http://dmirlab.com

VALSE2019 Sharing

Speaker : 章浩

Time : 2019/05/17

Outline

- About VALSE
- 弱监督视觉理解与主动学习
- 以人为中心的视觉理解
- 视觉中的知识推理
- CV+NLP 2.0
- Poster 选讲

Outline

http://dmirlab.com

- About VALSE
- 弱监督视觉理解与主动学习
- 以人为中心的视觉理解
- 视觉中的知识推理
- CV+NLP 2.0
- Poster 选讲

Outline

VALSE 2019 合肥

VALSE年度研讨会的主要目的是为计算机视觉、图像处理、模式识别与机器学习研究领域内的中国青年学者（以80后为主）提供一个深层次学术交流的舞台。在这个舞台上，我们恪守并倡导理性批判、勇于探索、实证、创新等科学精神；在这个舞台上，我们倡导自由平等原则下、理性而纯学术的百家争鸣和思想交锋；这个舞台上，我们期望欣赏到国内青年学者越来越优美的学术华尔兹（VALSE）。通过这个舞台，我们期望促进国内青年学者的思想交流和学术合作，从而在相关领域做出重量级学术贡献，提升中国学者在国际学术舞台上的学术影响力。

截至目前，VALSE已成功举办8届，分别为 VALSE 2011（杭州），VALSE 2012（西安），VALSE 2013（南京），VALSE 2014（青岛），VALSE 2015（成都），VALSE 2016（武汉），VALSE 2017（厦门），VALSE 2018（大连）。VALSE 2019 将于2019年4月在合肥举行，由中国人工智能学会与安徽大学汤进教授团队联合合肥工业大学、中国科技大学等团队共同承办。

Outline

http://dmirlab.com

- About VALSE
- 弱监督视觉理解与主动学习
- 以人为中心的视觉理解
- 视觉中的知识推理
- CV+NLP 2.0
- Poster 选讲

Weakly | Active Learning

- What Weakly ?
- Mingming Chen, NKU
  - Multi-scale Backbone Architecture
  - 通用视觉基元属性感知
  - 互联网大数据自主学习
- Qixiang Ye, CAS
  - 粗粒度的弱监督标记
  - Towards Self-Learning

Weakly | Active Learning

- Shengjun Huang, NHU
  - Active Learning
- Summary (个人观点，真伪自辩)

Weakly | Active Learning

- What Weakly ?
- Mingming Chen, NKU
  - Multi-scale Backbone Architecture
  - 通用视觉基元属性感知
  - 互联网大数据自主学习
- Qixiang Ye, CAS
  - 粗粒度的弱监督标记
  - Towards Self-Learning

Weakly | Active Learning

- 弱监督分类：
  - 不完全监督：只有**一部分**训练数据具备**标签**
  - 不确切监督：训练数据只具备**粗粒度标签**
  - 不准确监督：给出的**标签并不总是真值**（标签有噪声）
- 弱监督的含义：
  - 弱监督给出的**标签**会在某种程度上**弱于**我们面临的任务所要求的**输出**
- 研究背景：
  - 数据集很重要，但是标注成本很大
  - 研究方法难选：深度神经网络**共性**技术，视觉**基元属性**感知。

