第6届全国高校大数据与人工智能教学研讨会

2023.05.12–2023.05.13　中国·厦门

主办单位：教育部高等学校计算机类专业教学指导委员会

承办单位：厦门大学 XIAMEN UNIVERSITY　浙江大学 ZHEJIANG UNIVERSITY　中山大学 SUN YAT-SEN UNIVERSITY　中南大学 CENTRAL SOUTH UNIVERSITY　湖南大学 HUNAN UNIVERSITY　郑州大学 ZHENGZHOU UNIVERSITY　华侨大学 HUAQIAO UNIVERSITY　集美大学　人民邮电出版社 POSTS & TELECOM PRESS

协办单位：美林数据 MERITDATA　头歌 在线实践 educoder.net　普开数据 PUKAI DATA　山东省计算中心（国家超级计算济南中心）　荔峰科技 LIFENG　Tedu.cn 达内教育　高慧强学 HIGHSTRONG.CN

清華大學
Tsinghua University

第6届大数据与人工智能教学研讨会

大数据机器学习教学及科研体会

袁春
清华大学深圳国际研究生院
2023/05/13

清华大学深圳国际研究生院（Tsinghua Shenzhen International Graduate School，简称Tsinghua SIGS），是在国家深化高等教育改革和推进粤港澳大湾区建设的时代背景下，由清华大学与深圳市合作共建的公立研究生教育机构。

- 致力于建设成为世界一流的研究生院，
- 优先布局：能源材料、信息科技、医药健康、海洋工程、未来人居、环境生态和创新管理。
- 截至2022年12月，专职教师217人，其中院士4名，15%为国家级人才，70%拥有海外知名大学博士学位，博士后159人。
- 获国家级奖励13项，省部级奖励64项，其他科技奖励151项；已建成5个国家级和15个省部级科研机构。
- 现有硕士、博士研究生共5100多人。

清华大学深圳国际研究生院是国家教育部正式批准的

录取标准、培养要求、学位授予与清华大学研究生院**完全一致**

录取通知书、毕业证书和学位证书由**清华大学**颁发

入学和毕业院系为**清华大学深圳国际研究生院**

大数据机器学习教学及科研体会

01 课程基本情况和挑战

02 课程特色

03 科研成果

04 问题与思考

课程基本情况

- **《大数据机器学习》**
  - 2015年我院开设“大数据工程硕士培养项目”
  - 最受学生欢迎的专业基础理论课程之一
- **课程难点**
  - 如何协助教学资源受限的院校开展系统的AI课程教学
  - 如何满足终身学习以及社会各阶层人士对AI的学习需求
  - 如何在研究生课程上针对不同本科专业不同基础的学生
  - 如何将算法和软件教学和具体的硬件和应用场景结合
  - 如何撰写优秀的教辅书籍
  - 如何针对留学生进行课程设计
  - 如何通过讨论活跃课堂气氛

课程难点&特色

如何协助教学资源受限的院校开展系统的AI课程教学以及如何满足终身学习以及社会各阶层人士对AI的学习需求

如何在研究生课程上针对不同本科专业不同基础的学生

如何将算法和软件教学和具体的硬件和应用场景结合

如何撰写优秀的教辅书籍

如何针对留学生进行课程设计

如何通过讨论活跃课堂气氛

线上线下课堂

课程特色：线上线下课堂

- 课程名称：大数据机器学习
- 平台网址：https://www.xuetangx.com/course/THU08091001026
- 首轮上线时间：2018-10-15
- 视频数量（个）：113
- 视频总时长（分钟）：939.57

