

会议纪要

汇报人：江星洵
2021年12月21日



东南大学



汇报提纲

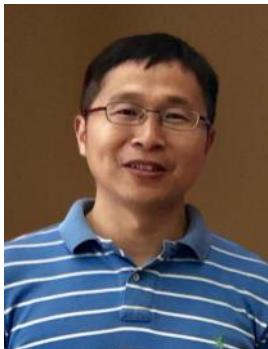
- 一. 青椒到老教
- 二. 可信计算机视觉
- 三. 深度学习大模型
- 四. 视觉分析与多模态数据理解
- 五. 腾讯视觉技术创新与行业应用专题
- 六. 硬件友好的轻量深度神经网络

汇报提纲

- 一. 青椒到老椒**
- 二. 可信计算机视觉
- 三. 深度学习大模型
- 四. 视觉分析与多模态数据理解
- 五. 腾讯视觉技术创新与行业应用专题
- 六. 硬件友好的轻量深度神经网络

一. 青椒到老椒

■人物介绍(青椒到老椒)



MSRA, 百度
王井东研究员



中科院自动化所
刘成林研究员



西北工业大学
韩军伟教授



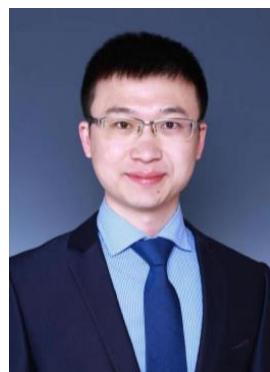
中科院计算机所
陈熙霖研究员



北京大学
王奕森
助理教授



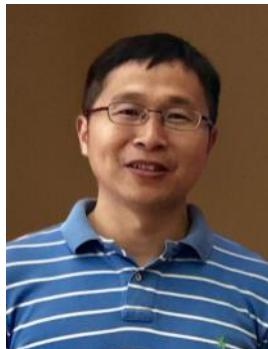
北京大学
林宙辰教授



清华大学
黄高
助理教授

一. 青椒到老椒

■人物介绍(青椒到老椒)



MSRA, 百度
王井东研究员

一. 青椒到老椒- 王井东: 如何写好一篇论文

■论文应该展现的几点

Problem

Categorization of related works

Positioning

Approach

Experimental verification

Open-sourcing

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, MARCH 2020

Deep High-Resolution Representation Learning for Visual Recognition

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao

Abstract—High-resolution representations are essential for position-sensitive vision problems, such as human pose estimation, semantic segmentation, and object detection. Existing state-of-the-art frameworks first encode the input image as a low-resolution representation through a bottleneck that is formed by connecting high-to-low resolution convolutions in series (e.g., ResNet, VGGNet), and then recover the high-resolution representation from the encoded low-resolution representation. Instead, our proposed network, named as High-Resolution Network (HRNet), maintains high-resolution representations throughout the whole process. There are two key characteristics: (i) Connect the high-to-low resolution convolution streams in parallel; (ii) Repeatedly exchange the information across resolutions. The benefit is that the resulting representation is semantically richer and spatially more precise. **Keywords**—Visual recognition, the proposed HRNet has a wide range of applications, including human pose estimation, semantic segmentation, and object detection, supporting that that HRNet is a powerful backbone for computer vision problems. All the codes are available at <https://github.com/HRNet>.

Index Terms—HRNet, high-resolution representations, low-resolution representations, human pose estimation, semantic segmentation, object detection.

1 INTRODUCTION

Deep convolutional neural networks (DCNNs) have achieved state-of-the-art results in many computer vision tasks, such as image classification, object detection, semantic segmentation, human pose estimation, and so on. The strength is that DCNNs are able to learn richer representations than conventional hand-crafted representations.

Most recently-developed classification networks, including AlexNet [1], VGGNet [1], GoogLeNet [1], ResNet [2] etc., follow the design rule of LeNet-5 [3]. The rule is depicted in Figure 1 (a): gradually reduce the spatial size of the feature maps, connect the convolutions from high resolution to low resolution in series, and lead to a low-resolution representation, which is harder processed for classification.

High-resolution representations are needed for position-sensitive tasks, e.g., semantic segmentation, human pose estimation, and object detection. The previous state-of-the-art methods adopt the high-resolution recovery process to raise the representation resolution from the low-resolution representation outputted by a classification or classification-like network as depicted in Figure 1 (b), e.g., Hourglass [4], SegNet [5], DeconvNet [6], U-Net [7], StackedHourglass [8], and encoder-decoder [9]. In addition, dilated convolutions are used to remove some down-sample layers and thus yield medium-resolution representations [4] [5] [6].

We present a novel architecture, namely High-Resolution Network (HRNet), which is able to maintain high-resolution representations through the whole process. We start from a high-resolution convolution stream, gradually add high-to-low-resolution convolution streams one by one, and connect the

multi-resolution streams in parallel. The resulting network consists of several (4 in this paper) stages as depicted in Figure 1, and the i th stage contains n streams corresponding to n resolutions. We conduct repeated multi-resolution fusions by exchanging the information across the parallel streams over and over.

The high-resolution representations learned from HRNet are not only semantically strong but also spatially precise. This comes from two aspects: (i) Our approach connects high-to-low resolution convolution streams in parallel rather than in series. Thus, our approach is able to maintain the high resolution instead of recovering high resolution from low resolution, and accordingly the learned representation is potentially spatially more precise. (ii) Most existing human schemes aggregate high-resolution low-level and high-level representations obtained by upsampling low-resolution representations. Instead, we repeat multi-resolution fusions to boost the high-resolution representations with the help of the low-resolution representations, and vice versa. As a result, all the high-to-low resolution representations are semantically strong.

We present two versions of HRNet. The first one, named as HRNetv1, only exploits the high-resolution representation composed from the high-resolution convolution streams. We apply it to human pose estimation by following the heatmap estimation scheme [10] [11]. We empirically demonstrate the superior performance on the COCO dataset.