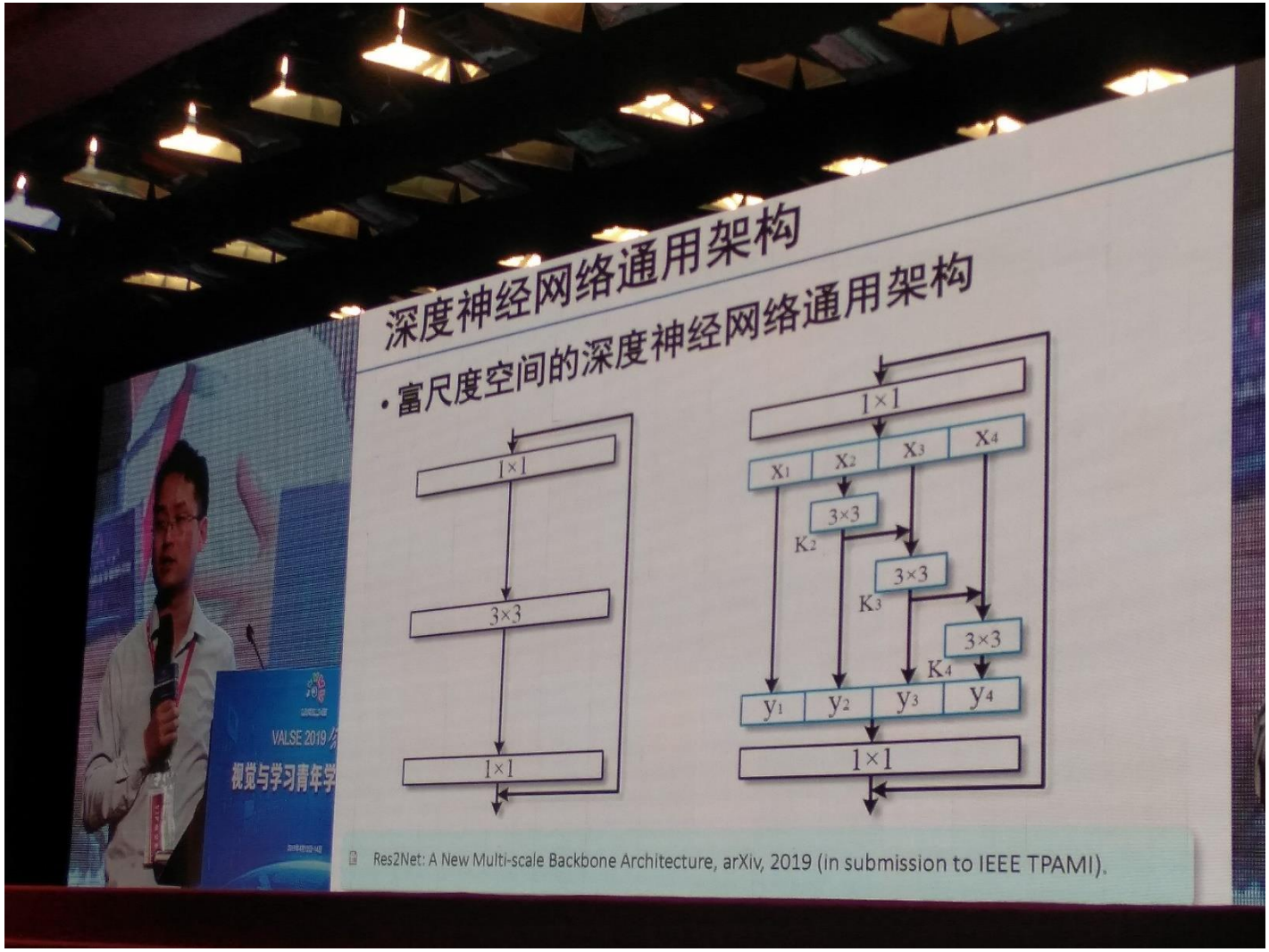
Weakly | Active Learning

- What Weakly ?
- Mingming Chen, NKU
  - Multi-scale Backbone Architecture
  - 通用视觉基元属性感知
  - 互联网大数据自主学习
- Qixiang Ye, CAS
  - 粗粒度的弱监督标记
  - Towards Self-Learning

