关键字检索监控场景中的人物，将极大提高效率。

本文的主要贡献点有：

1、引入图像描述领域中的编码-解码框架到图像检索中，解决了图像标签需人工标注的难题。

2、生成监控场景中各人物的多属性标签，标签类别包

收稿日期：2107-08-01　　**修回日期**：2017-10-10
基金项目：国家自然科学基金项目(No.61572061、No. 61472020)，国家“八六三”高技术研究发展计划项目基金(No.2015AA016403)
作者简介：**李亚栋**(1992-)，男，山西省，汉族，硕士研究生，研究方向为计算机视觉；**刘永超**(1992-)，男，北京市，汉族，硕士研究生，研究方向为计算机视觉；**庚琦川**(1989-)，男，黑龙江省，汉族，博士研究生，研究方向为图像语义理解；**周　忠**(1978-)，男，江苏省，汉族，教授，博士，研究方向为虚拟现实等；**吴　威**(1961-)，男，黑龙江省，汉族，教授，博士，研究方向为网络与信息安全技术、分布式虚拟现实与可视化技术。

括性别、服饰颜色和行为，有助于提高检索目标人物的速度。

3、提出一个包含 1864 张，带有人物描述标注的监控图像数据集，用于监控场景中的人物检索。并在该数据集上验证了本文方法，其中性别预测的准确率达到 86.5%，服饰颜色粗匹配的准确率达到 93.5%。

2 相关工作

图像检索的相关研究可以分为三个阶段，初期主要是基于文本的图像检索方法，利用数据库存储图像，对每幅图像添加人工标注，但该方法数据量大，代价过高；之后是基于内容的图像检索方法[1-5]，主要思想是比较图像特征间的相似性，初期主要提取图像的颜色、纹理和形状等低级特征，不包含语义信息，因此准确率不高。现阶段该问题的研究主要致力于减小低级视觉特征和高级语义特征间的空白区间[6,7]，然而这类方法受限于要先获得待检索的人物图像，所以并不可行。

图像描述主要有直接生成法和视觉空间检索法。直接生成法运用语义信息完成图像到描述映射[8]，这种方法严重依赖于对视觉内容评估的正确性和语句描述视觉内容的准确性。而视觉空间检索法是将图像描述问题等价为检索相似图像的问题[9]，这种方法总能产生符合语法规则的描述，但主要问题在于需要大规模的包含人工标注描述的图像数据集。

随着编码-解码框架在机器翻译领域的成功应用，该框架逐渐成为图像描述的主流方法，Google 在 2015 年提出的 Show and Tell 模型将编码阶段的递归神经网络（Recurrent neural network, RNN）换成卷积神经网络（Convolutional neural network, CNN），提取图像视觉特征，在解码阶段将 RNN 改为性能更好的长短期记忆网络（Long Short-Term Memory, LSTM）[10]。Kelvin Xu 等人在上述框架中加入注意力机制，每次生成的单词过程运用不同的视觉特征[11]。后来武齐等人提取图像的高层语义作为解码器的输入，并将高层语义理解为一个多标签分类问题[12]。Mind's eye 模型改动了解码器的结构，不仅能将图像特征翻译为文字，还能反过来从文字得到图像特征[13]。

综上所述，图像描述的现有研究成功实现了图像到文本的转换，且准确性接近人类的描述语句，因此我们引用图像描述的方法，实现自动生成人物图像的属性标签，提高检索效率。

3 方法

本文的主要方法是在视频流中，按特定间隔抓取图像，然后检测并裁减出其中的人物区域，接着将得到的人物图像依次输入到编码（Inception）+解码（LSTM）结构中，生成图像的文字描述，进一步提取文字描述中的多属性标签，由此获得人物图像文本级别的属性，也就是生成了人物图像和属性的映射关系，最终将所有的图文配对信息存入到统一的数据库中，即可用文字检索监控场景的特定人物。框架如图 2 所示。

其中，特定间隔抓取图像目的在于保证视频内容完整性的前提下剔除冗余图像，间隔的长度由监控场景的性质决定，检测并裁减人物区域图像是由于多数监控图像的人数不定（可能包含 0 个、1 个或者多个人物），且噪声较多（周围环境复杂且多变），我们使用的检测算法是 Faster RCNN[14]+ Resnet101[15]模型，该模型是目标检测领域通用的模型，检测速度较快且准确性较高。