The other is HRNetv2, which combines the representations from the high-resolution convolution streams. We apply it to semantic segmentation, object detection, and estimating segmentation uncertainty from the combined high-resolution representation. The proposed approach achieves state-of-the-art results on PASCAL Context, Cityscapes, and

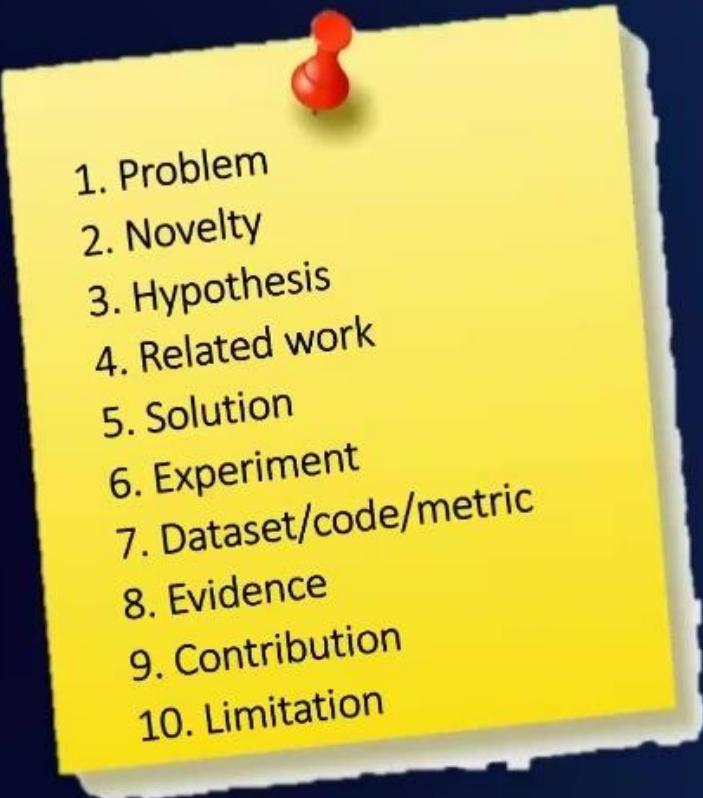
• J. Wang is with Microsoft Research, Beijing, P.R. China.
E-mail: jwang@microsoft.com



一. 青椒到老椒 – 王井东: 如何写好一篇论文

■ 检查论文是否可明确地看到几点

 ReadPaper.com
专业的学术讨论社区

- 
1. Problem
 2. Novelty
 3. Hypothesis
 4. Related work
 5. Solution
 6. Experiment
 7. Dataset/code/metric
 8. Evidence
 9. Contribution
 10. Limitation

论文搜索

Finding

文献管理

Managing

在线笔记

Noting

学术讨论

Discussing



一. 青椒到老椒

■人物介绍(青椒到老椒)



中科院自动化所
刘成林研究员

一. 青椒到老椒 – 刘成林:面向问题的研究选题

■选题及成功的标准 – 有应用价值嘛?问题解决了嘛?

科研的目的

- 为什么从事科研
 - 兴趣: 真兴趣, 假兴趣 (难以深入持久)
 - 求职或谋生: 不喜欢而从事一行, 只会徒增痛苦
 - 最好是以喜欢的工作为职业
 - 要花时间从事教学和申请项目? 那也是科研的一部分 (能促进科研)
- 选题和成功的标准
 - 选题标准: 是否真问题 (有理论意义或应用价值)
 - 成功标准: 问题解决了没有, 是否好用。或者在现有条件下是否做到最优
 - 附属标准: 发表论文, 关键看质量
 - 什么样的文章可以发表: 把研究意义、创新性和效果说清楚了 (让人看懂) 一定可以发表

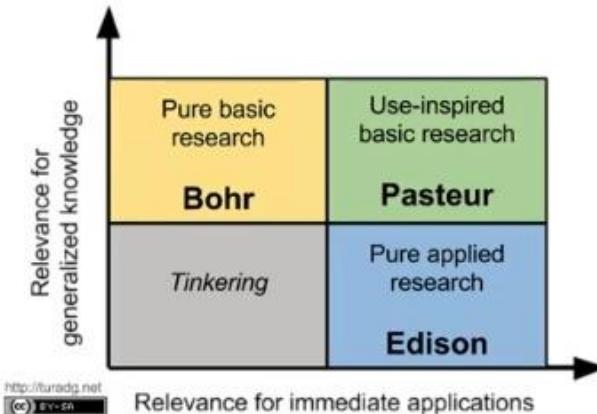
一. 青椒到老椒 – 刘成林:面向问题的研究选题

■科研问题的种类 – 基础 vs 应用

科研问题的种类

- 基础vs应用

- 基础研究(Basic/Fundamental Research): 发现或解释自然规律(why, what, how), 提出假说
- 应用研究(Applied Research): 针对实际问题提出解决办法(how to do)。
- 应用基础研究: 面向应用的通用性原理或方法
(比如机器学习)



Donald E. Stokes, Pasteur's Quadrant:
Basic Science and Technological
Innovation, 1997.

一. 青椒到老椒 – 刘成林:面向问题的研究选题

■如何发现应用问题中的问题 – 技术问题扩展延伸

如何发现问题

- 按理论意义/应用价值的标准，从需求发现
- 理论问题
 - 好奇心驱动，多问“为什么”“什么联系”，比如神经网络分类器跟贝叶斯分类器的联系
 - 已有模型方法的理论解释，比如半监督学习的收敛性、泛化性的理论保障，深度神经网络的不鲁棒性是如何引起的、如何克服等
- 应用问题
 - 围绕具体应用（如手写文档识别），可不断提出系统结构、模型设计、学习和推理算法、自适应、上下文融合等技术问题
 - 应用问题的完全解决需要长期坚持，技术问题可不断扩展、延伸，并可引申到基础理论方法的研究

