会议论文收录信息可视化与分析

——2021 春数据可视化课程报告

学号：20181001095　　班级：191181　　姓名：常文瀚

报告成绩：_____

功能实现（40）	目标网址符合要求（5）	
	对文本信息进行分词、词频统计等预处理（5）	
	对研究热点词汇、高产作者、论文数量、论文标题长度等信息进行统计分析（10）	
	对上述分析结果进行可视化展示和说明（20）	
报告格式（40）	报告是否按照规定模板撰写（5）	
	是否符合科技文献撰写的格式规范（5）	
	是否使用了折线图、饼状图、词云等多种可视化方式（10）	
	可视化（图、表的格式）是否简洁、高效、美观、规范（10）	
	篇幅是否饱满（5）	
	有核心源代码（5）	
总结和分析（20）	是否包含对可视化结果的分析和理解（10）	
	上述分析和理解是否正确、深刻（10）	

任课教师：孙 琨

原创性声明：

本人声明报告者中的内容和程序为本人独立完成，引用他人的文献、数据、图件、资料均已明确标注出。除标注内容外，不包含任何形式的他人成果，无侵权和抄袭行为，并愿意承担由此而产生后果。

作者签字：常文瀚　　时间：2021.05.21

要求及评分标准

1. 功能（40 分）

主要考查：是否按要求实现了主要功能，即：（1）对网址 https://openaccess.thecvf.com/CVPR2020 中所列文献信息进行分析；（2）对文本进行预处理，包括分词、词频统计等；（3）对研究热点词汇、高产作者、论文数量、论文标题长度等信息进行统计分析；（4）对上述分析结果进行可视化展示和说明；（5）可选项，对不同年份上的数据进行分析，揭示其变化和关联规律。

2. 报告格式（40 分）

主要考查：（1）报告是否按照规定模板撰写；（2）是否符合科技文献撰写的格式规范；（3）是否使用了折线图、饼状图、词云等多种可视化方式；（4）可视化（图、表的格式）是否简洁、高效、美观、规范；（5）篇幅是否饱满；（6）核心源代码。

3. 总结和分析（20 分）

主要考查：（1）是否包含对可视化结果的分析和理解；（2）上述分析和理解是否正确、深刻。

中国地质大学

题　　目：	会议论文收录信息可视化与分析
姓　　名：	常文瀚
院　　系：	计算机学院
班　　级：	191181
学　　号：	20181001095

2021 年 5 月 8 日

目录

CVPR 会议论文收录信息可视化与分析

常文瀚

（中国地质大学，武汉，2021）

摘要 数据可视化素养是数据经济时代公民应具备的基本素养之一,是大数据情境下素养研究和教育的新方向,是时代赋予计算机学科的新命题。以当前我们对人脑思维和认知的了解,抽象思维与形象思维是我们联系一切事物的根本。在计算机技术日益成熟的今天,大数据可视化是当代最为热门的话题之一,随着大数据时代的到来,大数据可视化技术与可视分析已逐渐成为科学发现与创新的重要方式。本文利用数据可视化技术，对近三年来 CVPR 会议的研究热点词汇、高产作者、论文数量进行了统计分析、可视化展示与说明。

关键词：计算机技术，大数据，可视化

CVPR 是首屈一指的年度计算机视觉活动，包括主要会议和几个共同举办的研讨会和短期课程。它以其高质量和低成本，为学生，学者和行业研究人员提供了一个特殊的平台。

在各种学术会议统计中，CVPR 被认为有着很强的影响力和很高的排名。目前在中国计算机学会推荐国际学术会议的排名中，CVPR 为人工智能领域的 A 类会议。在巴西教育部的排名中排名为。基于微软学术搜索(Microsoft Academic Search)2014 年的统计，CVPR 中的论文总共被引用了 169,936 次。

目前，正式发表的学术论文题目中一般都含有关键词。这些关键词既反映研究成果的核心内容，又提供重要的检索途径。一个学术研究领域较长时域内的大量学术研究成果的关键词集合，可以揭示研究成果的总体内容特征、研究内容之间的内在联系、学术研究的发展脉络与发展方向等许多重要课题。关键词分析已经成为文献计量学研究的一种行之有效的方法。本文将利用数据可视化技术，对近三年来 CVPR 会议的研究热点词汇、高产作者、论文数量进行了统计分析、可视化展示与说明。

一、样本介绍

2018年CVPR会议共收到3309篇文章，其中979篇被录用。投录比约为29.5%。收录论文按专家评分，分为三个层次：Poster、Spotlight、Oral。Spotlight(亮点论文)一共有 224 篇，占收录论文(224/979)的 22.88%。Oral(演示论文)一共有 70 篇，占收录论文(70/979)的 7.1%。

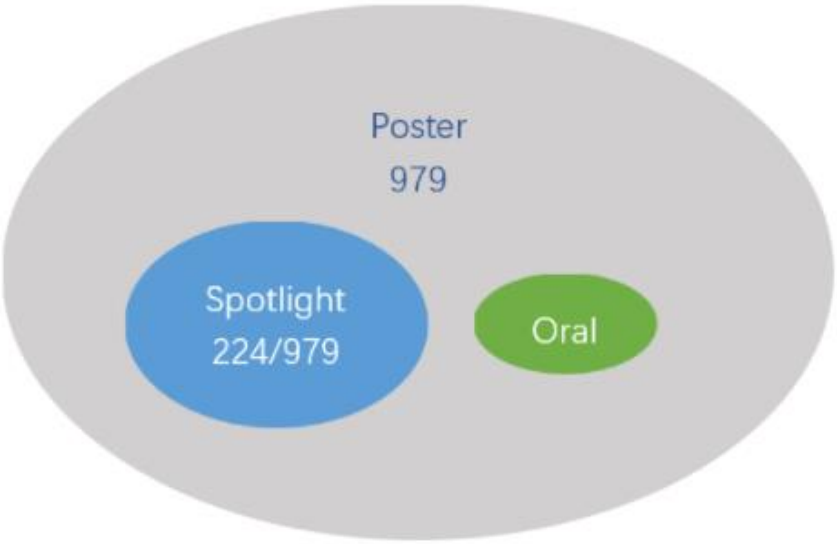


图 1 2018 年 CVPR 收录文章等级占比

CVPR 2019 于美国洛杉矶举办，接收结果公布后，又引起了 CV 届的一个小高潮，根据 CVPR 官网论文列表统计的数据，2019 年度共有 1300 篇论文被接收，而这个数据在 2019 年的过去 3 年分别为 643 篇（2016）、783 篇（2017）、979 篇（2018）。这从一个方面也说明了计算机视觉这个领域的方兴未艾，计算机视觉作为机器认知世界的基础，也作为最主要的人工智能技术之一，正在受到越来越多的关注。