2018**清华大学首批“混合式教学试点课程”**

课程特色：线上线下课堂

获得学分认可的培养单位：

北京师范大学、西南民族大学、陕西服装工程学院、重庆交通大学、中原工学院、郑州大学、上海第二工业大学、青岛理工大学、空军工程大学、河北大学

- 修读人数：50161
- 近一年修读人数： 4897


光明地方 difang.gmw.cn 首页>地方频道>社会

清华大学发布首批在线认证证书项目

2018-06-21 17:19 来源：光明网

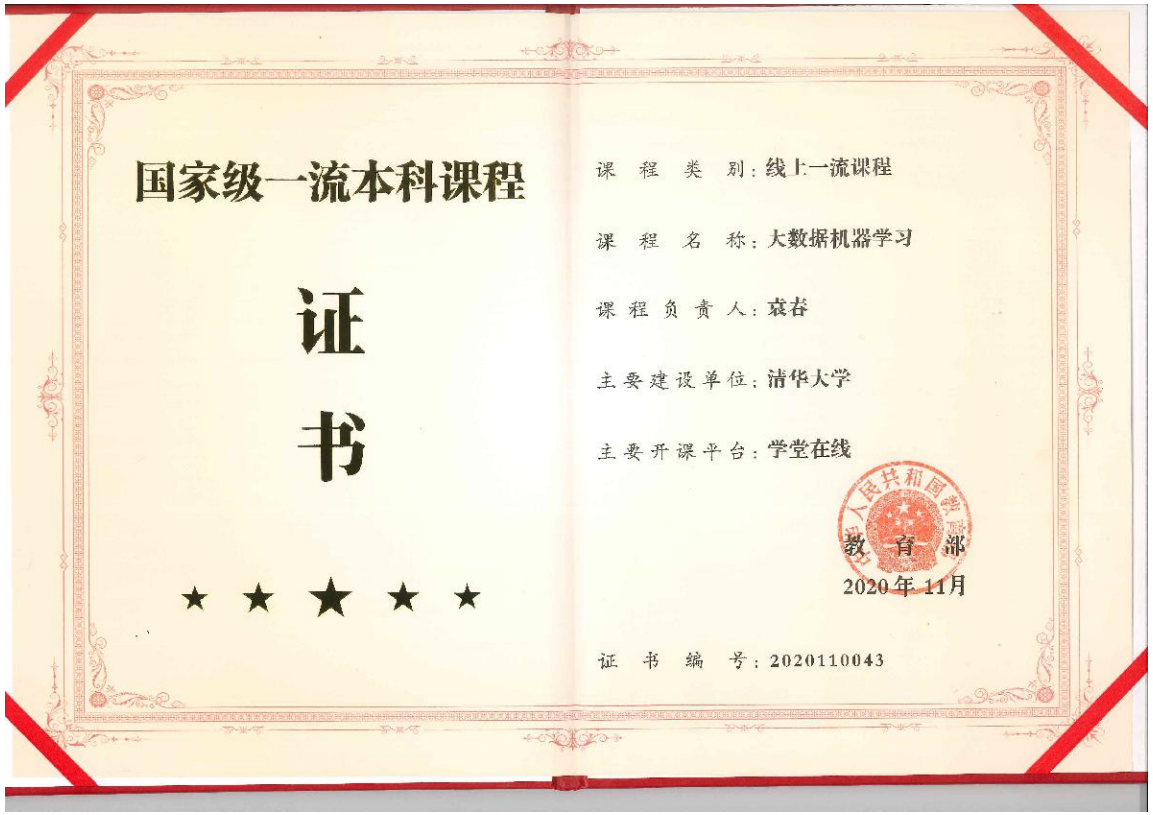
6月21日，清华大学发布首批两个在线认证证书项目，分别以公共管理认证和数据科学认证为题，并在学堂在线慕课平台证书频道上线。学习者可付费学习系列线上课程、在线完成作业，通过考试后将获得清华大学在线教育办公室及相关院系联合认定的证书。

2018年清华大学首批在线证书项目，41270选课数

课程特色：线上线下课堂

2019年**中宣部“学习强国”平台“每日慕课”推荐课**

3,279,195播放



2020年国家级线上一流课程

课程难点&特色

如何协助教学资源受限的院校开展系统的AI课程教学以及如何满足终身学习以及社会各阶层人士对AI的学习需求

如何在研究生课程上针对不同本科专业不同基础的学生

如何将算法和软件教学和具体的硬件和应用场景结合

如何撰写优秀的教辅书籍

如何针对留学生进行课程设计

如何通过讨论活跃课堂气氛

层次化教学

清華大學 Tsinghua University

本章重点

- AdaBoost算法
- 加法模型
- Boosting tree
- 残差Residual
- Gradient boost
- XGboost

课程特色：层次化教学—课堂和实验

本章重点

- AdaBoost算法
- 加法模型
- Boosting tree
- 残差Residual
- Gradient boost
- XGboost

残差网络

$$L(y, f(x)) = (y - f(x))^2$$

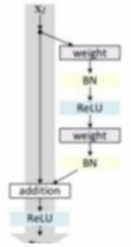
$$L(y, f_{m-1}(x) + T(x;\Theta_m)) = [y - f_{m-1}(x) - T(x;\Theta_m)]^2$$
$$= [r - T(x;\Theta_m)]^2$$

残差

$$r = y - f_{m-1}(x)$$

- 深度学习中，网络层数增多一般会有以下问题：
- 计算资源的消耗
- 模型容易过拟合
- 梯度消失/梯度爆炸问题的产生

$$x_{l+1} = x_l + \mathcal{F}(x_l, W_l)$$



EM与生成模型的关系

- 新的算法来估计参数，
- 近似算法：马尔科夫蒙特卡洛MCMC算法
- 确定性算法：变分推理，变分法是将后验分布通过某种方式分解或者假设后验概率分布有一个具体的参数形式，如高斯分布（VAE）。

无监督生成模型

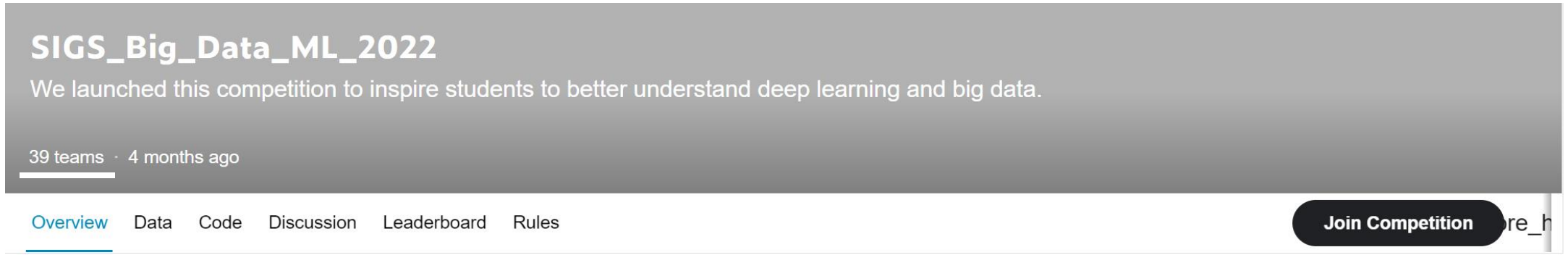
- K均值（K-Means）算法
- 自编码器（Auto-Encoder）
- 主成分分析（Principal Component Analysis）
- 玻尔兹曼机（BM）
- 生成对抗网络GAN
- 变分自编码VAE
- Flow-based Model
- DDPM:Denoising Diffusion Probabiblistic Model

EM和深度学习

Expectation-Maximization Attention Networks for Semantic Segmentation

Xia Li[1,2], Zhisheng Zhong[2], Jianlong Wu[2,3], Yibo Yang[2,4], Zhouchen Lin[2], Hong Liu[1,✉]
[1] Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University
[2] Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
[3] School of Computer Science and Technology, Shandong University
[4] Academy for Advanced Interdisciplinary Studies, Peking University
{ethanlee,zszhong,jlwu1992,ibo,zlin,hongliu}@pku.edu.cn

课程特色：层次化教学—课堂和实验



SIGS_Big_Data_ML_2022

We launched this competition to inspire students to better understand deep learning and big data.