下面将重点介绍框架的其余两部分内容，3.1 介绍编码-解码结构用于生成图像描述，3.2 介绍多属性标签的设计。

3.1 图像描述

文字描述是已抽象出信息的结构化数据，比图像更有利于人类统计的方式，在图像检索中具有重要意义，从图像到描述的转换是人物检索的核心内容。

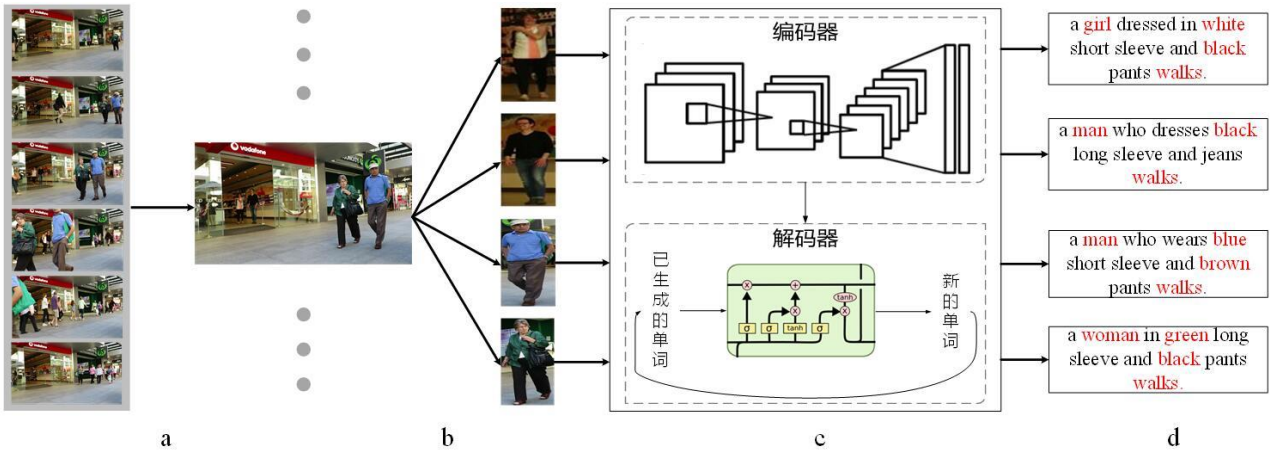

图 2 基于图像描述的人物检索的整体框架图。(a)从监控视频中截取视频帧图像；（b）检测并裁减出监控图像中的人物区域；（c）图像描述模型生成人物图像的文字描述，包含编码器和解码器两个模块；（d）提取描述语句中的多属性标签；

当前标注中含有图像描述标签的数据集有 MSCOCO[17] 和 Flickr30k 等，但所有数据集中的描述都是针对整幅图像的，极少包含图像中特定人物的详细属性信息，不满足局部人物检索的实际需求。

因此，在 show and tell 模型[10]的基础上，我们改进生成描述的句式，以配合后续提取属性关键字的步骤。具体结构如图 2（c）所示，编码器模块中，选择 Inception 网络[16]提取图像语义特征，Inception 子网络的并联结构可以有效保持图像的语义信息；解码器模块中，由图像语义特征和每一时刻之前生成的单词序列输入解码器，依次产生该时刻的新单词，如公式所示：

$$\log p(S \mid I) = \sum_{t=0}^{N} \log p(S_t \mid I, S_0, ..., S_{t-1}) \qquad (1)$$

其中 I 表示图像语义特征，$S = (S_0, ..., S_t, ..., S_n)$ 长度为 n 的描述语句，S_t 表示 t 时刻新生成的单词。通过在时间域上的迭代，最终生成图像的完整文字描述。解码器采用 LSTM 结构，较之前的递归网络，具备遗忘保存机制的 LSTM 拥有了学习能力，使生成的描述语法正确且贴近图像内容。

在训练阶段，我们采用如下公式

$$\theta^* = \arg\max \sum_{(I,S)} \log p(S \mid I; \theta) \qquad (2)$$

来最大化正确描述图像的概率，θ 表示模型中的所有变量参数。

3.2 多属性标签

在检索过程中，属性标签是人物图像的唯一索引，正确标签会有效缩小人物的搜索范围、加快检索速度。为了确定正确的属性标签，我们以调查问卷的方式展开调研，共收集到 50 份问卷，根据问卷结果选取了三类关键字信息，分别是性别、服饰颜色和行为，部分人物示例如图 3。