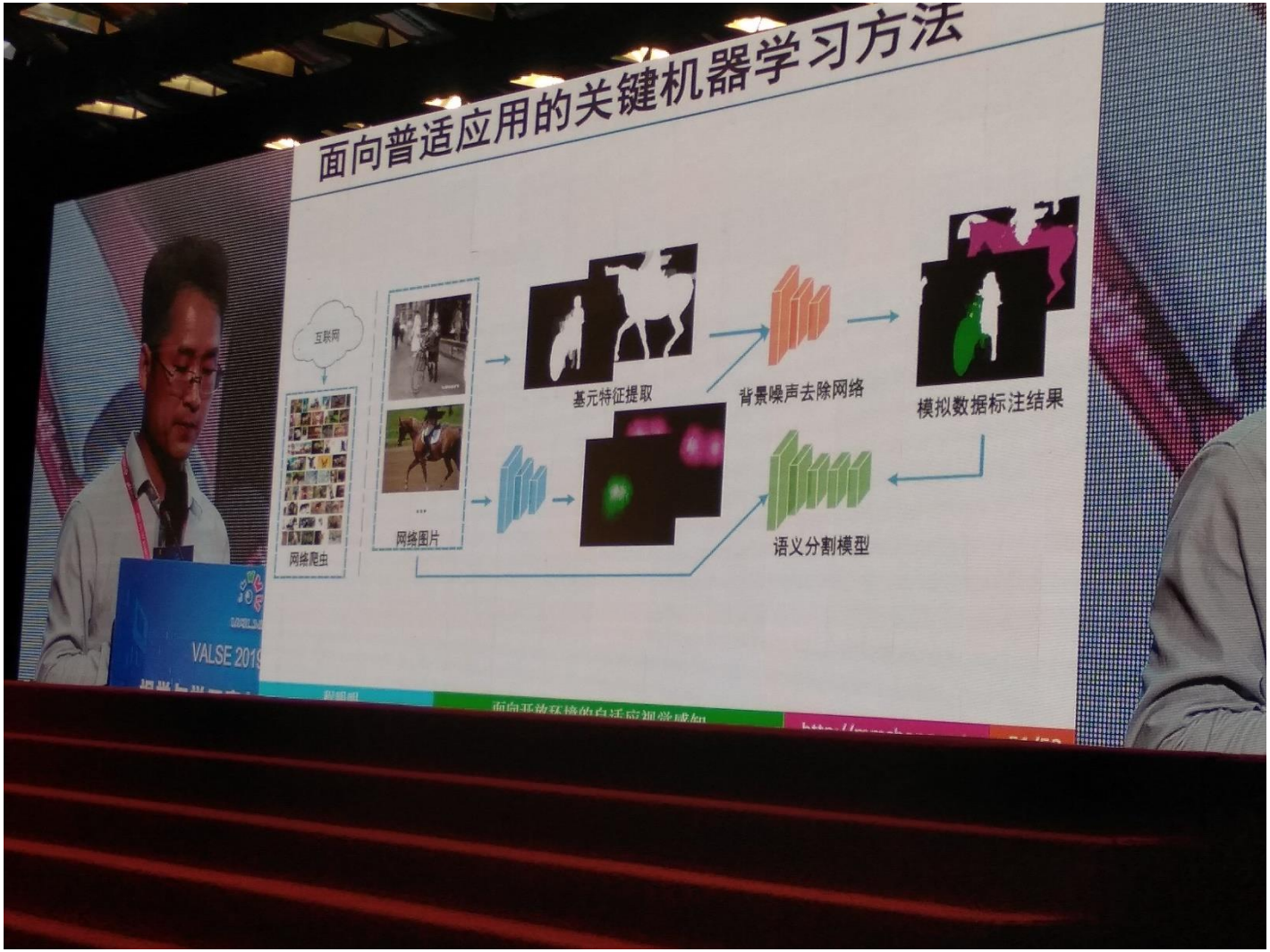
Weakly | Active Learning

- Motivation:
  - 当前各种深度网络的进步得益于网络**多尺度**信息综合能力的提升
- 报告核心：
  - **富尺度**空间神经网络架构：多任务协同求解，鲁棒性提高
  - 显著性物体检测：**预设基元属性**感知能力，减少数据依赖
  - 互联网大数据**自主**学习：减少人工标注，自动学习