CVPR 2020 公布“开奖”结果，6656 篇有效投稿中最终有 1470 篇论文“中奖”，接收率为 22.1%左右。据悉，本届 CVPR 的评审阵容包括 198 位领域主席和 3664 位审稿人。

虽然在近三年来，CVPR 的论文投稿量都在持续大涨（CVPR 2018 有 3300 篇有效投稿、CVPR 2019 有 5160 篇有效投稿、CVPR 2020 有效投稿达 6656），然而在录用率方面，已是“二连降”（CVPR 2018 收录论文 979 篇、接收率为 29%

左右；CVPR 2019 收录论文 1300 篇，接收率为 25%左右；CVPR 2020 收录论文 1470 篇、接收率为 22%左右）。具体趋势可以参考下图（引自知乎）。

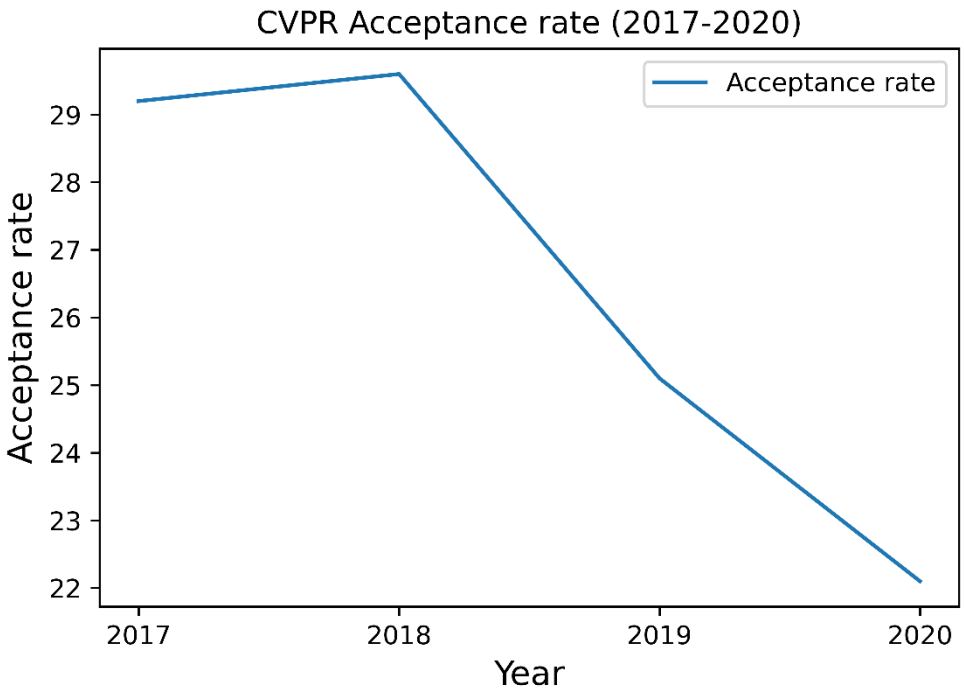


图 2 近 4 年 CVPR 会议文章录取率变化

二、数据分析

2.1 对 2018 年样本的分析

2018 年 CVPR 大会有超过 3300 篇的大会论文投稿，录取 979 篇（接受率约为 29%，其中包括 70 篇 Oral 和 224 篇 Spotlight 论文），注册参会人员也达到了 6512 位。

引用”指的是在论文中引述前人的研究成果，是作者表明其方法、观点和发现来源的标准方式。评价一篇论文的重要性，除了论文是否被顶级会议收录这一维度，论文的被引数也是不可或缺的维度。虽然引用量具体到不同学科的数据相差很多，但在计算机视觉这一单个学科内，论文的被引用量是评价某篇论文是否得到推崇的重要量化指标。

根据谷歌学术上的数据，我们统计出了 CVPR 2018 收录的 979 篇论文中被

引用量最多的前五名，希望能从引用量这个数据，了解到这些论文中，有哪些最为全球的学者们所关注。

	论文名称	引用量
1	Squeeze-and-Excitation Networks	554
2	Learning Transferable Architectures for Scalable Image Recognition	335
3	ShuffleNet:An Extremely Efficient Convolutional Neural Network for Mobile Devices	332
4	MobileNetV2:Inverted Residuals and Linear Bottlenecks	256
5	Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering	227

表 1 2018 年 CVPR 引用量最多的五篇文章

CVPR 2018 的高被引数论文都是获得学术界较大关注和推崇的论文，这主要在于他们的开创性。例如，排名第一的 Squeeze-and-Excitation Networks（简称 SE-Net）构造就非常简单，很容易被部署，不需要引入新的函数或者层，并且在模型和计算复杂度上具有良好的特性。

此外，还有 Google Brain 带来的 Learning Transferable Architectures for Scalable Image Recognition，提出了用一个神经网络来学习另一个神经网络的结构，也为许多学者所关注。

从引用量这项数据我们可以看出，全球的研究者在对图像处理与计算机视觉关注的同时也非常关注神经网络底层的结构，并有很大的从神经网络结构入手以改善计算机视觉算法应用效果的局势。

在对引用量分析的同时，我们也需要关注整体计算机视觉的研究趋势，这一点可以对文章题目切分后的关键词词频进行分析，同时会研究热点词汇、高产作者、论文数量进行了统计分析、可视化展示与说明。