图 3 监控场景的人物示例

性别是区别行人基本的属性，检索中加入性别属性可以筛掉不符要求的人物。同时在监控场景中，通常得到的是人物的整体轮廓，而非面部特写信息，同时由于摄像头的分辨率差异较大，所以行人的服饰颜色在检索中占主要作用，正确的服饰颜色信息将精准聚焦在视频的个别人物上。另一方面，预判危险行为是重要的安防目标，因此我们也将行为加入到行人的关键字当中。

目前，这三种类型的图像属性有很多识别方法。但联合识别所有属性的研究非常少，我们设计了多属性标签的统一识别框架。图像描述阶段，得到保留图像语义信息且包含属性关键字的文字描述，从描述中查询即可得到属性关键字。

4 实验

下面首先介绍数据集，然后介绍本文方法在该数据集上的实验情况。

4.1 数据集

为了实现文字检索图像的目标，数据集中必须带有图像的描述信息，且描述信息对应人物区域，当前还没有满足上述要求的图像数据集，我们标定了一个新的数据集（Surveillance Person Caption Dataset, SPCD）。

SPCD 主要用于监控场景，来源于北航某些监控相机（50%）、国外街景（25%）和 MSCOCO（25%）数据集[17]等。筛选掉不包含人物或者人物区域较小的图像，最终数据集图像总量为 1010 张，每张图像中都包含 1 名或多名人物，数据集的人物总数达 1864 人。

数据集中的所有人物都含有文字描述标签，每条描述标签中包含人物的性别、服饰颜色和行为三种属性信息，每种属性的类别如下：

表 1 数据集中各属性的类别划分

属 性	类 别
性 别	男/女
服装颜色	黑/白/红/绿/蓝/黄/粉/紫/灰/棕
行 为	站/坐/走/弯腰/蹲/打电话

实验部分，随机选择 200 张人物图像作为测试集，其余 1664 张作为训练集。

4.2 实验细节

我们的实验平台为 Ubuntu16.04，显卡为 NVIDIA TITAN X(Pascal)，CPU 为 24 核 Intel Xeon E5-2643@ 3.4GHz，内存为 128G。

关于抓取视频帧的时间间隔，我们实验得到室内场景间隔 30 秒，室外场景间隔 10 秒。

编码-解码模型的在 Show and tell 模型 200 万次迭代的基础上进行训练。训练阶段，词表共有 12000 个单词，批处理的尺寸为 64，阻止率为 0.7，共计迭代 8 万次，优化策略选择随机梯度下降（Stochastic Gradient Descent，SGD）。模型推理阶段，对每幅测试图像生成 5 句描述用于查询关键字。

4.3 其他方法的比较

我们对比了基于文本的方法和方向梯度直方图（Histogram of Oriented Gradient, HOG）—颜色直方图

（Color Histogram，CH）的方法，其中基于文本的方法以人工标记为主，我们的方法则可以自动生成图像的多属性标签用于检索，较传统方法优势明显。而 HOG-CH 的方法无法解决监控场景难以获得待检索人物图像的情况。另一方面，HOG-CH 特征经过主成分分析（Principal components analysis，PCA）处理之后，特征点对应的图像的行为分布情况如图 4 所示，可以看出 6 种行为不能直接根据上述特征分类，同时颜色属性的判别属于多分类任务（每个人可以有多个颜色属性值），更增加了分类的难度，由此看出基于图像内容的方法无法适应多属性标签的分类任务。

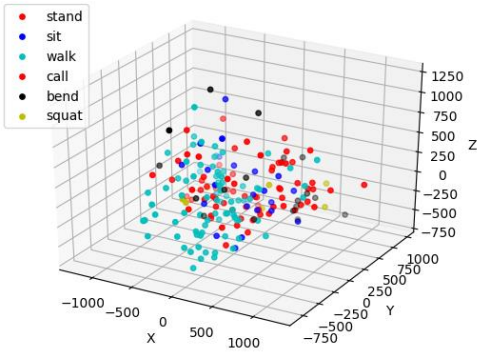

图 4 SPCD 测试集的行为属性分布

我们的方法在 SPCD 测试集上的表现如图 5 所示，横轴表示训练的迭代次数（单位：万次），纵轴为属性识别的准确率。

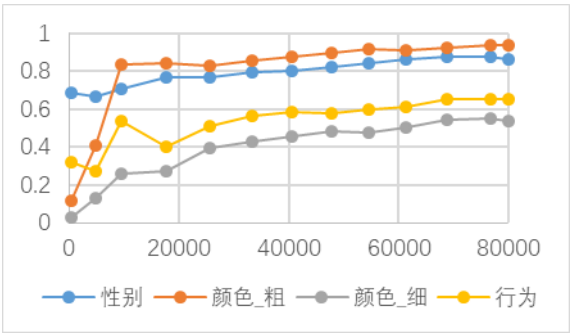

图 5 属性识别准确率与迭代次数的关系

计算过程中，单人的性别判定准则如下：

$$A_{i\_sex}=\begin{cases}1 & Num(sex_i)>5/2\\ 0 & Num(sex_i)\le 5/2\end{cases} \tag{3}$$