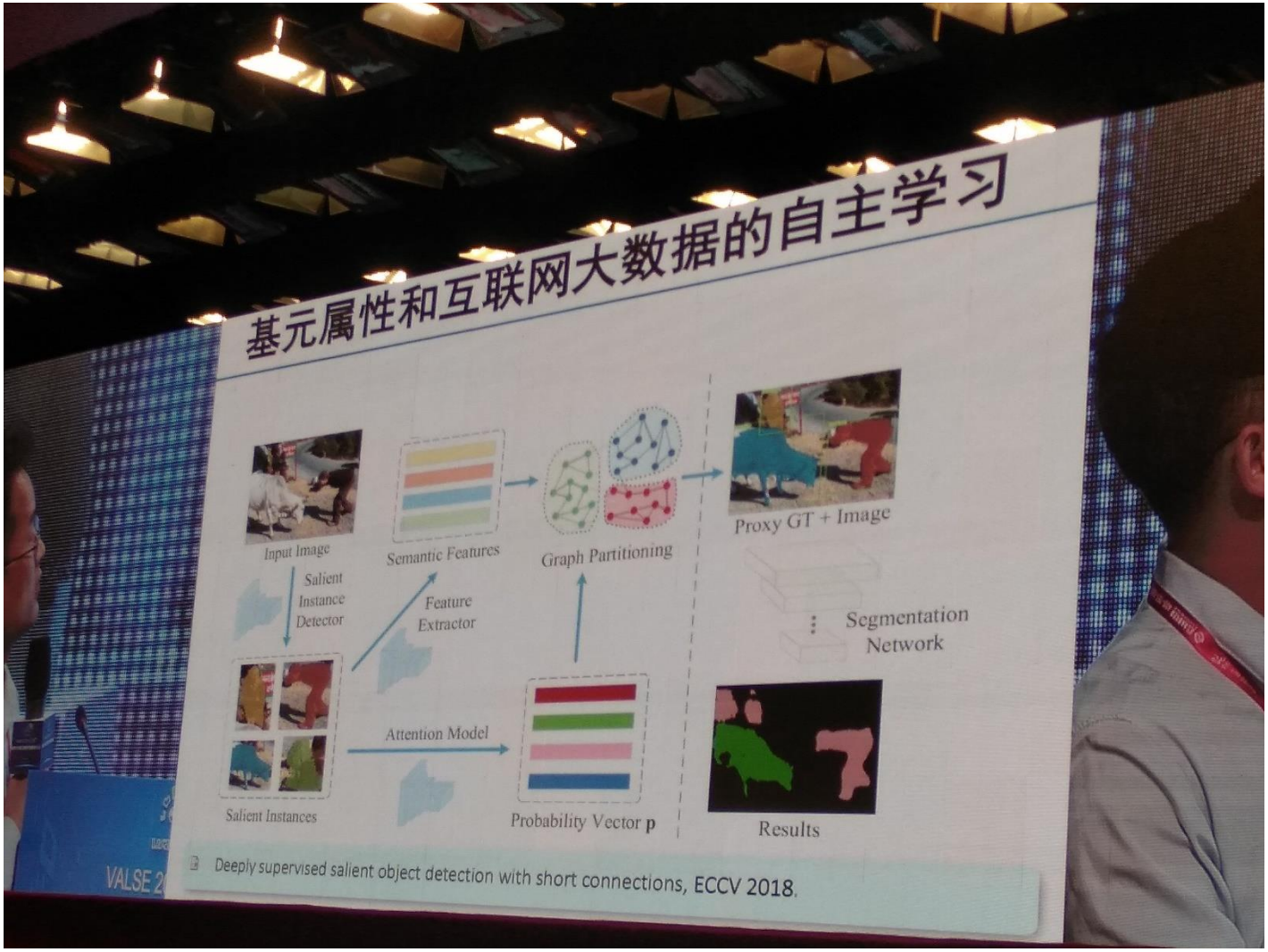
Weakly | Active Learning



Weakly | Active Learning



Weakly | Active Learning



Weakly | Active Learning

Weakly | Active Learning

- What Weakly ?
- Mingming Chen, NKU
  - Multi-scale Backbone Architecture
  - 通用视觉基元属性感知
  - 互联网大数据自主学习
- Qixiang Ye, CAS
  - 粗粒度的弱监督标记
  - Towards Self-Learning