39 teams · 4 months ago

Leaderboard

file_download Raw Data refresh Refresh

search Search leaderboard

Public Private

This leaderboard is calculated with approximately 80% of the test data. The final results will be based on the other 20%, so the final standings may be different.

#	Team	Members	Score	Entries	Last	Solution
1	cxr&wyz&yx		0.50402	194	4mo	
2	Ping Tang		0.49383	47	4mo	
3	RVVV		0.48469	57	4mo	
4	lxl-yjs		0.47940	100	4mo	
5	test		0.47704	46	4mo	

课程难点&特色

如何协助教学资源受限的院校开展系统的AI课程教学以及如何满足终身学习以及社会各阶层人士对AI的学习需求

如何在研究生课程上针对不同本科专业不同基础的学生

如何将算法和软件教学和具体的硬件和应用场景结合

如何撰写优秀的教辅书籍

如何针对留学生进行课程设计

如何通过讨论活跃课堂气氛

大数据实践教学

课程特色：大数据实践教学

HUAWEI

MindStudio 环境搭建实验手册 第 2页

1 实验介绍

MindStudio 提供您在 AI 开发所需的一站式开发环境，支持模型开发、算子开发以及应用开发三个主流程中的开发任务。

依靠模型可视化、算力测试、IDE 本地仿真调试等功能，MindStudio 能够帮助您在一个工具上就能高效便捷地完成 AI 应用开发。

MindStudio 采用了插件化扩展机制，开发者可以通过开发插件来扩展已有功能。

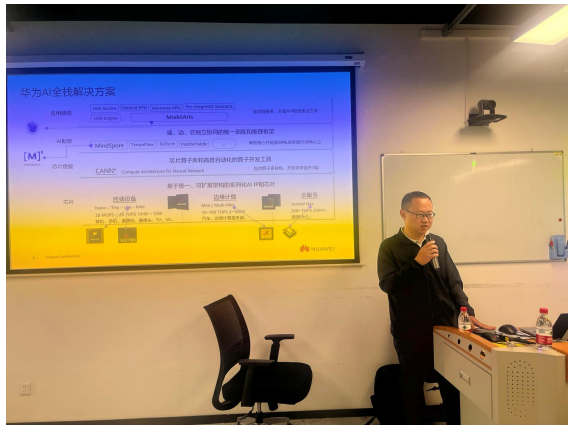
关于 MindStudio 的更多功能介绍、特性介绍，请查阅产品官网：
https://www.hiascend.com/software/mindstudio

关于 MindStudio 的详细安装与管理、常用操作、关键功能入门等，请查阅产品文档：
https://www.hiascend.com/document/detail/zh/mindstudio/50RC3/progressiveknowledge/index.html

1.1 实验目的

MindStudio 可安装在 Windows 或 Linux 上，本手册将介绍两种使用 MindStudio 开发环境的方式：

一： 通过 PC 机安装 MindStudio 开发环境；另因开发需要，同时安装深度学习框架 MindSpore1.5 和 Python3.7.5。



Ascend
昇腾万里，让智能无所不能

昇腾众智金质量奖·2022

开发者：江邦睿

清华大学深圳研究院

指导老师：袁春

获奖事迹

参加基于昇腾CANN技术栈Pytorch模型众智开发任务，独立完成TSM模型训练和推理项目，提前完成了模型交付任务，具有较强的学习能力。

开发者：陈瑶

清华大学深圳研究院

指导老师：袁春

获奖事迹

参加基于昇腾CANN技术栈TensorFlow模型众智开发任务，快速理解和学习项目知识及流程，在实践过程中能及时提出问题、且和华为积极有效互动，高质量的完成模型提前验收。