其中 sex_i 表示第 i 个人的真实性别，$Num(sex_i)$ 表示 5 句描述中真实性别关键字出现的次数，如果超过一半，则认为第 i 个人的性别判断正确。性别判别的整体准确率 P_{sex} 的计算方法如下：

$$P_{sex}=\frac{\sum_{i=1}^{n}A_{i\_sex}}{n} \tag{4}$$

n 表示测试集的总数人，行为的判定方法与性别类似。

需要强调的是，每个人物的颜色属性可能有多种，因此颜色判定准则分粗匹配和精匹配两种。计算方式分别如下：

$$A_{i\_rough}=\begin{cases}0 & P_{i\_color}\cap T_{i\_color}=\varnothing\\ 1 & P_{i\_color}\cap T_{i\_color}\neq\varnothing\end{cases} \tag{4}$$

$$A_{i\_fine}=\begin{cases}0 & P_{i\_color}\cap T_{i\_color}=T_{i\_color}\\ 1 & P_{i\_color}\cap T_{i\_color}\neq T_{i\_color}\end{cases} \tag{5}$$

其中，P_{i_color} 为第 i 个人预测的颜色集合，T_{i_color} 为第 i 个人真实的颜色集合。

我们的方法最终在小规模数据集上取得了超越其他方法的结果，性别预测准确率达到 86.5%，服饰颜色的粗匹配和精匹配的准确率分别达到 93.5%和 55%，行为分类的准确率达到 65.5%，之后的实验将会进一步分析原因。

4.4 其他方法的比较

4.4.1 目标检测

我们随机选取 100 幅监控图像并计算出总人数，针对五种当前主流的网络模型，分别测试了人物类别的检测准确率和单幅图像的平均检测时间，结果如表 2 所示。

表 1 目标检测模型的时间和准确率测试

模　型	时间(s)	准确率
ssd_mobilenet_v1[18]	0.127	0.72
ssd_inception_v2[18,16]	0.1317	0.81
rfcn_resnet101[19,15]	0.284	0.86
faster_rcnn_resnet101[14,15]	0.3212	0.87
faster_rcnn_inception_resnet_v2[14,20]	1.1564	0.93

由表可以看出，检测准确率和检测时间大致成反比，权衡时间成本和准确率后，我们选用 faster_rcnn_resnet101 模型。

4.4.2 图像描述

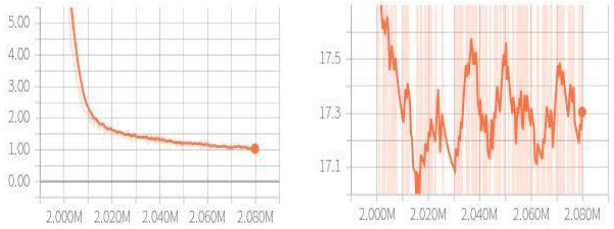

图 6 图像描述模型实验图

图 6 左表示模型损失（loss）随迭代次数的增加而不断下降，迭代 8 万次后趋于收敛；图 6 右表示描述语句每次迭代的最大长度，由图可知，句子最大长度在 17.3 附近小幅波动，这对训练的稳定有重要作用。