在对 2018 年整体文章标题分析分析中，我们可以去除掉一些常见名词以及连接词，例如：“for”、“or”、“and”、“learning”，结果如下图：

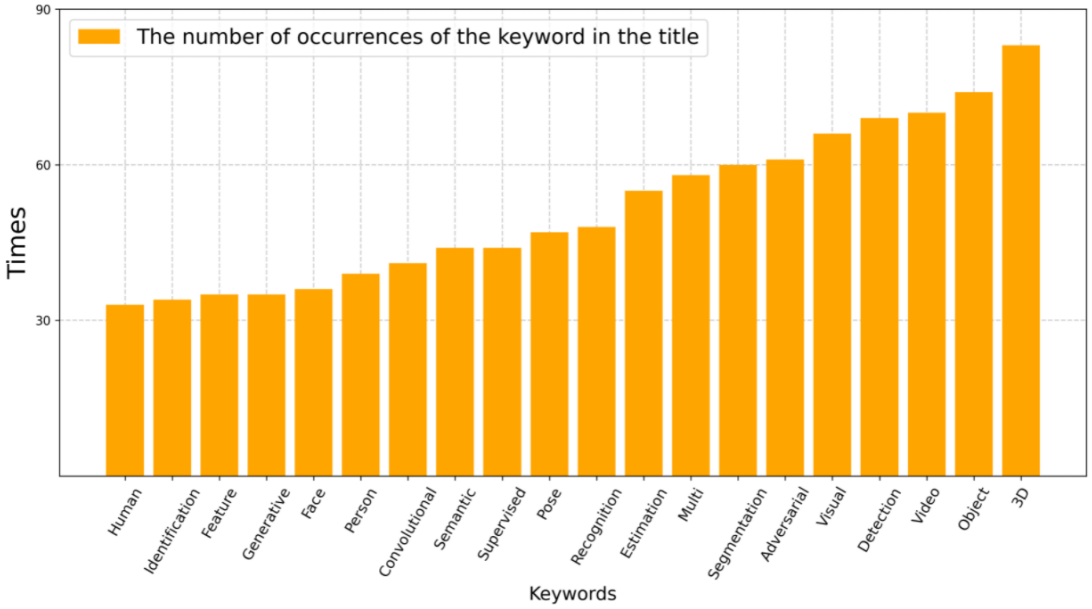


图 3 2018 年 CVPR 会议标题高频词汇

将图像中出现次数最多的十位转换为表格形式可以得到下表：

序号	关键词	词频	序号	关键词	词频
1	3D	83	6	Adversarial	61
2	Object	74	7	Segmentation	60
3	Video	70	8	Multi	58
4	Detection	69	9	Estimation	55
5	Visual	66	10	Recognition	48

表 2 2018 年 CVPR 会议文章标题中出现最多的十个关键词

通过图像与表格可以看到出现次数最多的前十个词汇中主要包含了：“OBject”、“Detecton”、“Estimation”等词，在 2018 年的计算机视觉研究者关注的主题主要包含了对对象的检测与图像的预测。

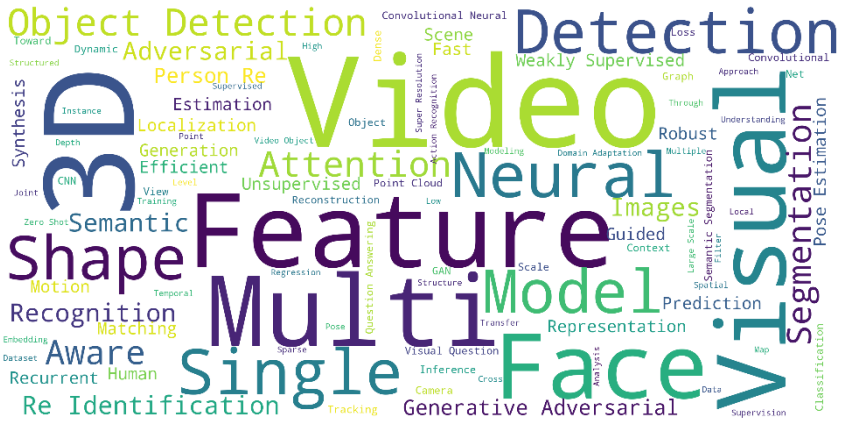
其中可以了解到 3D(第一位)成为了 2018 年计算机视觉界最为火热的主题，三维重建经过数十年的发展,已经取得巨大的成功，基于视觉的三维重建在计算机领域是一个重要的研究内容,主要通过使用相关仪器来获取物体的二维图像数据信息,然后,再对获取的数据信息进行分析处理，最后，利用三维重建的相关理

论重建出真实环境中物体表面的轮廓信息。基于视觉的三维重建具有速度快、实时性好等优点，能够广泛应用于人工智能、机器人、无人驾驶、虚拟显示和 3D 打印等领域，具有重要的研究价值，也是未来发展的重要研究方向。

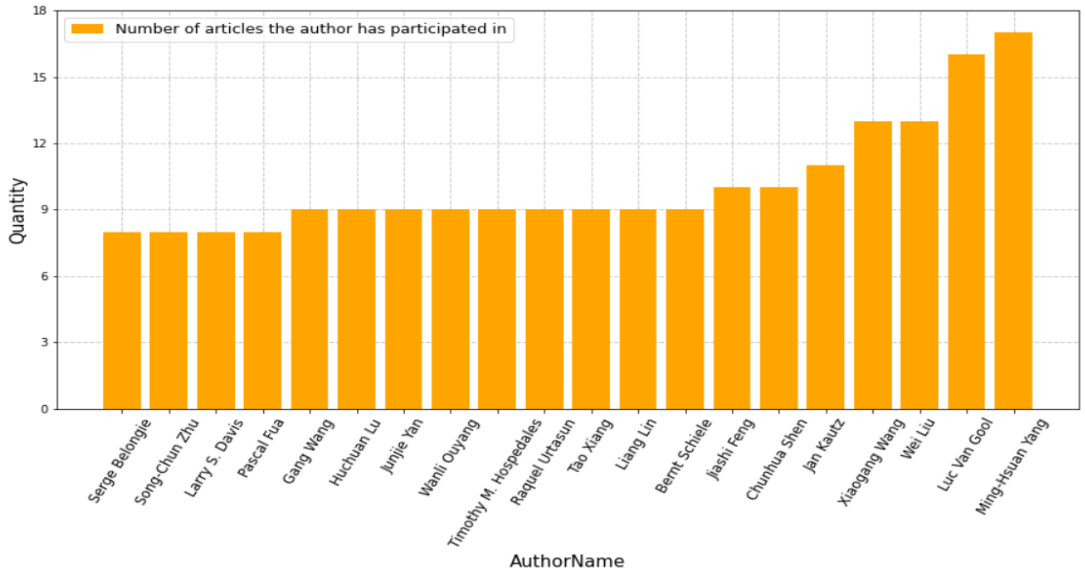
第三位高频词汇-“Video”，说明计算机视觉的研究领域有很大一部分已经从静态图像转移到了动态视频的处理，动态计算机视觉也可以被应用到很多实际的领域，如自动驾驶、机器人开发等。

在前十位中我们也会关注到排名第六位的词汇：Adversarial，与他相关的研究领域包含了 GAN 生成对抗网络、安全聚合、隐私保护等领域，这反映出即使在计算机视觉发展的同时，研究者们也丝毫没有忘记信息安全对于计算机发展的重要意义。