课程难点&特色

如何协助教学资源受限的院校开展系统的AI课程教学以及如何满足终身学习以及社会各阶层人士对AI的学习需求

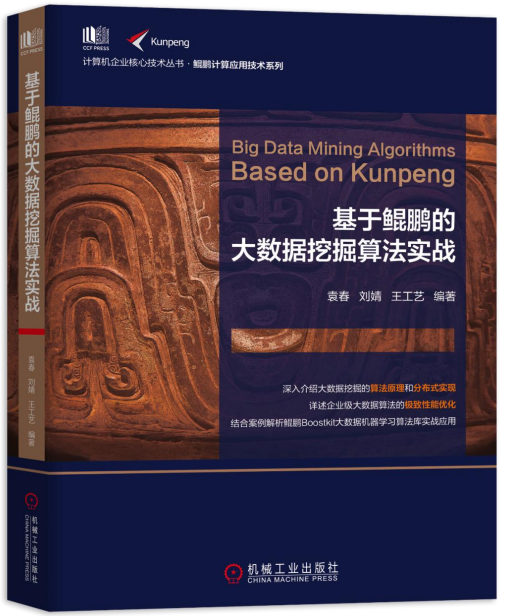
如何在研究生课程上针对不同本科专业不同基础的学生

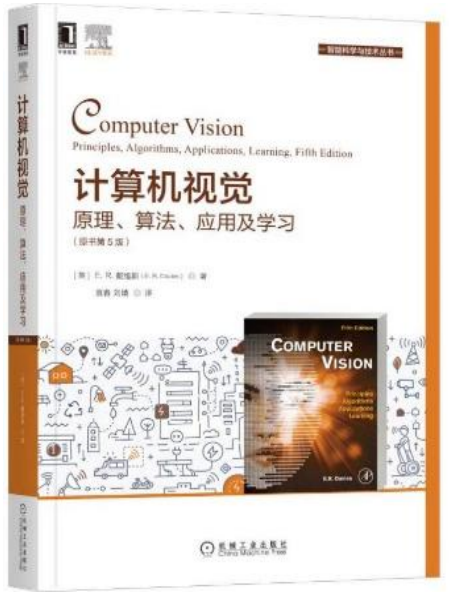
如何将算法和软件教学和具体的硬件和应用场景结合

如何撰写优秀的教辅书籍

如何针对留学生进行课程设计

如何通过讨论活跃课堂气氛



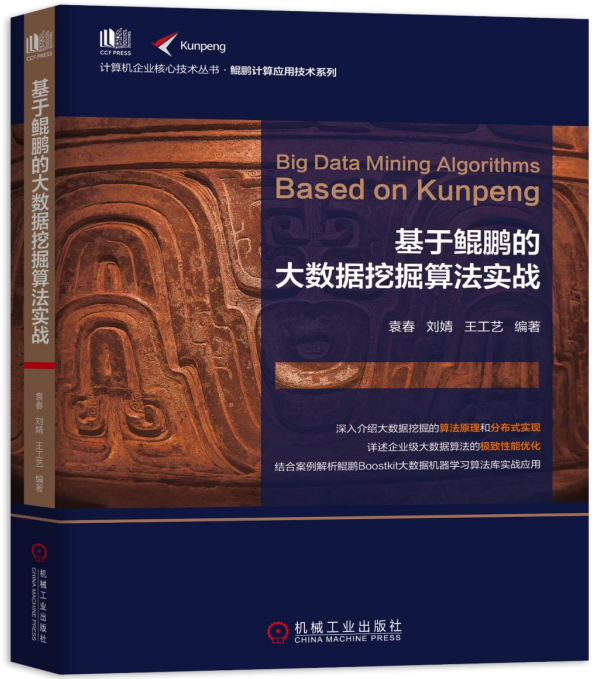


教辅书籍编写

课程特色：教辅书籍编写

- 《基于鲲鹏的大数据挖掘算法实战》机械工业出版社 2022
  - 编著 袁春 刘婧 王工艺
  - 郑炜民 院士 作序

- 图书销量 2022双12图书节
  - 大数据互联网 排行榜 第一名
  - 计算机大类　　排行榜 第二名

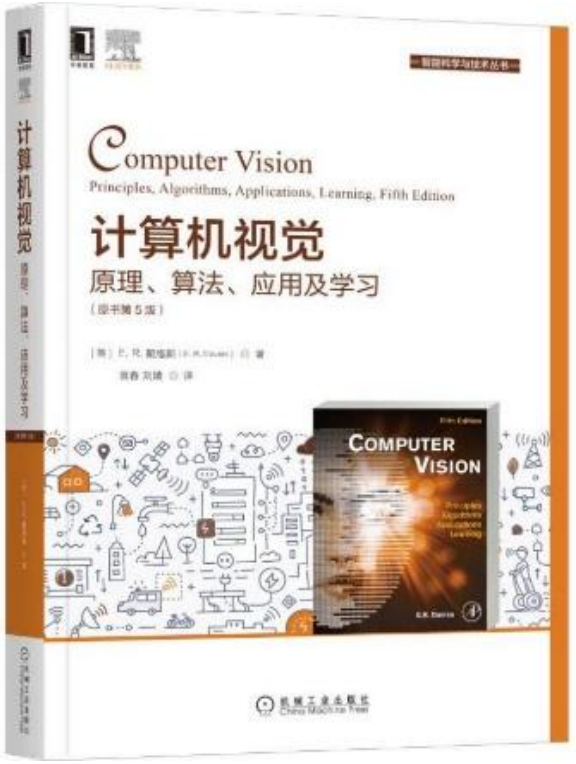


课程特色：教辅书籍编写

翻译出版：《计算机视觉：原理、算法、应用及学习》（机械工业出版社）

袁春、刘婧翻译

安徽大学	海南大学	山东建筑大学
北京师范大学珠海分校	河北大学	山西大学
电子科技大学	吉林大学	四川大学
东北大学	江南大学	天津大学
东华大学	南京航空航天大学	武汉理工大学（马房山校区）
东南大学	南京理工大学	西南交通大学
广西大学	南开大学	中国石油大学（北京）
贵州师范大学	厦门大学嘉庚学院	中国石油大学（华东）
哈尔滨工业大学（威海）	山东大学	中山大学
哈尔滨工业大学 (深圳)	山东大学齐鲁软件学院	重庆大学
哈尔滨理工大学（西区）	山东大学（威海）	