Weakly | Active Learning

- What Weakly ?
- Mingming Chen, NKU
  - Multi-scale Backbone Architecture
  - 通用视觉基元属性感知
  - 互联网大数据自主学习
- Qixiang Ye, CAS
  - 粗粒度的弱监督标记
  - Towards Self-Learning

Weakly | Active Learning

- Motivation:
  - 目前detection / segmentation需要为每一个任务制定详细而具体的标注，而且无法共享，故而**标注成本极大**
- 核心观点：
  - 粗粒度的弱监督标记：比如，
    - 只给目标物体上画一条线，
    - 只在目标物体上打一个点，
    - 仅仅告诉模型一系列图片中包含什么而不给位置，让模型自己学习找到这些目标

Weakly | Active Learning

Weakly | Active Learning

Weakly | Active Learning

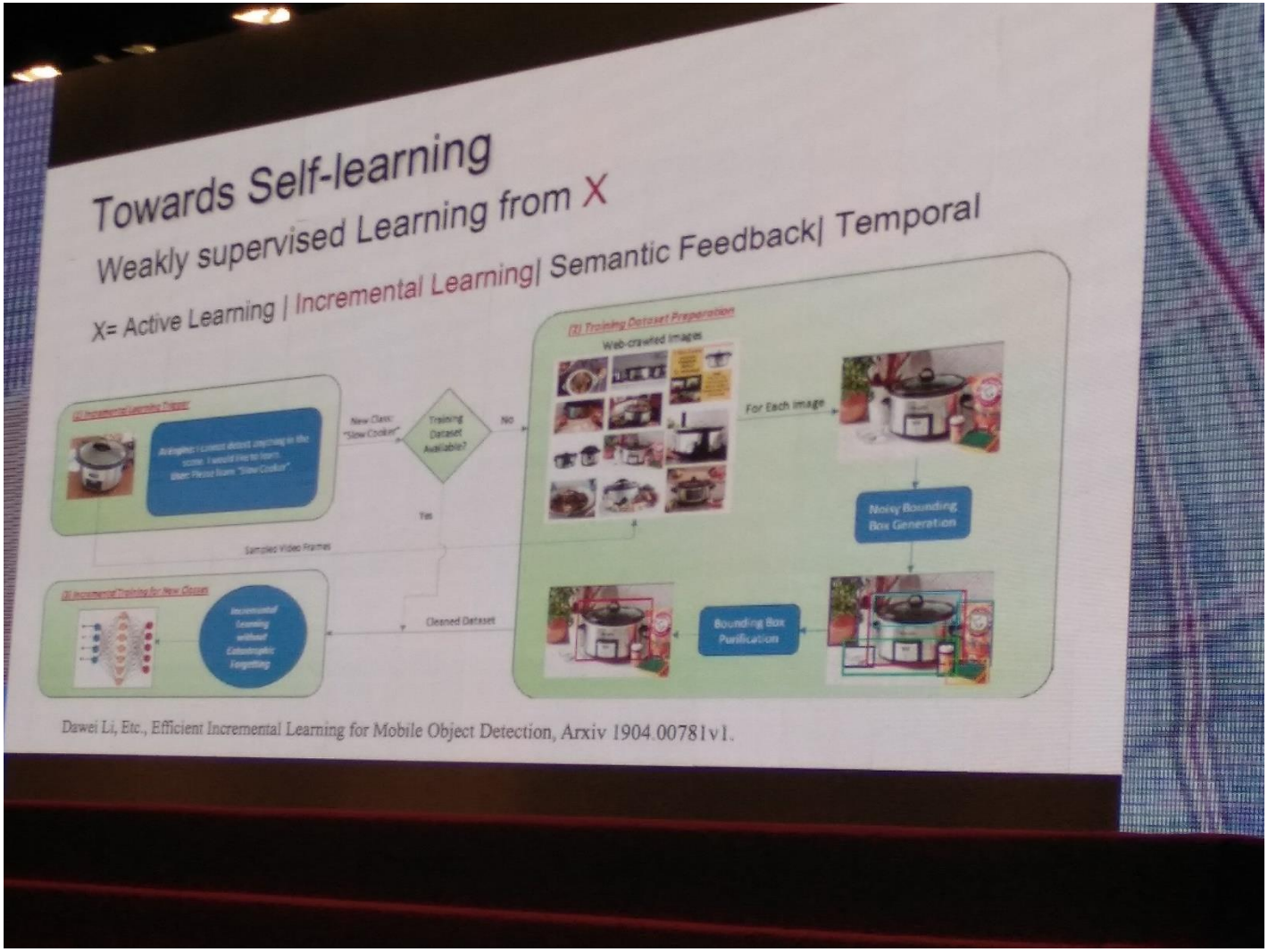
- Min-entropy Latent Model for Weakly Supervised object Detection, (CVPR2018)
- CMIL: Continuation Multiple Instance Learning for Weakly Supervised Detection (CVPR2019 **Oral**)
- SPN: Soft Proposal Network for Weakly Supervised Object Localization (ICCV2017)
- Learning Instance Activation Maps for Weakly Supervised Instance Segmentation (CVPR2019)
- PAMI2019: Recurrent Learning(MELM+RecurrentLearning)
- PeakResponseMapping(PRM) (CVPR2018)