2018 年 CVPR 会议关键词词频分析词云见下图：



最后我们同样可以分析了参会作者的文章数量和以第一作者发表的文章数量，具体数据参考下图：



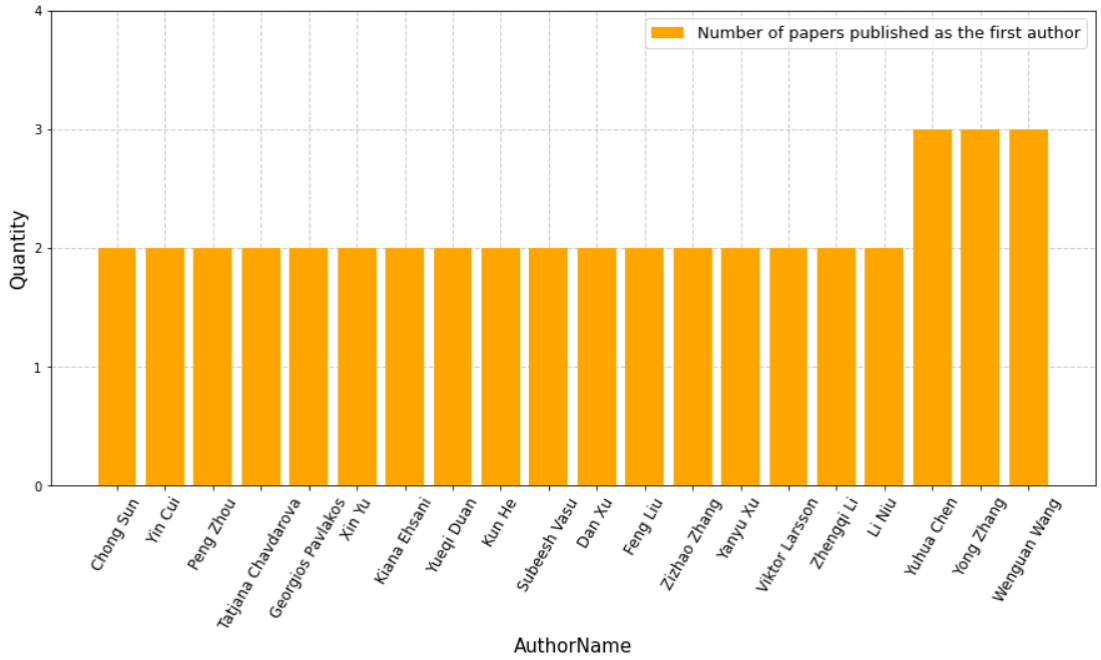


图 5 和图 6 发表文章与第一作者发表文章数量排序

2.2 对 2019 年样本的分析

CVPR2019 于 2019 年 6 月 16 日在美国召开，此次会议共收到来自全球 14104 位研究者提交的 5160 篇文章，同比 2018 年增长 56%，一举打破记录，受欢迎程度可见一斑。

盘点其中影响力最大的 20 篇论文，这里的影响力以谷歌学术上显示的论文的引用量排序，截止时间为 2020 年 7 月 22 日，可以得到这些结论：

1）这 20 篇论文全部开源了。不开源的论文复现代价大，别人参考的门槛会高很多，维护好论文对应的开源软件能极大提高论文影响力。

2）方向分布在 GAN、人脸识别、神经架构搜索、语义分割、图像合成、姿态估计、迁移学习、3D 目标检测、全景分割、目标跟踪、图像分类、网络结构设计（可变形卷积）、对抗学习、三维重建等方向。

3）这些论文绝大多数有工业界巨头的身影，英伟达贡献 2 篇 （第一名来自英伟达），谷歌贡献 5 篇，Facebook 贡献 8 篇， 国内商汤 1 篇，京东 1 篇。

4）目标检测是计算机视觉领域非常火的方向，但入选的两篇全是 3D 点云目标检测。

5）人脸识别在工业界应用很火，但只有一篇论文入前 20（大名鼎鼎的 ArcFace），说明这个领域的技术也许已经趋于成熟。

对所有接收的文章的题目关键词分析可以参考下图：

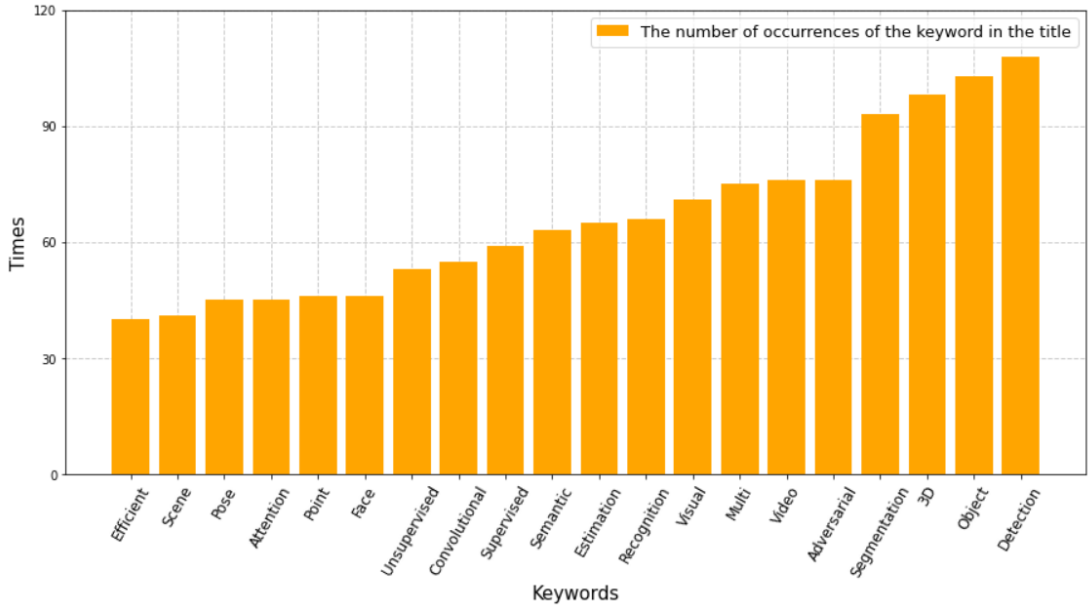


图 7 2019 年 CVPR 会议标题高频词汇

将图像中出现次数最多的十位转换为表格形式可以得到下表：

序号	关键词	词频	序号	关键词	词频
1	Detection	108	6	Video	76
2	Objection	103	7	Multi	75
3	3D	98	8	Visual	71
4	Segmentation	93	9	Recognition	66
5	Adversarial	76	10	Estimation	65

表 3 2019 年 CVPR 会议文章标题中出现最多的十个关键词

对比 2018 年和 2019 年出现最多的前十个关键词，出现最多的变为 Detection，这说明 2019 年计算机视觉领域最火热的研究方向方向是目标的检测，其余的高热度俺就放箱并没有特别大的变化。