86万字，70多高校采用，第三次印刷

课程难点&特色

如何协助教学资源受限的院校开展系统的AI课程教学以及如何满足终身学习以及社会各阶层人士对AI的学习需求

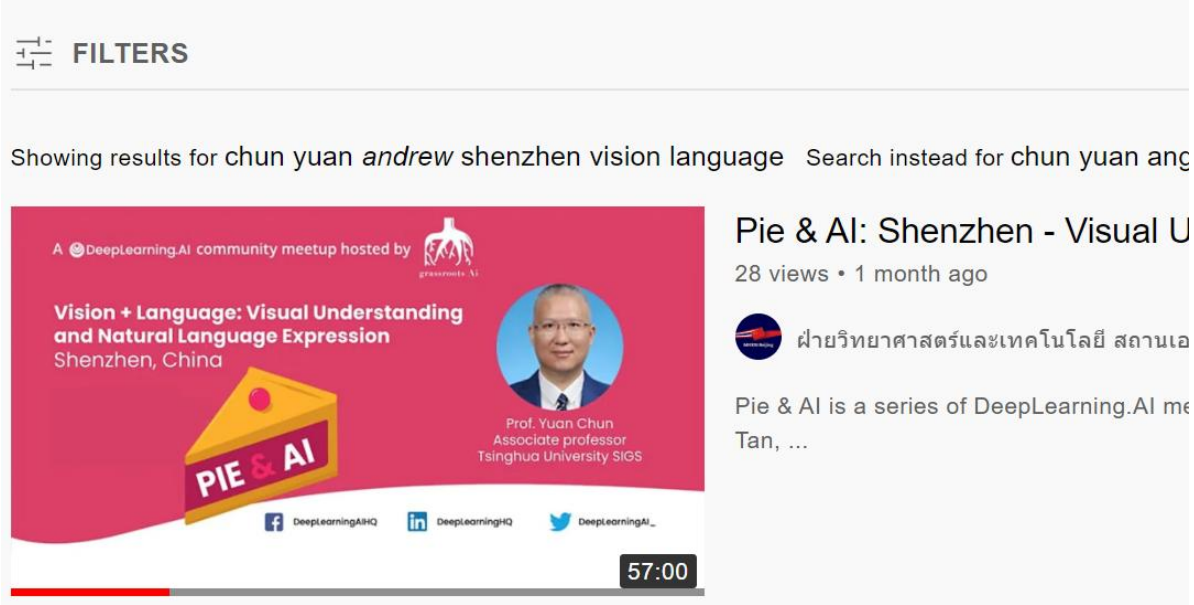
如何在研究生课程上针对不同本科专业不同基础的学生

如何将算法和软件教学和具体的硬件和应用场景结合

如何撰写优秀的教辅书籍

如何针对留学生进行课程设计

如何通过讨论活跃课堂气氛



⇨ **中英文课堂**

课程特色：中英文课堂

2017-2018 秋季/2018-2019 秋季/2018-2019 春季 TBSI/2019-2020 秋季《大数据机器学习》课程内容对比

内容	2017-2018 年度秋季中文版课程	2018-2019 年秋季英文版课程（中文讲解）	2018-2019 年度春季 TBSI 英文版课程（英文讲解）	2019-2020 年度秋季中文版课程
教材	《统计学习方法》李航 清华大学出版社 《机器学习》周志华 清华大学出版社	《Pattern Recognition and Machine Learning》Christopher M. Bishop，2006，Springer 出版	《Pattern Recognition and Machine Learning》Christopher M. Bishop，2006，Springer 出版	《统计学习方法-第二版》李航 《Reinforcement learning：An Introduction》Sutton, The MIT Press
教学/作业/考试	PPT：中文 课堂讲解：中文 平时作业：中文（12 次） 大作业：2 次 中文 期中和期末考试：2 次，中文	PPT：英文 课堂讲解：中文 平时作业：英文（11 次） 大作业：2 次，中英文 期中和期末考试：2 次，英文	PPT：英文 课堂讲解：英文 平时作业：英文（13 次） 大作业：1 次，英文 大作业口头报告：英文 期中和期末考试：2 次，英文	PPT：统计学习部分中文，强化学习部分英文 课堂讲解：中文 平时作业：11 次，统计学习部分中文，强化学习部分英文， 大作业：2 次，分层教学(基础类和竞赛类，中英文 考试：1 次考试
教学内容	第一讲：概述 第二讲：机器学习基本概念 第三讲：模型性能评估 第四讲：感知机 第五讲：KNN 第六讲：决策树和随机森林 第七讲：贝叶斯分类器和图模型 第八讲：SVM 第九讲：隐马尔科夫模型 第十讲：逻辑斯蒂回归与最大熵 第十一讲：期望最大算法 第十二讲：神经网络和深度学习 第十三讲：CNN 第十四讲：RNN	第一讲：介绍 第二讲：概率分布（一） 第三讲：概率分布（二） 第四讲：线性回归模型 第五讲：线性分类模型（一） 第六讲：线性分类模型（二） 第七讲：神经网络 第八讲：核方法 第九讲：稀疏核方法 第十讲：图模型（一） 第十一讲：图模型（二）	第一讲：介绍 第二讲：概率分布（一） 第三讲：概率分布（二） 第四讲：线性回归模型 第五讲：线性分类模型（一） 第六讲：线性分类模型（二） 第七讲：神经网络 第八讲：核方法 第九讲：稀疏核方法 第十讲：图模型（一） 第十一讲：图模型（二） 第十二讲：采样理论	第一讲：机器学习和统计学习 第二讲：感知机 第三讲：K 近邻算法 第四讲：贝叶斯分类器 第五讲：决策树 第六讲：逻辑斯蒂回归与最大熵 第七讲：SVM 与核函数 第八讲：adaboost 第九讲：EM 和隐马尔科夫模型 第十讲：条件随机场 第十一讲：无监督学习概论 第十二讲：聚类方法 第十三讲：奇异值分解和主成分分析