一. 青椒到老椒 – 刘成林:面向问题的研究选题

■如何应对变化的学术热点 – 从热点出发，寻找自己的方向

如何应对变化的学术热点

- 模式识别领域的热点
 - 80年代: 神经网络
 - 90年代: 支撑向量机, 集成学习
 - 2000s: 概率图模型, 迁移学习, 深度学习
 - 应用问题: 人脸识别, 目标检测与识别, image captioning, re-identification, VQA
- 不要为错过热点懊悔
 - 一旦成为热点再去跟已经晚了
 - 跟起来很吃力
- 怎么办?
 - 从热点方向学习理论方法, 为我所用
 - 寻找和坚持自己的方向
 - 理论: 及时发现、预见现有方法的不足并提出解决办法, 比如深度神经网络训练的收敛性(ResNet)、鲁棒性(对抗学习、鲁棒性模型设计)、可解释性
 - 应用: 永远在路上, 可不断扩展, 坚持就好(这条路适合大多数人)

一. 青椒到老椒 – 刘成林:面向问题的研究选题

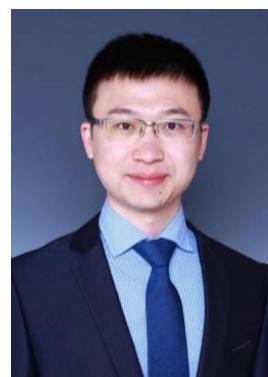
■学习之路的总结 – 博士阶段需要形成独立研究能力

总 结

- 科研路径
 - 独立开展研究: 选择问题（理论意义、应用价值），创新性方案、路线，验证效果，总结分析（包括遗留问题和扩展）
 - 有独立思考，独特的研究方向和路线，形成独特的影响
- 青椒到老椒之路
 - 博士生：形成独立研究能力
 - 青椒：开展独立研究
 - 上升：扩展研究面和深度，初步形成影响
 - 老椒：全面把握领域态势，有长远规划和大问题研究布局
- 长期坚持，不断提升

一. 青椒到老椒

■人物介绍(青椒到老椒)



清华大学
黄高
助理教授

一. 青椒到老椒 – 黃高:科研中的合作

■合作

合作的定义

合作: 指共同创作、从事；二人或多人一起工作以达到共同目的。

“collaboration is an effective interpersonal process that facilitates the achievement of goals that cannot be reached when individual professionals act on their own”

- Bronstein, L. R. 2003. A model for interdisciplinary collaboration. *Social Work*, 48(3): 297–306.

合作的动机: 通过与他人共事而或者更大的成功，达到 $1+1>2$ 的效果。



清华大学
Tsinghua University



一. 青椒到老椒 – 黄高:科研中的合作

■合作的成功要素

1. 明确的合作理由
2. 花时间寻找合适的合作对象
3. 明确各自的预期
4. 有明确的交付物
5. 要有明确的时间节点
6. 提升开会的效率
7. 利用好效率工具
8. 明确各自的职责
9. 负责任、诚实、宽容

一. 青椒到老椒 – 黄高:科研中的合作

■合作的成功要素

科研合作的成功要素



(1) 明确合作的理由

- 为了正确的理由而合作：有共同的诉求和期待，且能形成密切的合作关系。
- 充分考虑合作的风险（时间、空间约束、沟通成本、文化差异、个性特点）
- 适时地、有选择性地开展合作

一. 青椒到老椒 – 黄高:科研中的合作

■合作的成功要素

科研合作的成功要素



(2) 花时间寻找合适的合作对象

- 找最优质的合作对象：“二流的科研人员凑在一起很难做出一流的研究”
- 寻找合作对象的时间往往远小于与不合适对象开展合作耽误的时间
- 优势互补、气场相投
- 正式合作前，可以尝试性进行非正式的合作
- “青椒”要学会说“no”

一. 青椒到老椒 – 黄高:科研中的合作

■合作的成功要素

科研合作的成功要素

(3) 明确各自的预期

- 预期的产出
- 预期投入多少? (人力、时间、经费)
- 预期利益分配多少? (作者排序、知识产权)
- 在合作开始之前, 充分讨论, 消解分歧



清华大学
Tsinghua University



一. 青椒到老椒 – 黄高:科研中的合作

■合作的成功要素

科研合作的成功要素

(4) 有明确的交付物

- 有利于更好地理解双方诉求，以及各自的职责
- 有利于定义合作的边界



一. 青椒到老椒 – 黄高:科研中的合作

■合作的成功要素

科研合作的成功要素

(5) 要有明确的时间节点

- 没有DDL就没有生产力
- 了解合作项目整体进展
- 要形成定期沟通的机制，定期同步双方进展
- 协调好与其他事情的关系



一. 青椒到老椒 – 黄高:科研中的合作

■合作的成功要素

科研合作的成功要素

(6) 提升开会的效率

- 清楚会议的主要目的，做好PPT
- 线上 or 线下？
- 做好会议纪要，形成 to-do-list



清华大学
Tsinghua University



一. 青椒到老椒 – 黄高:科研中的合作

■合作的成功要素

科研合作的成功要素



(7) 利用好效率工具

- 通讯软件 (Slack、飞书)
- 在线会议软件 (腾讯会议、Zoom)
- 在线文档 (GoogleDoc、腾讯文档、Overleaf)
- 文件共享 (百度云、华为云等)
- 代码共享 (GitHub)
- 共享日历 (Google/Apple/Exchange Calendar,)

一. 青椒到老椒 – 黄高:科研中的合作

■合作的成功要素

科研合作的成功要素

(8) 明确各自的职责

- 错误的期待严重影响合作进程和效果。
- 职责尽可以清晰、明确，并且形成书面文件



一. 青椒到老椒 – 黄高:科研中的合作

■合作的成功要素

科研合作的成功要素

(9) 负责任、诚实、宽容

- 信守承诺
- 尊重彼此
- “合作不成人情在”



一. 青椒到老椒

■人物介绍(青椒到老椒)



西北工业大学
韩军伟教授

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

国科基金项目申请中的几点建议

韩军伟

西北工业大学

2021年12月

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

科学基金的基本定位

基础性:

资助基础研究，不资助工艺研究；

越是基础科学问题研究，越容易获得资助；

注重共同关心的基础科学问题。

创新性:

理论创新、原理创新、方法创新、材料创新、对象创新……

有价值:

能解决实际问题，不能闭门造车。

可行性:

要有解决问题的可行的思路和方法。

知己知彼
投其所好

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

如何选题

如何做到与基金的资助范围和学科性质相符合?