在第十位到第二十位的排名中，对于人脸的研究仍然是重要的话题，注意到关键词“Scene”首次出现在了排名中，这也表明研究者们对场景分析的关注。

2019 年 CVPR 会议关键词词频分析词云见下图：

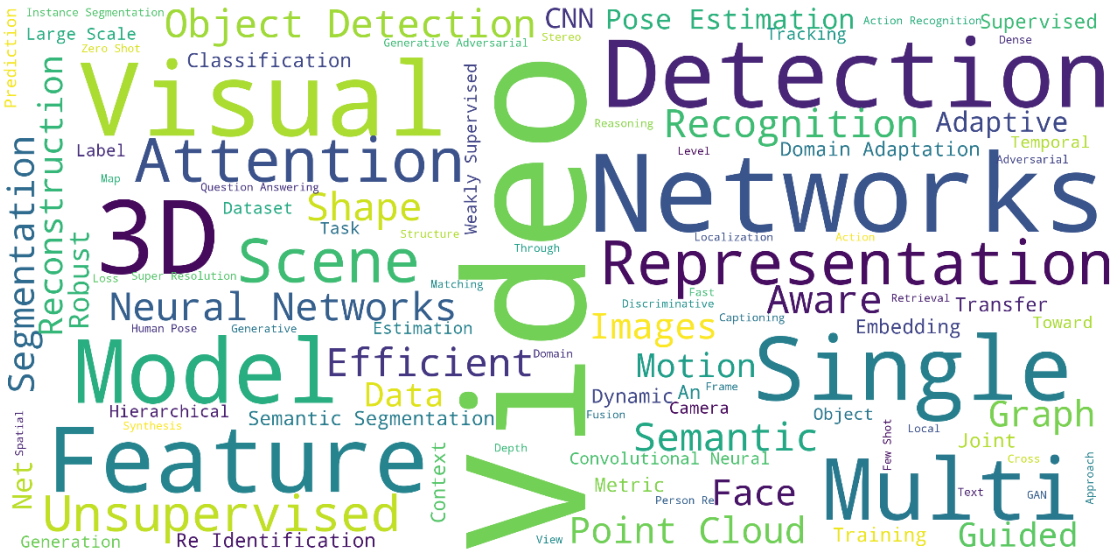


图 8 2019 年 CVPR 关键词词云

下图则是对高产作者的分析，两张图片分别标注了产出量最高的作者和以第一作者的身份发表文章量最多的的研究者：

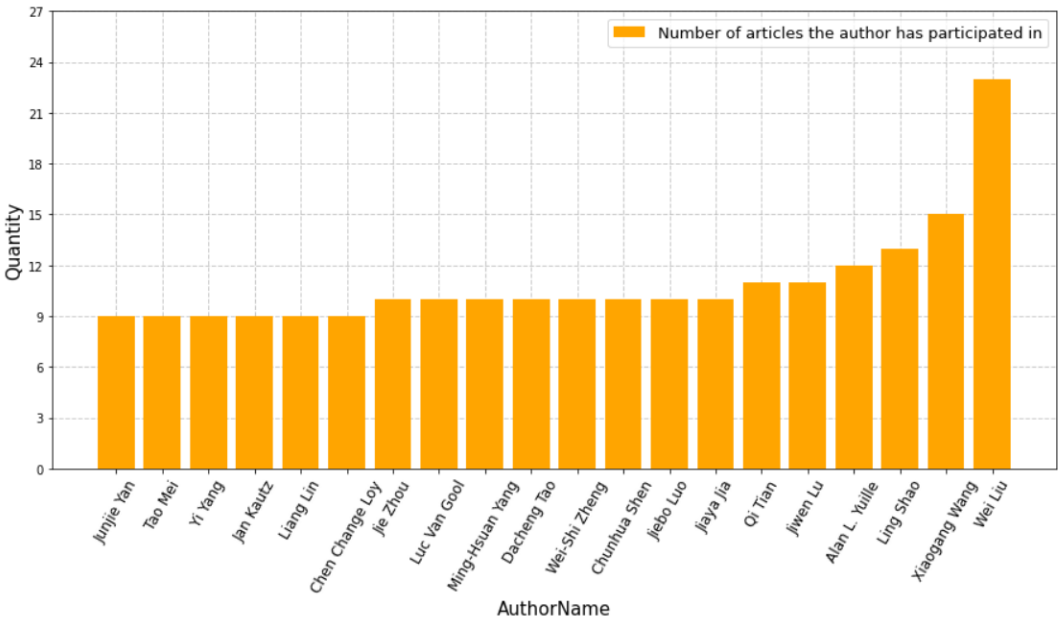


图 9 CVPR2019 年作者发表文章数

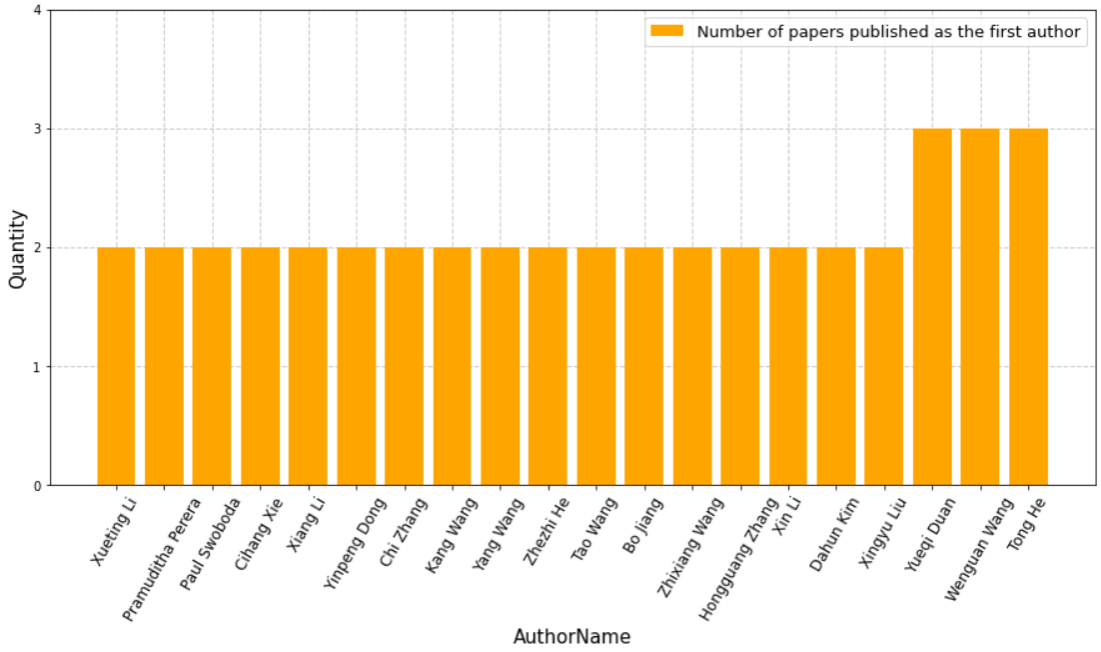


图 10 2019CVPR 第一作者发表文章数

2.3 对 2020 年样本的分析

CVPR 2020 公布接收论文结果，从 6656 篇有效投稿中录取了 1470 篇论文，录取率约为 22%。

对所有录取文章的题目关键词分析如下图：

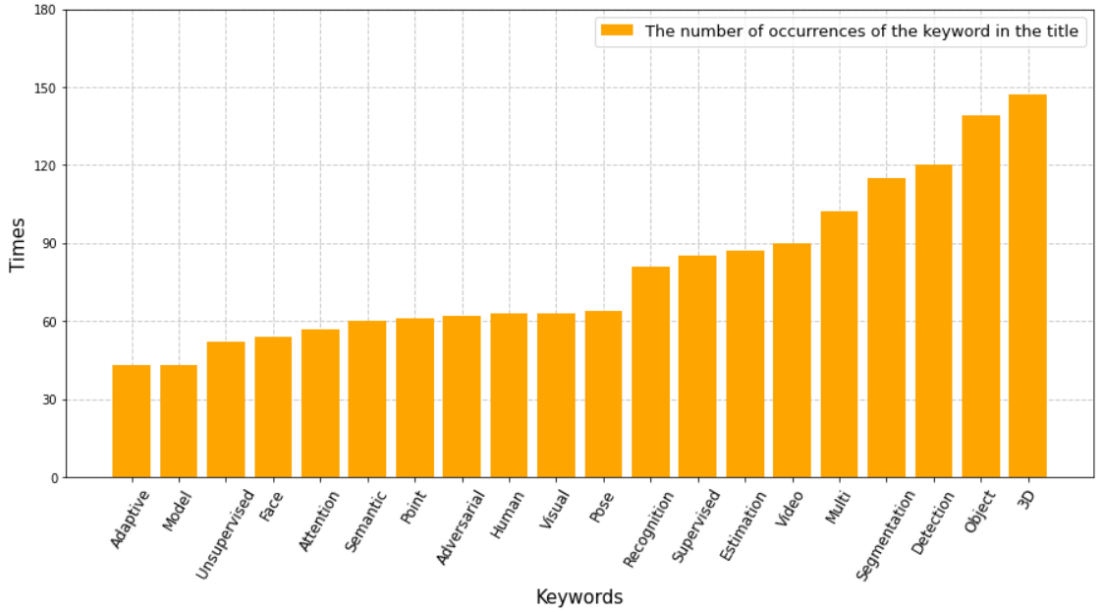


图 11 2020 年 CVPR 会议标题高频词汇