课程难点&特色

如何协助教学资源受限的院校开展系统的AI课程教学以及如何满足终身学习以及社会各阶层人士对AI的学习需求

如何在研究生课程上针对不同本科专业不同基础的学生

如何将算法和软件教学和具体的硬件和应用场景结合

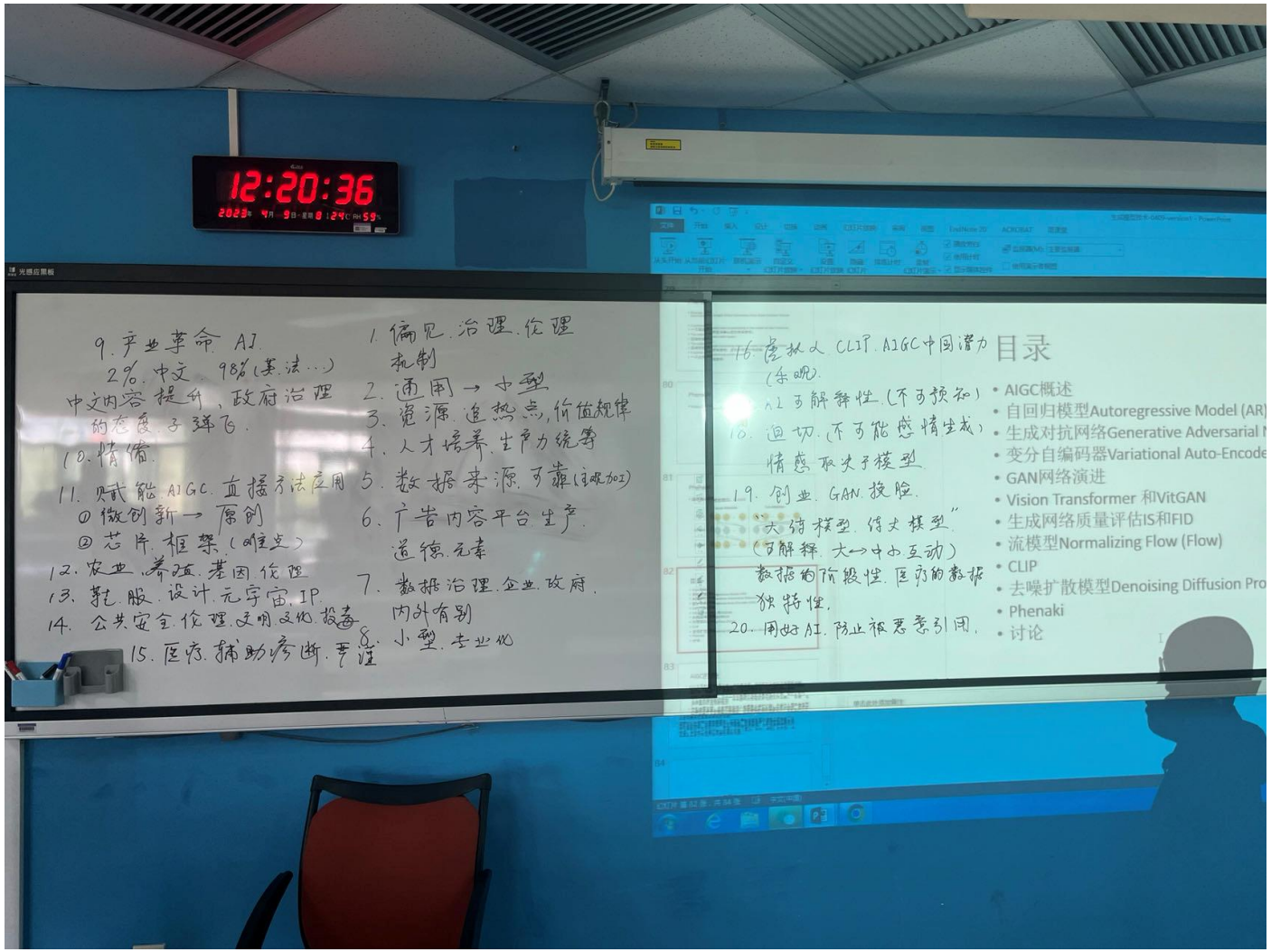
如何撰写优秀的教辅书籍

如何针对留学生进行课程设计

如何通过讨论活跃课堂气氛

课堂讨论

课程特色：课堂讨论


12:20:36
目录
AIGC概述
自回归模型Autoregressive Model (AR)
生成对抗网络Generative Adversarial
变分自编码器Variational Auto-Encode
GAN网络演进
Vision Transformer 和VitGAN
生成网络质量评估IS和FID
流模型Normalizing Flow (Flow)
CLIP
去噪扩散模型Denoising Diffusion Pro
Phenaki
讨论