- 申请科学基金项目时应首先注意阅读《项目指南》，了解重点与优先领域，以利于确定选题范围。
- 研究类型主要支持基础研究与应用基础研究。
- 开发性、纯应用研究、工程或软科学中的可行性论证研究不建议申请科学基金。

基金申请成败的关键在于选题!

抓住两个关键点：创新性和研究基础

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

如何选题

选题的创新性

- 提出新的理论、新学说、新方法，或进行开创性的研究工作
- 在前人（也包括自己）工作的基础上有所发现，有所发明，有所前进
- 将国际科学前沿理论、方法与中国实际相结合，创造性的发展理论方法

一般分为四个层次：

新理论（新方法）研究新问题 ★★★★★

新理论（新方法）研究老问题 ★★★★

旧理论（老方法）研究新问题 ★

旧理论（老方法）研究老问题 ☆

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

如何选题

选题如何发挥自己的研究基础与学术优势？

- 选择自己有研究基础、能发挥本人学术优势的方向。
- 本着“扬长避短”的原则，尽量结合自己的研究基础；缺乏一定科学（研究）基础的“创新”是不成立的，许多情况甚至是“空想”。
- 以问题为导向，不要新以技术、新方法的应用为导向！
- 忌盲目追求“学科前沿”和“研究热点”问题，每年选题都改变，打一枪换一个地方。

在熟悉的领域做擅长的事情！

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

如何写好申请书

为什么要写这个课题?

重大需求、现有研究水平、存在问题和挑战



我怎么做这个课题?

创新研究方案、技术路线、研究内容、研究目标



为什么我能做好这个课题?

研究基础、研究团队、支撑条件、可行性分析

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

如何写好申请书

- 题目
- 摘要
- 立论依据的论述与撰写
- 项目的研究内容、研究目标、以及拟解决的关键问题
- 拟采取的研究方案及可行性分析
- 项目的特色与创新之处
- 年度研究计划及预期研究成果
- 研究基础与工作条件

逻辑清晰，突出重点，一条主线贯穿始终！

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

题目名称

- 题目是你对评审专家说的第一句话。需要创新、创新、再创新！！尽量回答“干什么、对象是什么、用什么方法、解决什么问题”。
- 简洁明确，具体清楚。不宜过长，不宜出现过多的关键词，但最好要有新意的关键词出现，包含研究视角、方法和研究对象的创新，最好能让专家一看到“题目名称”就能基本了解本申请重点要研究的问题。
- 忌讳项目名称重复，即使所提出的与以前资助项目研究内容有所不同，甚至有所创新，但名称重复很难给人以新意。

一定要到NSFC检索类似课题历年资助情况，避免重复！！

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

摘要（限400字）

包括：研究方法、内容、目标、科学意义等

- 如：“用……方法（手段）进行……研究，探索/证明……问题，对阐明……机制 / 揭示……规律有重要意义，为……奠定基础 / 提供……思路 ”
- 作用：画龙点睛
- 效果：引发评议专家兴趣，使其产生探个究竟的好奇心——“他到底要怎么做？”摘要字少，但切忌平淡无奇。（要勾起评委浓厚兴趣）
- 一定要语气坚定，旗帜鲜明，字数有限，资源宝贵，要特别注意重点突出，讲明现状、意义、课题主要研究目标、内容、思路和预期结果

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

立项依据

(研究意义、国内外研究现状及发展动态分析，需结合科学的研究发展趋势来论述科学意义；或结合国民经济和社会发展中迫切需要解决的关键科技问题来论述其应用前景。附主要参考文献目录)

为什么要这个课题？

重大需求、存在问题、有解决思路

让评审者读了申请书以后要有如下感觉：

这个研究很重要，国内外都在做，但有要害问题没有解决，申请人提出了很好的解决途径，思路很独特且合理，若沿着这条思路做几个方面的研究，有解决希望。

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

立项依据

编写建议: 分四大部分, 太多段落会逻辑费解

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

立项依据

第一部分：大背景——价值、意义、重要性

从研究所涉及的领域入手，描述项目的背景意义。要研究站得高，应用背景重大，科学意义普遍！

第二部分：国内外研究概况——提出存在问题

从“国内外对此进行了大量研究开始”，论述历史—现状—最近进展；分析存在的问题—原因；指出未来研究的方向—这样做的科学价值。

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

立项依据

第三部分：国内外研究的不足，应该从什么新思路去研究，解决的关键科学问题。理论分析：指出你思路的理论依据；实践分析：前期探索，证明用这个思路能解决问题。

第四部分：总结

指出在这个思路下的研究内容、研究目标，研究的意义

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

项目的研究内容、研究目标、以及拟解决的关键问题

研究内容要适度，即**有限目标**，研究任何新的科学问题都是在前人工作基础上开拓，且科学研究是无止境的，不可能设想在一个项目或一次研究中将所有或众多的问题都解决，因此研究内容并不是写得越多越好。**有3-4项即可，最多5项，应突出重点，有1-2项重点内容（可能取得突破）就够了。**

注意：

- (1) 详细，多写
- (2) 突出科学问题
- (3) 一般表达：研究什么，阐明什么，研究什么，揭示什么，表征什么，弄清什么；一般不提及具体方法和技术

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

拟采取的研究方案及可行性分析

如何撰写研究方案?

存在问题:

过于简单: 在方案中只有方法名称而无具体步骤。

过于繁杂: 大量罗列一些常规研究方法。

撰写要求:

研究方案与技术路线必须具体、正确、合理、可行，与研究内容对应，将其操作步骤和关键环节体现出来，说清楚如何用、用什么方法研究什么内容，主要体现对研究问题如何展开研究的一个逻辑关系。

建议：用一个流程图阐述研究方案！

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

拟采取的研究方案及可行性分析

如何撰写技术路线?

存在问题: 不清楚, 不详细。

撰写要求: 清晰、详细、逻辑性强。 切忌太具体!