和前两年的排名相比，3D 建模和目标检测等领域依旧热度不减，2020 年度最为火热的话题包括了：目标检测、目标跟踪、图像分割、人脸识别、姿态估计、三维点云、视频分析、模型加速、GAN、OCR 等方向。

其中三维点云重建在视觉方面具有很高的研究价值，在机器人视觉导航、智能车环境感知系统以及虚拟现实中被广泛应用。三维重建经过数十年的发展，已经取得巨大的成功。基于视觉的三维重建在计算机领域是一个重要的研究内容，主要通过使用相关仪器来获取物体的二维图像数据信息，然后，再对获取的数据信息进行分析处理，最后，利用三维重建的相关理论重建出真实环境中物体表面的轮廓信息。

而目标检测是计算机视觉的基础,也是计算机视觉研究领域的挑战。当今时代,深度学习的思想已经逐渐融入人们的生活,为人们带来了诸多便利,基于深度学习的目标检测算法也成为目标检测研究领域的一项热门。

经常出现在排名前五位的“图像分割”是计算机视觉研究中的一个经典难题，已经成为图像理解领域关注的一个热点，所谓图像分割是指根据灰度、彩色、空间纹理、几何形状等特征把图像划分成若干个互不相交的区域，使得这些特征在同一区域内表现出一致性或相似性，而在不同区域间表现出明显的不同。简单的说就是在一副图像中，把目标从背景中分离出来。对于灰度图像来说，区域内部的像素一般具有灰度相似性，而在区域的边界上一般具有灰度不连续性。