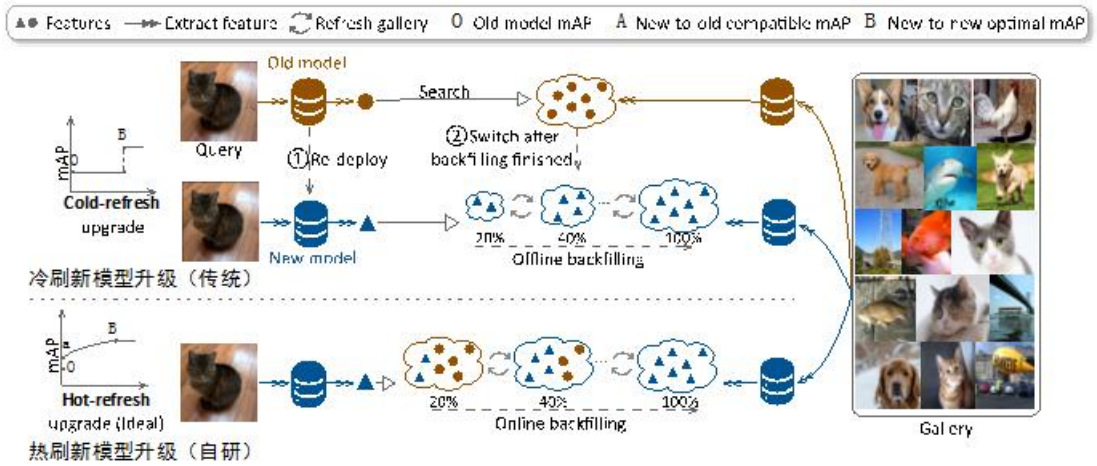

FANET: FACTOR ANALYSIS NEURAL NETWORK

课程难点及课程特色

课程难点	课程特色
如何协助教学资源受限的院校开展系统的AI课程教学以及如何满足终身学习以及社会各阶层人士对AI的学习需求	线上线下课堂
如何在研究生课程上针对不同本科专业不同基础的学生	层次化教学
如何将算法和软件教学和具体的硬件和应用场景结合	大数据实践教学
如何撰写优秀的教辅书籍	教辅书籍编写
如何针对留学生进行课程设计	中英文课堂
如何通过讨论活跃课堂气氛	课堂讨论

科研成果：模型升级更新

热刷新模型升级 ICLR2022



Forward Compatible Training for Large-Scale Embedding Retrieval Systems

Vivek Ramanujan* University of Washington† Pavan Kumar Anasosalu Vasu Apple Ali Farhadi Apple Oncel Tuzel Apple

Hadi Pouransari* Apple

Towards Backward-Compatible Representation Learning

Yantao Shen* Yuanjun Xiong Wei Xia Stefano Soatto AWS/Amazon AI

Online Backfilling with No Regret for Large-Scale Image Retrieval

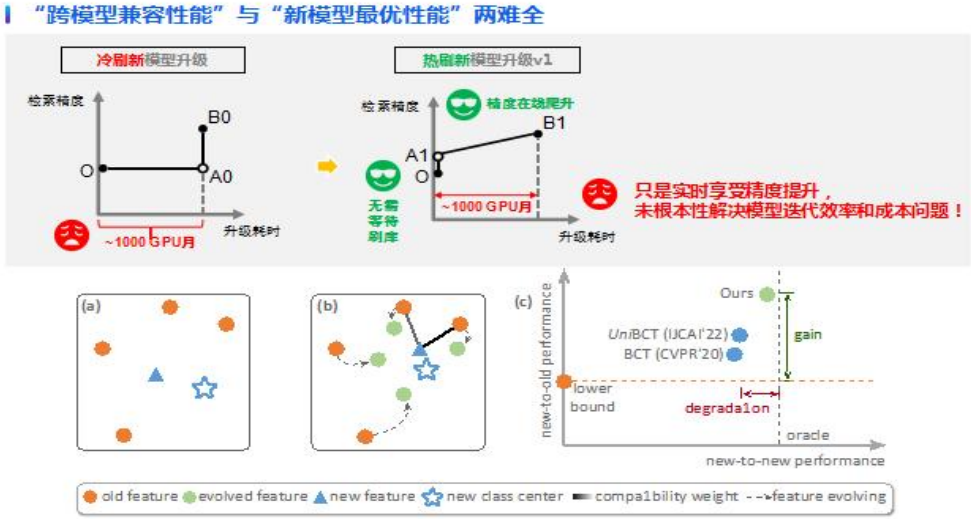
Seonguk Seo[1,†] Mustafa Gokhan Uzunbas[3] Bohyung Han[1,2] Sara Cao[3] Joena Zhang[3] Taipeng Tian[3] Ser-Nam Lim[3]

[1]ECE & [1,2]IPAI, Seoul National University [3]Meta AI

FastFill: Efficient Compatible Model Update

Florian Jaeckle* University of Oxford† Fartash Faghri Apple Ali Farhadi Apple Oncel Tuzel Apple Hadi Pouransari* Apple