撰写方法:

- ✓ 以时间顺序为主线设计技术路线
- ✓ 以研究内容为主线设计技术路线
- ✓ 分大小标题, 突出逻辑关系
- ✓ 详细地写清楚每个具体步骤
- ✓ 以图形加文字的方式表述

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

拟采取的研究方案及可行性分析

如何撰写可行性分析?

□ 研究方案可行性分析

- ✓ 理论分析
- ✓ 研究手段、方法分析
- ✓ 预实验结果分析

□ 研究队伍可行性分析

- ✓ 研究队伍的前期研究基础
- ✓ 研究队伍的知识结构

□ 研究条件可行性分析

- ✓ 所用特殊数据、文献资料等获取渠道和方法的分析
- ✓ 对所具备的研究条件进行分析，工作电脑、办公场所等
- ✓ 对项目组成员国内外合作和交流情况的分析

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

项目的特色与创新之处

- 所谓特色创新即在本项目研究领域中申请者与国内外同行所不同的，也即前人未曾有过的新学术思想、新理论、新方法、新问题或新结论。
- 包括项目的立论依据、研究内容、研究方法与手段、技术路线及实验方案上的研究与创新点进行概括、提炼并集中反映出来。
- 讲清楚项目的创新性和特色的东西，切忌乱编乱造。要么在3-4个问题层面上有1-2个创新，或在方法上有1-2个创新（我首先把这个理论从别的学科应用到另一个学科，也算创新）。
- 总体讲，创新部分不要写的太多，2个比较合适，最多3个，把独特的别人没有的东西写进去。

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

年度研究计划及预期研究结果

年度研究计划:

按年度列出, 计划要具体: 研究内容及其阶段目标; 拟组织的重要学术交流活动、国际合作与交流计划等。

预期研究结果:

理论成果: 建立/丰富/补充/填补

技术方法: 建立/完善

专利: 可望获得

论文: 国际、国内

人才培养: 青年科技骨干、博硕研究生

注意: 不要只孤零零的写发表多少篇论文, 关键是解决什么问题, 达到什么水平, 要与“研究目标”写得相吻合, 强调质量

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

研究基础与工作条件

- 这一部分论述的目的在于使评阅人对你完成所提出的课题能够有学术上的可能性判断，因此，就要尽力展示能够支撑别人对你信任的材料。不外乎是过去的研究成绩和相关经验积累。
- 很重要的一条是展示课题组的实力，特别是与国内外同行合作的支撑材料，不可能完全是一个人的材料，因此要全面。
- 应说明是否具备研究所必须的实验设备与条件，特别应说明是否具备必须的有关单位的配合。发表的高水平论文、特别是重要期刊及被SCI/EI检索的论文要有详细的目录。

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

研究基础与工作条件

存在问题:

内容过简；与申请项目无关，只写团队研究基础，不写或少写申请人的研究基础。

撰写要求:

- ✓ 要介绍与申请项目直接相关的研究结果；
- ✓ 提供有关的研究论文、成果及专利等材料；
- ✓ 工作积累也要包括项目组成员的所有信息；
- ✓ 发表过的论文，不要列低档次的论文（没有好处，反有坏处）；
- ✓ 如果文章较少，在前面介绍中列出总数，在下面的目录中就写：“**发表的与本项
目相关的论文**”。其实有几篇十分相关的论文就可以了（最好是英文国际期刊，
或国内本行最好期刊）。

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

努力做好申请书外的几件事情

1、要有充分的时间反复筛选目标

- 选题时，一般要提前半年以上，有充分的时间反复筛选目标和创新点。
- 选题时忌讳项目名称重复，即使所提出的与以前资助项目的研究内容有所不同，甚至有所创新，但名称重复则很难给人以新意。
- 检索类似课题历年资助情况，如果发现项目名称重复，则应尽可能从新的视角提出问题，首先在项目名称上尽可能给人以面目全新之感。

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

努力做好申请书外的几件事情

2、认真研究选报学科组

- 选题时，要和填报的申请代码联系起来考虑，实际上，申请代码不同，即学科组不同，竞争程度、评阅人也不同。
- 要参考近年来各学科代码下的资助数目来考虑选题。某些学科资助类似项目已很多，若无好的基础和很强的创新，就很难争取到。注意不同学部指南内容的“重迭”部分，相关学科从不同视觉看待同一个问题。
- 尽量选择熟悉圈子所对应的学科组。

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

努力做好申请书外的几件事情

3、加强学术交流，让同行认识你

- 要多与国内外同行开展学术交流，让同行认识和了解你的研究成果和基础条件。
- 青年教师更要多参加国内的学术会议，宣传自己，取得支持，特别是认识那些目前正在承担基金课题的同行，这些人很可能就是你申报题目的同行评议人。
- 相对长期的过程，早谋划，早积累，成功是留给有准备的人！

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

努力做好申请书外的几件事情

4、换位思考，多从评审人角度审视申请书

- 评审人时间有限，评审项目数量多，申请书要逻辑清楚，重点突出。
- 多用图表等形象展示的技巧，少用大篇幅文字和公式，要快速吸引评审人眼球，帮评审人节省时间。
- 请一位熟悉你的小同行专家给出建议。同时，请一位和你不是同一研究领域的专家给出建议。

一. 青椒到老椒— 韩军伟: 国科基金申请的建议

■国科基金申请中的几点建议

做好项目答辩的几点建议

- 基金项目进入答辩有70%成功率，但也意味着竞争更加激烈。
- 申请书主要是面向小同行专家，答辩一般主要是面向大同行专家，一定要用大同行能够理解的语言和角度讲述你的创新。
- 重视PPT与申请书的不同，是一种更形象、更生动展示你研究工作的方式，要善用巧用。
- 反复练习，严格控制时间，着正装，精神饱满，声音洪亮，有自信。
- 回答问题要有礼貌，不是辩论，更不是讲课！

汇报提纲

- 一. 青椒到老椒
- 二. 可信计算机视觉
- 三. 深度学习大模型
- 四. 视觉分析与多模态数据理解
- 五. 腾讯视觉技术创新与行业应用专题
- 六. 硬件友好的轻量深度神经网络

二. 可信计算机视觉

■人物介绍(可信计算机视觉)



北京邮电大学
邓伟洪



中科科学技术大学
黄怀波

二. 可信计算机视觉 – 邓伟洪: 人脸公平分析

■ 人脸公平性分析

腾讯会议

Adaptive margin (1): Reinforcement Learning

RL-RBN adaptively find optimal margins for different races with the deep Q-learning method. In deep Q-learning, action means the change of margin (increase, decrease or unchanged) in loss functions, i.e. Arcface and Cosface; rewards are designed according to the skewness of intra/inter-class distances between races.