下图是 2020 年 CVPR 的关键词词云：

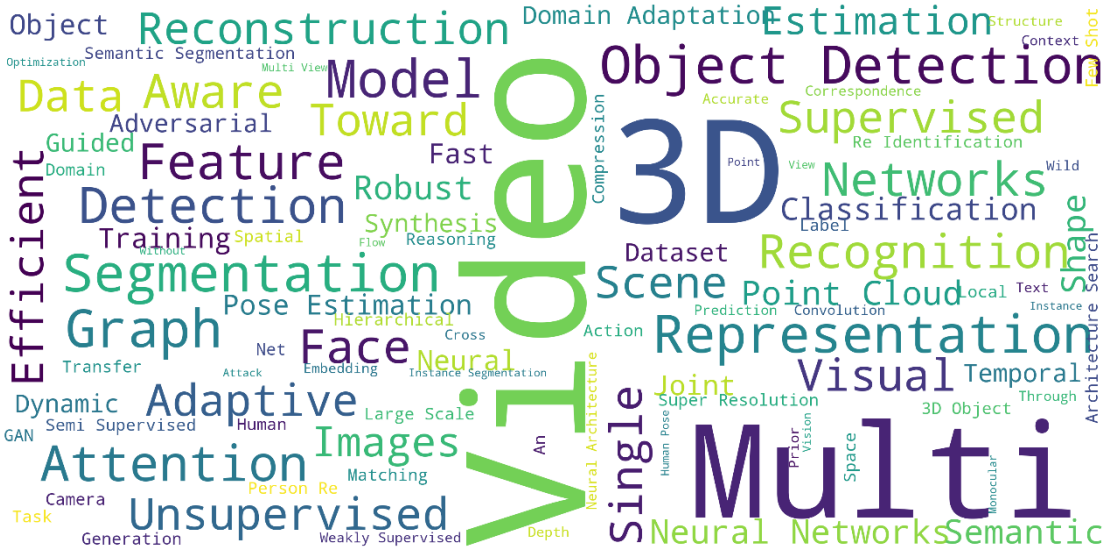


图 12 2020 年 CVPR 关键词词云

对 2018 年到 2020 年三年间的 CVPR 关键词热度进行分析后可以得到以下折线图，直接反映了不同领域的热度变化：

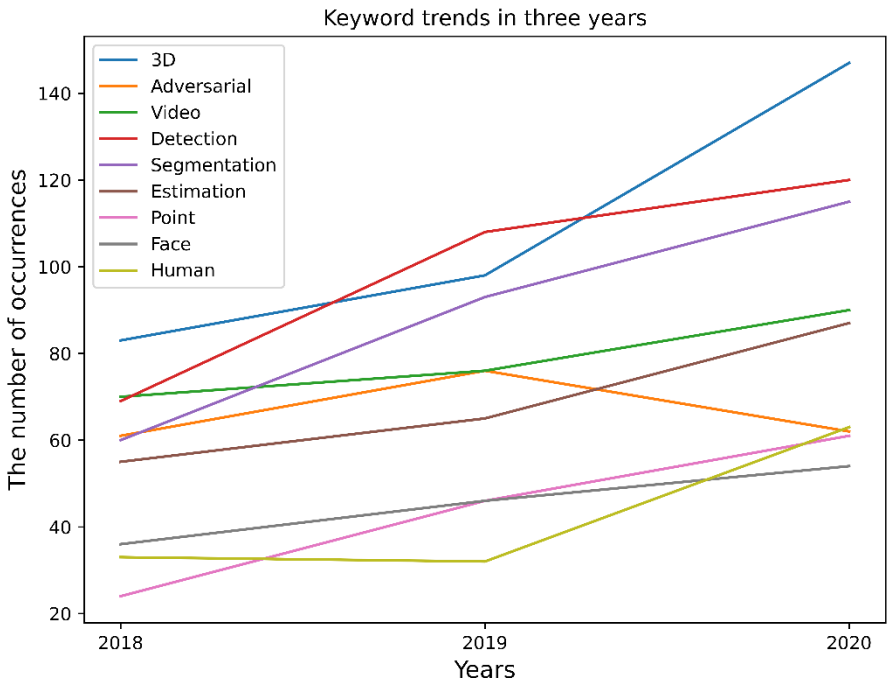


图 13 三年内 CVPR 不同领域的热度变化

对 2020 年 CVPR 的高产作者的分析如下图：

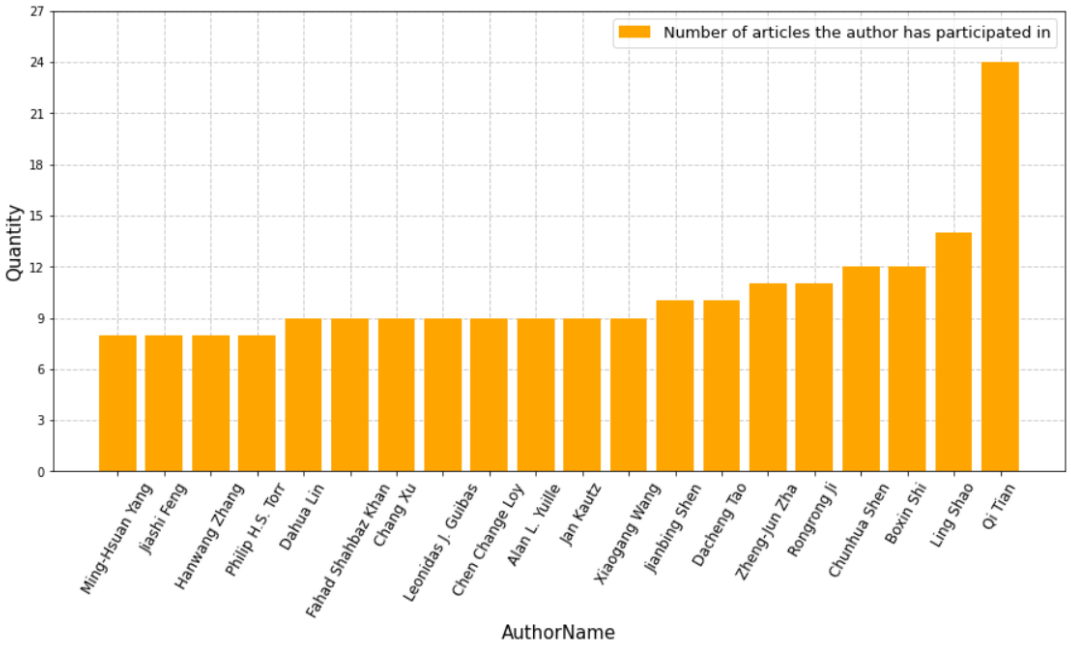


图 14 CVPR2020 作者发表文章数

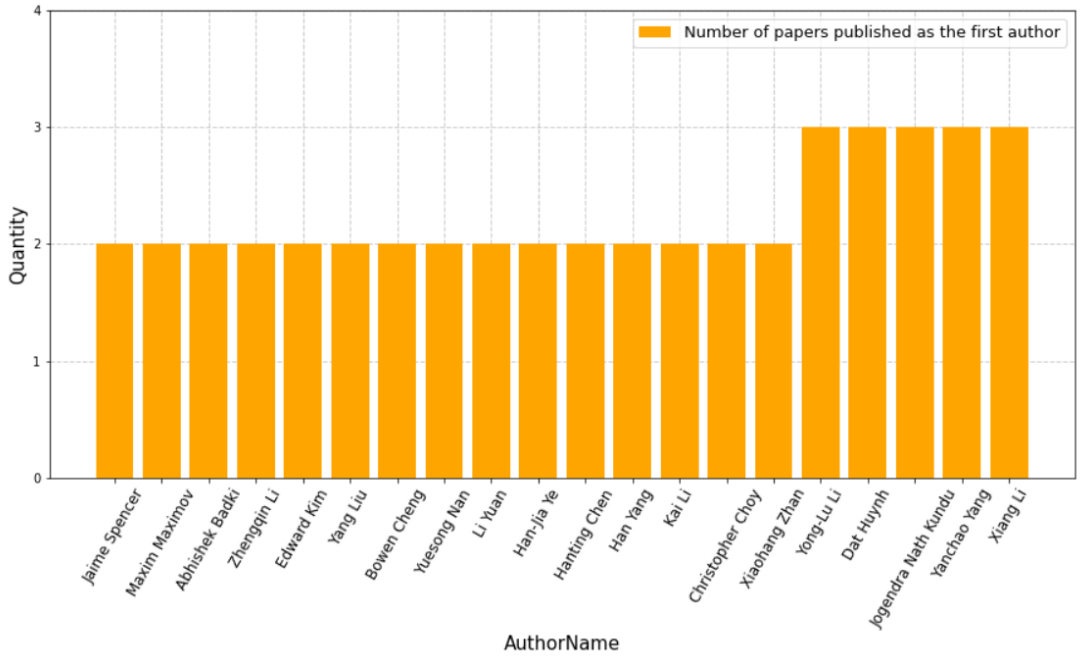


图 15 CVPR2020 第一作者发表文章数

三、 总结

3.1 总结

数据可视化是一个要求动手能力很强的一门实践课程，在课程设计期间我努力将自己以前所学的理论知识向实践方面转化，尽量做到理论与实践相结合，认真完成老师布置的任务 。

此次的结课作业，是对 CVPR 会议的文章标题关键词和高产作者进行分析 ，在设计过程中总是会遇到一些很小的问题，虽然不明显，却可以影响到整个程序的安全运作。这样一个小小的软件，却是通过一次次的搭建，修改之后的结果，真是令人感慨万千。

在设计与完善的过程中，第一个问题就是对数据可视化的操作工具不熟悉。虽然看过了一些参考视频，但是到了开始编写的时候还是会出现不熟练的问题，但是随着一次次的操作，我也开始越来越熟练了，虽然在这个过程中我出了很多的差错，但随着改进也会产生许多新的想法。

第一次的数据可视化并将结果进行分析虽然有所困难，却也使我更加深入的了解采用什么方法对数据进行可视化才能使他人更直观的接收到你想传达的信息。并将平时所学的知识第一次融会贯通。也明白了对数据可视化结果是需要花费很多精力去构思的，其间的财富是任何时候的上课实验所不可比拟的。

数据可视化是每一个计算机专业大学生在大学生涯中都不可或缺的，它使我们在实践中了巩固了所学的知识、在实践中锻炼自己的动手能力；它又是对每一位大学生所学专业知识的一种拓展手段，它让我们学到了很多在课堂上根本就学不到的知识,不仅开阔了自己的视野，增战了自己的见识，也为我们以后进一步走向社会打下了坚实的基础，是我们走向以后走向工作岗位的奠基石。

3.2 引用

[1] 李宝, 程志全, 党岗, 等. 三维点云法向量估计综述[J]. 计算机工程与应用, 2010, 46(23): 1-7.