4.4.3 结果展示与分析

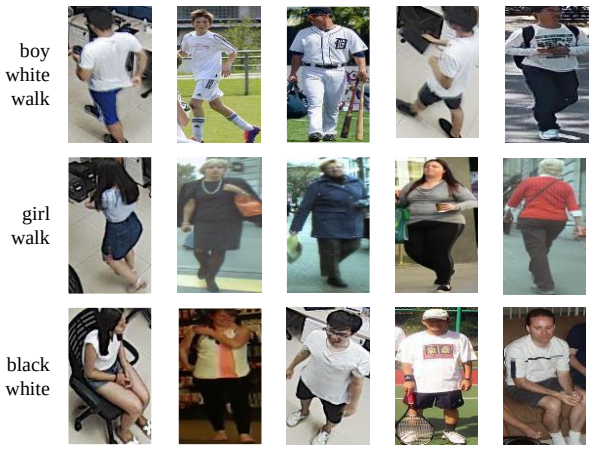

图 7 属性关键词检索的结果展示

图 7 展示了给定属性关键字进行检索时，数据库返回的结果。从图中可以看出返回结果在保证正确性的前提下，人物的丰富性也有较好表现，直接说明了我们方法的有效性。

图 8 属性识别失败的样例

如图 8 所示，在 SPCD 测试集中一些识别失败的情况，其中性别判断失败的主要原因是人脸及发型等性别的关键信息的缺失。颜色匹配失败的原因则是周围场景复杂和图案花纹丰富，数据集中的颜色类别有限。因为行为属性较另外两类属性的识别准确率较低，所以着重分析了这方面的原因，图 8 第二行从左至右，人物与场景相似度高、图像模糊、行为类别模糊，与动作类别高度相关的物体（如：手机）清晰度不够都会造成行为判别失败。

综合上述分析，想要提高人物检索的准确率，需要采集标记图像数量更多，分辨率更高的数据集，同时在标记时提供丰富的关键字信息。

5 总结

本文采用图像描述的编码-解码框架，自动生成针对监控场景中人物的文字描述，并从中提取属性关键字，提高了在图像检索问题中人物检索的效率。同时，本文标记了新的数据集 SPCD，并在此数据集上验证了上述方法的有效性，其中性别预测的准确率达到 86.5%，服饰颜色的粗匹配准确率达到 93.5%，目前实验结果受数据集规模和图像质量的影响，我们下一步工作将围绕优化模型和扩充数据集展开。

致 谢 本课题得到国家自然科学基金项目(No. 61572061、No. 61472020)、国家“八六三”高技术研究发展计划项目基金(No. 2015AA016403)资助。王世豪、马翰元和宁元星同学帮助标注了大量的数据，特此致谢。

参考文献:

[1] Jain A K, Vailaya A. Image retrieval using color and shape[J]. Pattern Recognition, 1996, 29(8):1233-1244.

[2] Swets D L, Weng J. Using Discriminant Eigenfeatures for Image Retrieval[M]. IEEE Computer Society, 1996.

[3] Datta R, Joshi D, Li J, et al. Image retrieval[J]. Acm Computing Surveys, 2008, 40(2):1-60.

[4] Kodituwakku S R, Selvarajah S. Comparison of Color Features for Image Retrieval[J]. Indian Journal of Computer Science & Engineering, 2010, 1(3):207211.

[5] Selvarajah S, Kodituwakku S R. Analysis and Comparison of Texture Features for Content Based Image Retrieval[J]. International Journal of Computers & Technology, 2011, 108(1):2045-5364.

[6] Rahman M M, Bhattacharya P, Desai B C. A Framework for Medical Image Retrieval Using Machine Learning and Statistical Similarity Matching Techniques With Relevance Feedback[J]. IEEE Transactions on Information Technology in Biomedicine, 2007, 11(1):58.

[7] Wang H H. Semantic Gap in CBIR: Automatic Objects Spatial Relationships Semantic Extraction and Representation[J]. International Journal of Image Processing, 2010, 4(3):192-204.

[8] ang H, Platt J C, Zitnick C L, et al. From captions to visual concepts and back[C]// Computer Vision and Pattern Recognition. IEEE, 2015:1473-1482.

[9] Ordonez V, Kulkarni G, Berg T L, et al. Im2Text: Describing Images Using 1 Million Captioned Photographs[J]. Advances in Neural Information Processing Systems, 2011:1143-1151.

[10] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[J]. 2014:3156-3164.

[11] Xu K, Ba J, Kiros R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[J]. Computer Science, 2015:2048-2057.

[12] Wu Q, Shen C, Liu L, et al. What Value Do Explicit High Level Concepts Have in Vision to Language Problems? [J]. 2016:203-212.

[13] Chen X, Zitnick C L. Mind's eye: A recurrent visual representation for image caption generation[C]// Computer Vision and Pattern Recognition. IEEE, 2015:2422-2431.

[14] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137.

[15] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]// Computer Vision and Pattern Recognition. IEEE,

2016:770-778.

[16] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[J]. Computer Science, 2015:2818-2826.

[17] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common Objects in Context[J]. 2014, 8693:740-755.

[18] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. 2015:21-37.

[19] Dai J, Li Y, He K, et al. R-FCN: Object Detection via Region-based Fully Convolutional Networks[J]. 2016.

[20] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[J]. 2016.