The diagram illustrates the RL-RBN framework. It starts with 'Ethnicity aware training datasets' showing four groups: Caucasian, Indian, Asian, and African. These feed into a 'CNN' block. The output of the CNN is processed by an 'Adaptive margin loss' block, which calculates $L_{RBN}(m_j(t))$. This loss is used to 'Give actions to change margin for different races'. The 'Agent' block receives these actions and the 'Current state for each group: $s^t = \{\text{Group, Margin, Bias}\}$ '. The 'Agent' also provides feedback to the 'Adaptive margin loss' block. The 'Deep Q-learning' block uses 'Offline samples $\{(s^t, a^t, r^t, s^{t+1})\}$ ' to train a 'DQN' network. The 'Offline sampling' block shows the process of collecting 'Offline samples' from 'Current state' $s^t = \{G, M^t, B^t\}$, where G is race group, M is margin, and B is bias. It involves a 'CNN' to calculate 'Inter distance' and 'Intra distance' from a face image, and then determining the 'Action' ($a^t = \{0, 1, 2\}$), 'Next state' ($s^{t+1} = \{G, M^{t+1}, B^{t+1}\}$), and 'Reward' ($r^t = R^{t+1} - R^t$).

Figure 6. An illustration of our method. **Offline sampling:** We varies margin for each race group to collect some training samples , i.e. (s^t, a^t, r^t, s^{t+1}) , before training DQN. **Deep Q-learning network:** With these samples, DQN is trained to approximate the Q-value function, and the reward is determined by the skewness of inter/intra-class distance between races. Then, adaptive margin policy for agent can be generated according to Q-value. **Adaptive margin:** We train a race balanced network with a fixed margin for Caucasians and adaptive margins for each colored-face which changes at each training step guided by agent.

[1] Mei Wang, Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. CVPR 2020: 9322-9331. 34

三分会场一的屏幕共享

场一 大学 吴祖煊-复旦大学 黄怀波-中科院自动化所

二. 可信计算机视觉 – 邓伟洪: 人脸公平分析

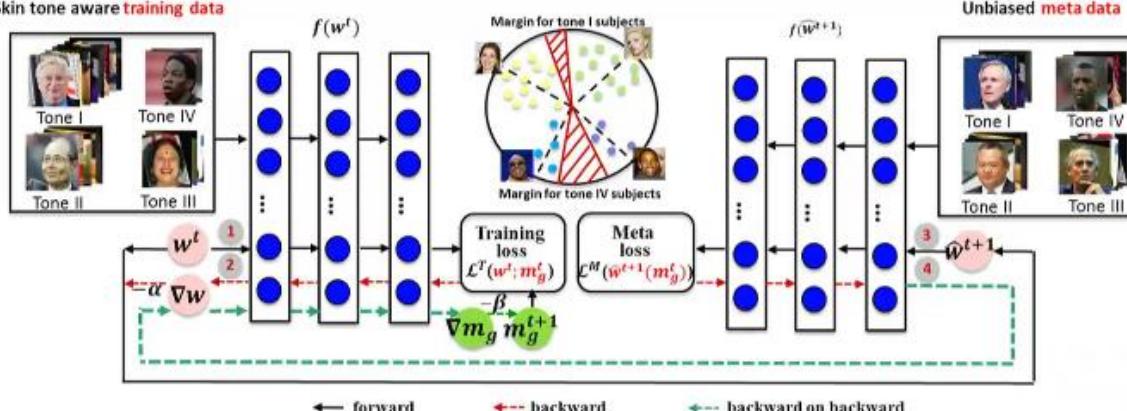
■ 人脸公平性分析

腾讯会议

Adaptive margin (2): Meta Learning

During training, we treat model optimization as the objective of inner algorithm trained by training data, and treat margin optimization as the objective of outer algorithm trained by meta data.

- Outer algorithm evaluate the bias of the learned model on meta data and dynamically output margins in adaptive margin loss function;
- Inner algorithm optimize the model guided by adaptive margin loss function on training data.



[1] Mei Wang, Yaobin Zhang, Weihong Deng. Meta Balanced Network for Fair Face Recognition. Accepted by PAMI.

38

场一
分会场一的屏幕...

大学
吴祖煌-复旦大学

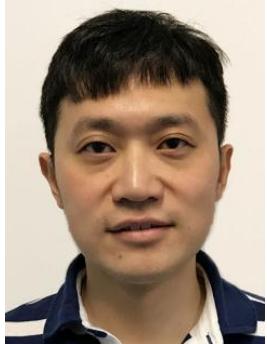
黄怀波-中科院自动化所

汇报提纲

- 一. 青椒到老椒
- 二. 可信计算机视觉
- 三. 深度学习大模型
- 四. 视觉分析与多模态数据理解
- 五. 腾讯视觉技术创新与行业应用专题
- 六. 硬件友好的轻量深度神经网络

四. 视觉与多模态数据理解

■人物介绍(可信计算机视觉)



浙江大学
杨易

四. 视觉与多模态数据理解 – 杨易：视频研究发展

■视频研究的发展

The screenshot shows a Tencent Meeting interface. At the top, it says '腾讯会议'. On the left, there's the Zhejiang University logo and the text '浙江大学 ZHEJIANG UNIVERSITY'. The main content area has a large watermark of the university's seal. The title 'Overview' is centered above three bullet points. To the right, there's a sidebar with two video thumbnails: one for '杨易的屏幕共享' (Yang Yi's screen sharing) showing a waterfall, and another for '分会场三-技术' (Subvenue 3 - Technology) showing a computer screen with a presentation slide.