[2] 王丽辉. 三维点云数据处理的技术研究[D]. 北京交通大学, 2011.

[3] 佟帅, 徐晓刚, 易成涛, 等. 基于视觉的三维重建技术综述[D]. , 2011.

[4] [1]陈科圻,朱志亮,邓小明,马翠霞,王宏安.多尺度目标检测的深度学习研究综述[J].软件学报,2021,32(04):1201-1227.

[5] [1]陆峰,刘华海,黄长缨,杨艳,谢禹,刘财喜.基于深度学习的目标检测技术综述[J].计算机系统应用,2021,30(03):1-13.

[6] [1]员娇娇,胡永利,孙艳丰,尹宝才.基于深度学习的小目标检测方法综述[J].北京工业大学学报,2021,47(03):293-302.

[7] [1]袁慧敏,张绪红.目标检测算法综述[J].科技经济导刊,2021,29(06):52-55.

[8] 罗希平, 田捷, 诸葛婴, 等. 图像分割方法综述[D]. , 1999.

[9] 王爱民, 沈兰荪. 图像分割研究综述[D]. , 2000.

[10] 林开颜, 吴军辉, 徐立鸿. 彩色图像分割方法综述[J]. 中国图象图形学报, 2005, 10(1): 1-10.

[11] https://www.zhihu.com/question/372070853

[12] https://zhuanlan.zhihu.com/p/112355481

附录 1 核心代码

```
index = 1


title=[]
author=[]


for i in data:
    if index % 2 != 0:
        title.append(i)
        index+=1
    elif index % 2 == 0:
        author.append(i)
        index+=1


print(len(title))
print(len(author))
# 导入 jieba 包
import jieba
#管理系统路径
import sys
import jieba.posseg as pseg


p = list(jieba.cut_for_search(title[0]))
print(p)
for i in p:
    if i == ' ':
        p.remove(i)
p.remove(p[len(p)-1])
```

```
print(p)


D=[]
for i in title:
    p = list(jieba.cut_for_search(i))
    p.remove(p[len(p)-1])   #删除转置符号

    for n in p:
        if n == ' ':          #删除空格
            p.remove(n)
    for n in p:
        if n == ':':          #删除冒号
            p.remove(n)

    for n in p:
        D.append(n)

print(len(D))

from collections import Counter
r = Counter(D)
print(r)
import pandas as pd
df = pd.DataFrame(r.items(), columns=['key', 'times'])
df.sort_values("times",inplace=True)

dropDict=['-','for','on','via','and','A','a','From',
          'from','of','to','in','With','with','the','Based','Domain','Graph','Aware','Single','Shot',
          'Self','Using','using',',','by','End','Learning','Deep','Neural','Network','Networks','Image']
```

```
for i in dropDict:
    df = df.drop(df[df['key']==i].index)

import matplotlib.pyplot as plt
import numpy as np

yticks=np.arange(0,210,30)

k = np.arange(len(x))  #首先用第一个的长度作为横坐标
width = 0.1      #设置柱与柱之间的宽度

fig,ax = plt.subplots(figsize=(15, 7))
ax.set_axisbelow(True)

ax.bar(k,y,label='The number of occurrences of the keyword in the title',alpha = 1,color= 'orange')

ax.set_xticks(k)#将坐标设置在指定位置
ax.set_xticklabels(x,size=12)#将横坐标替换成
ax.set_yticks(yticks)

ax.set_xlabel('Keywords',size=15)
ax.set_ylabel('Times',size=15)

plt.xticks(rotation=60)
plt.legend(fontsize=13)
plt.grid(linestyle = '--',linewidth =1, color= 'gray',alpha = 0.4)
# plt.savefig("The number of occurrences of the keyword in the title 2020.png",dpi=800)
plt.show()
```

```
author1=[]


authors=[]


for i in author:
    ls = i[:-1].split(", ")
#         print(len(ls))
#         print(ls)
    author1.append(ls[0])
    for k in ls:
        authors.append(k)
#print(author1)


t = Counter(author1)
countAuthors = Counter(authors)
#print(t)


au = pd.DataFrame(t.items(), columns=['key', 'times'])
at = pd.DataFrame(countAuthors.items(), columns=['key', 'times'])
au.sort_values("times",inplace=True)
at.sort_values("times",inplace=True)
at.tail(20)


yticks=np.arange(0,5)


k = np.arange(len(x))  #首先用第一个的长度作为横坐标
width = 0.1      #设置柱与柱之间的宽度


fig,ax = plt.subplots(figsize=(15, 7))
ax.set_axisbelow(True)
```

```
ax.bar(k,y1,label='Number of papers published as the first author',alpha = 1,color= 'orange')

ax.set_xticks(k)#将坐标设置在指定位置
ax.set_xticklabels(x1,size=12)#将横坐标替换成
ax.set_yticks(yticks)

ax.set_xlabel('AuthorName',size=15)
ax.set_ylabel('Quantity',size=15)

plt.xticks(rotation=60)
plt.legend(fontsize=13)
plt.grid(linestyle = '--',linewidth =1, color= 'gray',alpha = 0.4)
# plt.savefig("Number of papers published as the first author 2020.png",dpi=800)
plt.show()

np=df.key
string_paper = ' '.join(D)
import numpy as np
from PIL import Image
import re
from wordcloud import WordCloud,ImageColorGenerator,STOPWORDS
import    matplotlib.pyplot as plt

sw = {'-','for','on','via','and','A','a','From','Image','Deep','Network',
                'from','of','to','in','With','with','the','Based',
                'Self','Using','using',',','by','End','Learning'}


my_wordcloud = WordCloud(scale=8,stopwords=sw,background_color='white',max_words = 100,
```

```
                                    max_font_size = 60,random_state=20).generate(string_paper)
#显示生成的词云
plt.imshow(my_wordcloud)
plt.axis("off")
plt.show()

#保存生成的图片
my_wordcloud.to_file('Keywords wordcloud 2020.png')
```