Weakly | Active Learning

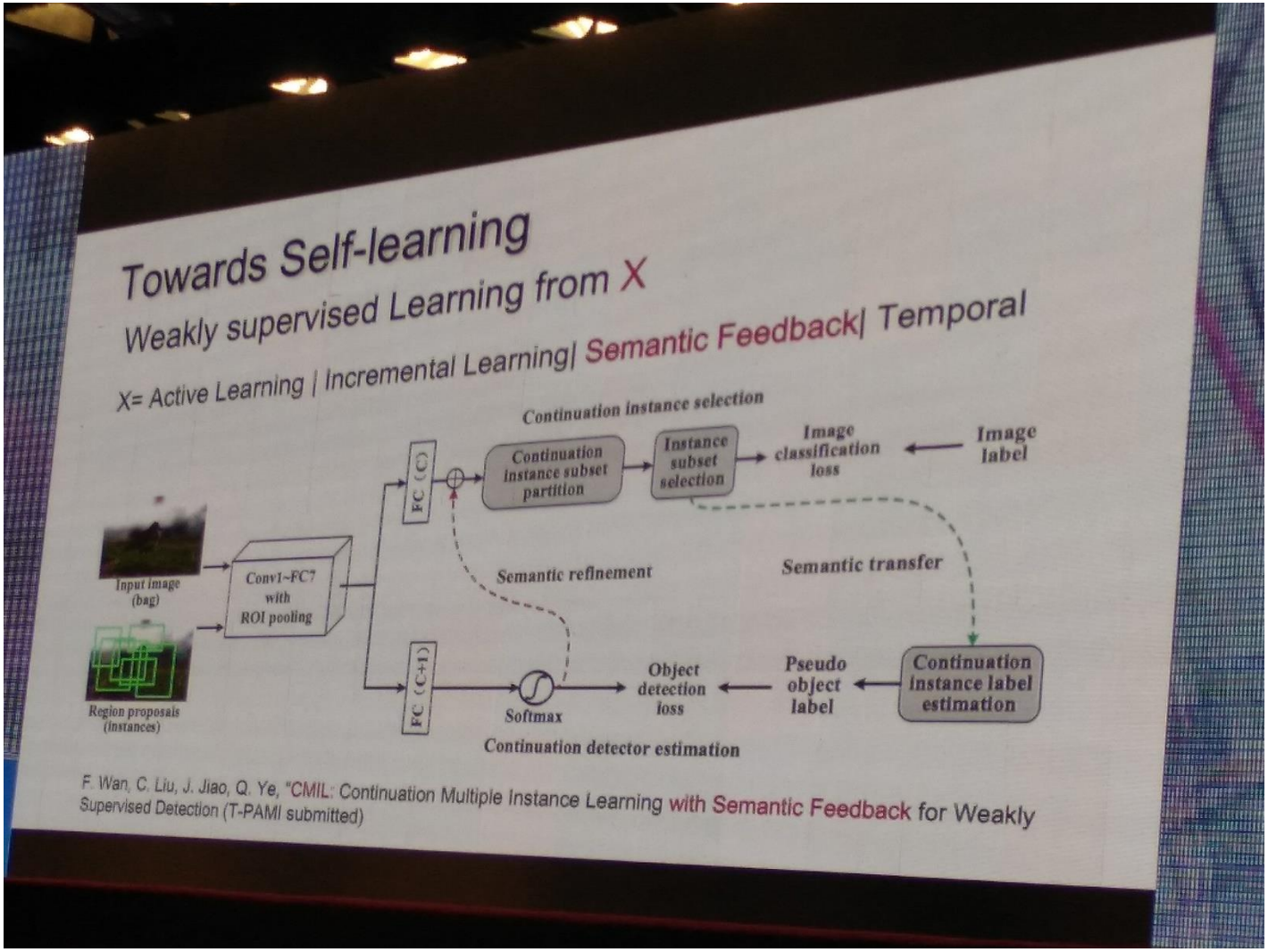
- What Weakly ?
- Mingming Chen, NKU
  - Multi-scale Backbone Architecture
  - 通用视觉基元属性感知
  - 互联网大数据自主学习
- Qixiang Ye, CAS
  - 粗粒度的弱监督标记
  - Towards Self-Learning