Overview

- New architectures
 - 2D/3D Conv to Transformer
- New datasets.
- More efficient methods
 - Replace expensive modules
 - Decouple expensive operations
 - Search for lightweight architectures
 - Fast inter-frame correlation operations
 - Sparse sampling
- Better semantic alignment

Towards large-scale video model training

Towards better speed-accuracy tradeoff

Towards learning with multi-modalities

Recognition, LEarning, Reasoning

杨易的屏幕共享

杨易的屏幕共享

分会场三-技术

四. 视觉与多模态数据理解 – 杨易：视频研究发展

■视频研究的发展

腾讯会议

2021 APR: Video Analysis
New architectures

ZHEJIANG UNIVERSITY

CNN

- Strong inductive bias: local connectivity and translation equivariance
- Capture short-range spatio-temporal information

Transformers

- Less inductive bias
- Model both local dependencies and global long-range dependencies

杨易的屏幕共享

四. 视觉与多模态数据理解 – 杨易：视频研究发展

■ 视频研究的发展

腾讯会议

The slide features the Zhejiang University logo and title '2021 APR: Video Analysis New architectures'. It includes three diagrams illustrating convolutional operations: (a) 2D convolution, (b) 2D convolution on multiple frames, and (c) 3D convolution. A blue arrow points from these diagrams to a detailed diagram of the ViViT architecture, which processes video frames through tokenization, multi-head dot-product attention, and an MLP head.

2021 APR: Video Analysis
New architectures

(a) 2D convolution

(b) 2D convolution on multiple frames

(c) 3D convolution

• Transformers

- Embed 3D tablet
- Consider global spatio-temporal dependencies

• However, the computational cost is significantly increased

- 5 times computational cost compared to 3D counterparts

ViViT: A Video Vision Transformer, ICCV 2021

Recognition, LEarning, Reasoning

杨易的屏幕共享

01/51

四. 视觉与多模态数据理解 – 杨易：视频研究发展

■视频研究的发展

The screenshot shows a presentation slide titled "2021 APR: Video Analysis Efficiency". The slide is displayed in a Tencent Meeting window. In the top left corner, the Zhejiang University logo and name are visible. The main content area contains a bulleted list of five items related to video analysis efficiency. The background features a large, faint watermark of the university's seal.

• Replace expensive modules
• Decouple expensive operations
• Search for lightweight architectures
• Fast inter-frame correlation operations
• Sparse sampling

腾讯会议

2021 APR: Video Analysis Efficiency

浙江大学
ZHEJIANG UNIVERSITY

杨易的屏幕共享

四. 视觉与多模态数据理解 – 杨易：视频研究发展

■ 视频研究的发展

腾讯会议

浙江大学 ZHEJIANG UNIVERSITY

2021 APR: Video Analysis
Efficiency: Replace expensive modules

The diagram illustrates a workflow for egocentric action recognition. It starts with a sequence of frames labeled 'Expensive'. These frames are processed by a 'Detector' module, which outputs to 'VerbNet' and 'NounNet'. 'VerbNet' leads to 'RoIAlign' and 'Local Alignment'. 'NounNet' leads to 'Global Pooling' and 'Broadcast'. 'RelAlign' connects the outputs of 'RoIAlign' and 'Global Alignment'. A large blue arrow points from this initial stage to a second stage. In the second stage, 'Distracting Objects' are identified (e.g., Bottle, Potato, Bowl, Pan, Bag). An active object ('Potato') is highlighted in a heatmap labeled 'Motion-Relevant Region'. A robot icon indicates the final output: 'What is the active object? Active object: "Potato"'.

• Object detectors or human gaze are expensive (computational cost, labor-intensive annotations)
• Leverage the information learned from the actor's motion to enable efficient egocentric action recognition

Symbiotic Attention for Egocentric Action Recognition with Object-centric Alignment , TPAMI 2020

Interactive Prototype Learning for Egocentric Action Recognition, ICCV 2021

Recognition, LEarning, Reasoning

杨易的屏幕共享

四. 视觉与多模态数据理解 – 杨易：视频研究发展

■ 视频研究的发展

The slide is titled "2021 APR: Video Analysis Efficiency: Replace expensive modules". It features two main sections:

- Video-level aggregation**: Shows a sequence of frames being processed by a network to aggregate video-level features.
- Replace expensive networks with cheap/shallow networks**: Shows a similar process where more complex modules are replaced by simpler ones.

A blue arrow points from the first section to the second. Below the sections is the title of a paper: "FASTER Recurrent Networks for Efficient Video Classification, AAAI 2020".

Below this is another section titled "Teach the shallow student network through confidence distillation". It shows a diagram of a teacher network processing clips of a video to produce probabilities and confidence, which are then used to train a student network. The student network processes full-sized videos and aggregates Top-K results to produce its own output.

The title of this paper is "Efficient Action Recognition Using Confidence Distillation, CVPR 2021".

At the bottom of the slide, the text "Recognition, LEarning, Reasoning" is displayed.

杨易的屏幕共享

四. 视觉与多模态数据理解 – 杨易：视频研究发展

■ 视频研究的发展

腾讯会议

浙江大学 ZHEJIANG UNIVERSITY

2021 APR: Video Analysis
Efficiency: Decouple expensive operations

Joint spatio-temporal attention is expensive

Query

Space-time neighborhood

Joint Space-Time Attention (ST)

Divide spatio-temporal attention into spatial attention and temporal attention

Temporal neighborhood

Query

Spatial neighborhood

Divided Space-Time Attention (T+S)

Time Att.

Space Att.