达尔文式模型升级 AAAI 2023

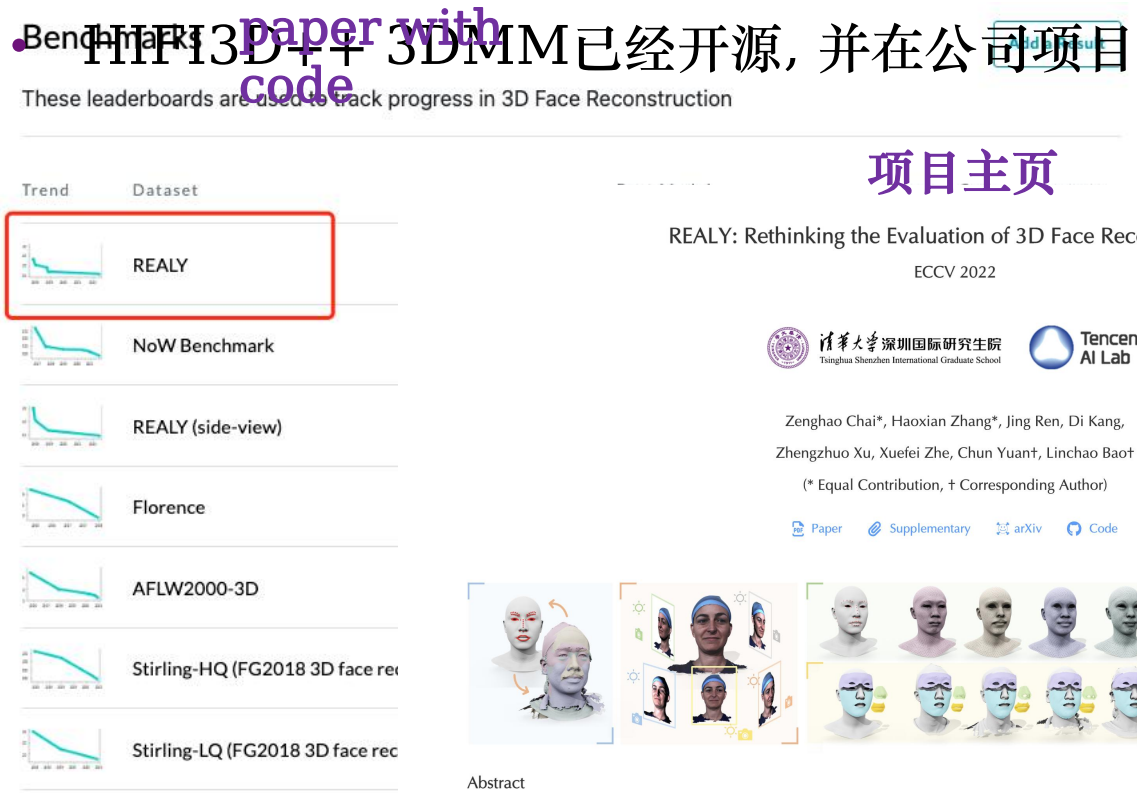


- 库存容量大
- 实时流量大
- 送检业务分布广
- 4+更新效率
- 节约1000 GPU月
- 节约200万机器成本

科研成果：3D重构

开源影响力

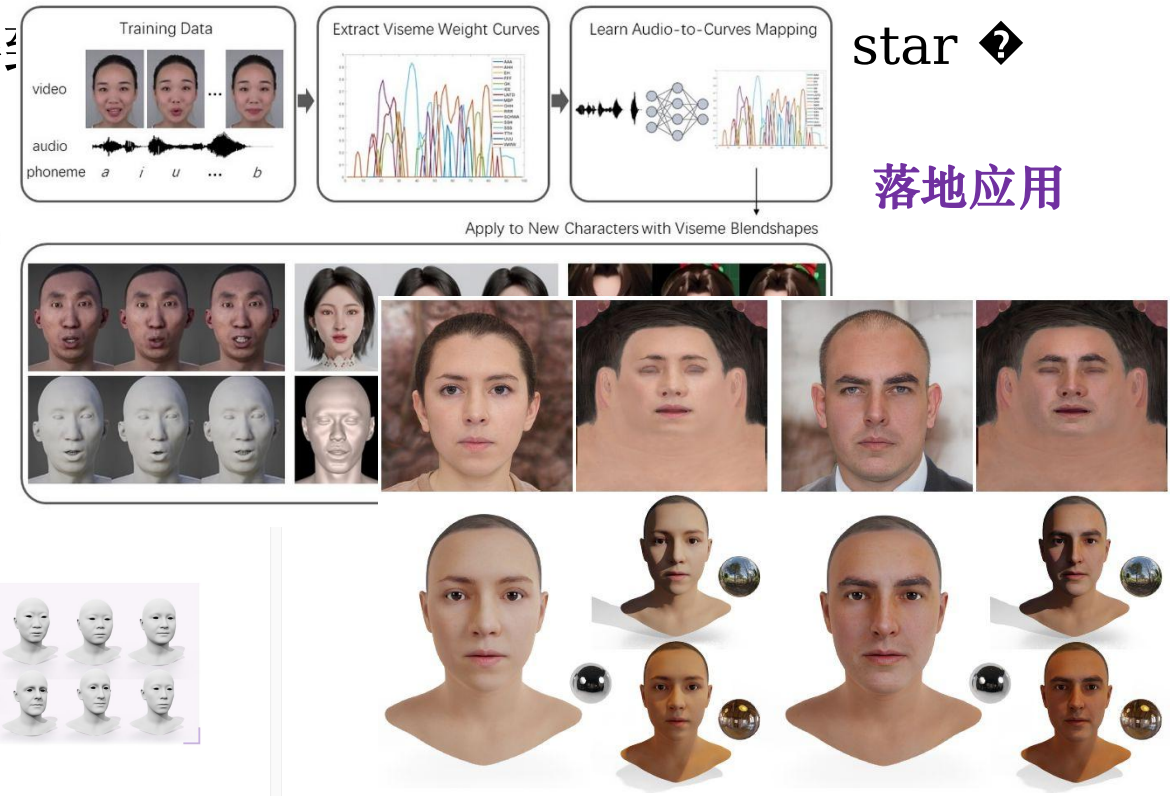
- 项目网站：http://realy3dface.com/，代码：https://github.com/czh-98/REALY
- REALY Benchmark已经开源，评测包括CVPR 2023在内17+重建方法，位列paper with code之首
- HIFI3D++ 3DMM已经开源，并在公司项目中得到 star ✪



paper with code

项目主页

落地应用



问题与思考

- 技术的迅猛发展
- 越来越多的学生从不同学科涌入本方向
- 越来越多的专业领域与本方向结合
- 实践教学任重道远
- 研究生教育与本科生教育的不同
- 与产业结合催生新的研究生培养模式

清華大學
Tsinghua University

感谢聆听

清华大学深圳国际研究生院
欢迎您

袁春

清华大学深圳国际研究生院

2023/05/13