MLP

$z^{(t-1)}$

z^t

$z^{(t)}$

Divided Space-Time Attention (T+S)

Is Space-Time Attention All You Need for Video Understanding? ICML 2021

Recognition, LEarning, Reasoning

杨易的屏幕共享

四. 视觉与多模态数据理解 – 杨易：视频研究发展

■ 视频研究的发展

腾讯会议

浙江大学 ZHEJIANG UNIVERSITY

2021 APR: Video Analysis
Efficiency: Search for lightweight architectures

The diagram illustrates the evolution from X3D to MoViNets. On the left, the X3D architecture is shown as a sequence of input frames (γ_t) processed by a stack of residual blocks (res_1, res_2, res_3) to produce a prediction. The dimensions are indicated as $\gamma_s \times \gamma_s \times \gamma_t$. On the right, the MoViNets architecture is depicted as a parallel processing pipeline. It takes input frames and processes them through causal convolutions and stream buffers. The output is then aggregated via 3D pooling and a final decision layer. A legend defines the Stream Buffer icon.

- Manually select architecture parameters

- A new search space for Neural Architecture Search
 - The diversity of architectures is improved
 - Introduce Stream Buffers
 - Enable constant memory
 - Enable online inference.

X3D: Expanding Architectures for Efficient Video Recognition , CVPR 2020

MoViNets: Mobile Video Networks for Efficient Video Recognition, CVPR 2021

Recognition, LEarning, Reasoning

杨易的屏幕共享

四. 视觉与多模态数据理解 – 杨易：视频研究发展

■ 视频研究的发展

腾讯会议

The slide is titled "2021 APR: Video Analysis Efficiency: Sparse sampling". It features two main diagrams illustrating different modeling approaches.

ActBERT Model (Left):

- Pre-extracted video features and region features
- Associate video and text via contextual learning
- Decouple action and object from a text description

The diagram shows the architecture of ActBERT. It takes Global stacked frames and Local object regions as input. These are processed through visual (action) embedding, token embedding, segment embedding, and position embedding. The resulting tokens are fed into an "ActBERT" block, which performs cross-modal matching, masked language modeling, masked action (verb) classification, and masked object (noun) classification.

CLIPBERT Model (Right):

- End-to-end modeling
- Contrastive Learning
- Clip Sparse Sampling

The diagram illustrates the CLIPBERT architecture. It starts with a "Video" input, which undergoes "Sparse Sampling" to produce "Clip c_{τ_2} ". This clip is processed by a "CNN" and "Spatial Downsampling" to extract "Clip Features". These features are combined with "Text Features" (from "Text Embedding") and "Type Embedding" (from "E[CLS], E[TXT], E[SEP], E[VID], E[VID]") via "Temporal Fusion" and "2D Position Embedding" to form the input for a "Transformer". The final output is a prediction p_{τ_1} .

Bottom Text:

ActBERT: Learning Global-Local Video-Text Representations, CVPR 2020 (Oral)

Less is More: CLIPBERT for Video-and-Language Learning via Sparse Sampling, CVPR 2021 (Oral)

Recognition, LEarning, Reasoning

四. 视觉与多模态数据理解 – 杨易：视频研究发展

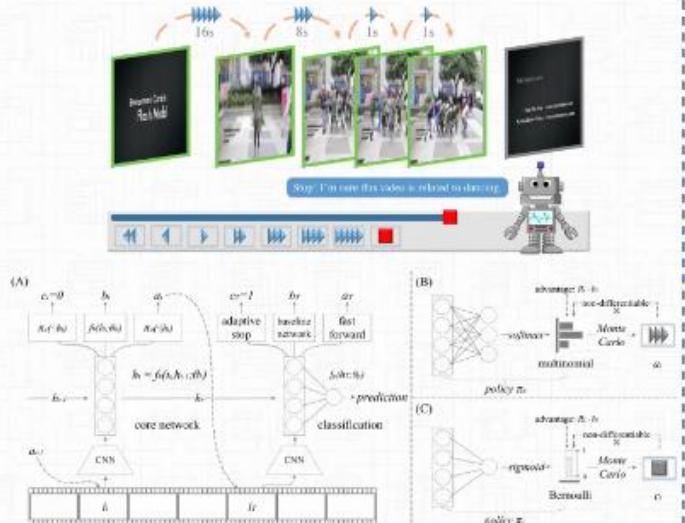
■ 视频研究的发展

腾讯会议



浙江大学
ZHEJIANG UNIVERSITY

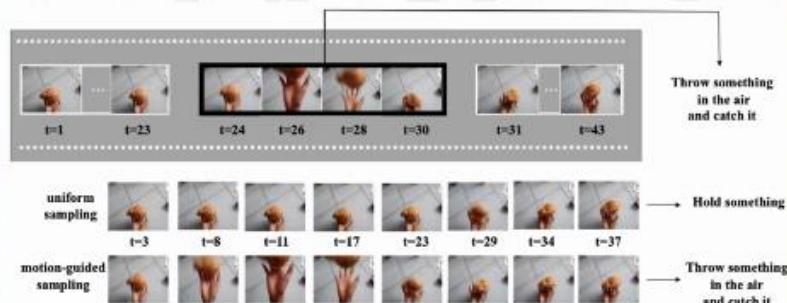
- Leverage RL to adaptive sample useful frames



Watching a small portion could be as good as watching all:
Towards efficient video classification, IJCAI 2018

2021 APR: Video Analysis Efficiency: Sparse sampling

- Use motion to guide the selection of important clips
- Enable end to end training



MGSSampler: An Explainable Sampling Strategy for Video Action Recognition, ICCV 2021

Recognition, LEarning, Reasoning

四. 视觉与多模态数据理解 – 杨易：视频研究发展

■视频研究的发展

腾讯会议

ZHEJIANG UNIVERSITY

Discussion

- Towards large-scale video model training
 - New architectures that can be adapted to large datasets
 - Modeling spatio-temporal relation, global (long) dependencies as well as local (short) dependencies
 - Larger datasets & more diverse (open) domains
- Towards better speed-accuracy tradeoff
 - Efficient architecture: mobile-friendly (consider memory, GPU/CPU constraints)
 - Adaptative and sparse frame/clip sampling
- Towards learning with multi-modalities
 - Effectively leverage weak / noisy supervisions
 - Larger clean multi-modal datasets; noisy-robust multi-modal feature learner
 - Improve modal interaction
 - Consider more modalities (audio, text, OCR, speech, ...)
 - Better multi-modal semantic alignment

Recognition, LEarning, Reasoning

杨易的屏幕共享

Q&A