Weakly | Active Learning

Weakly | Active Learning



Weakly | Active Learning



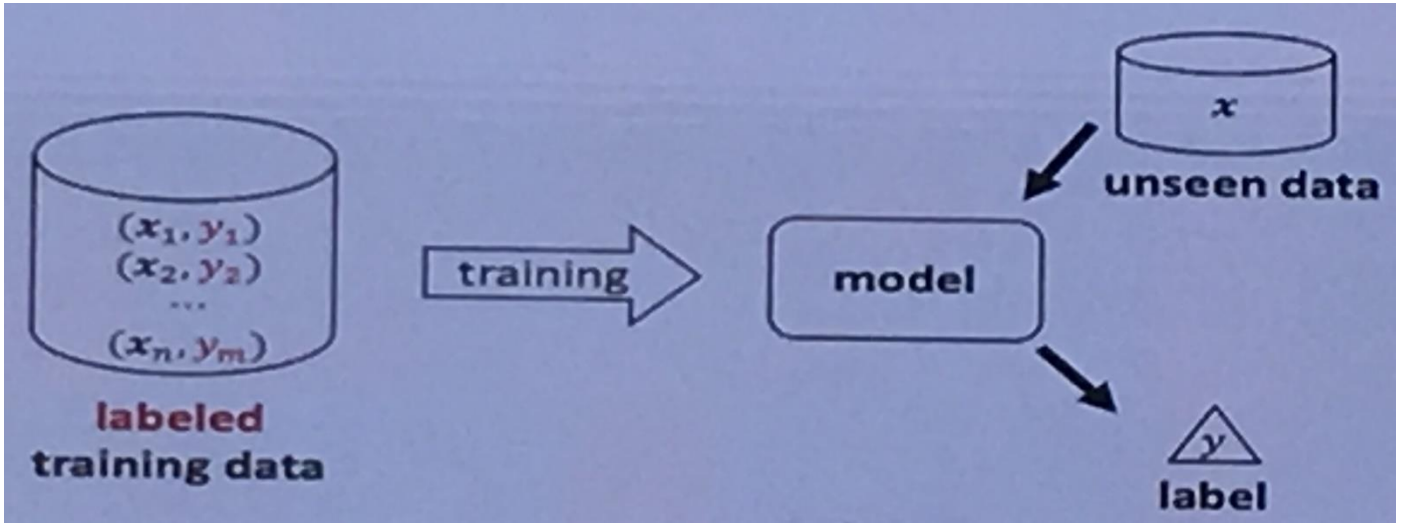
Weakly | Active Learning

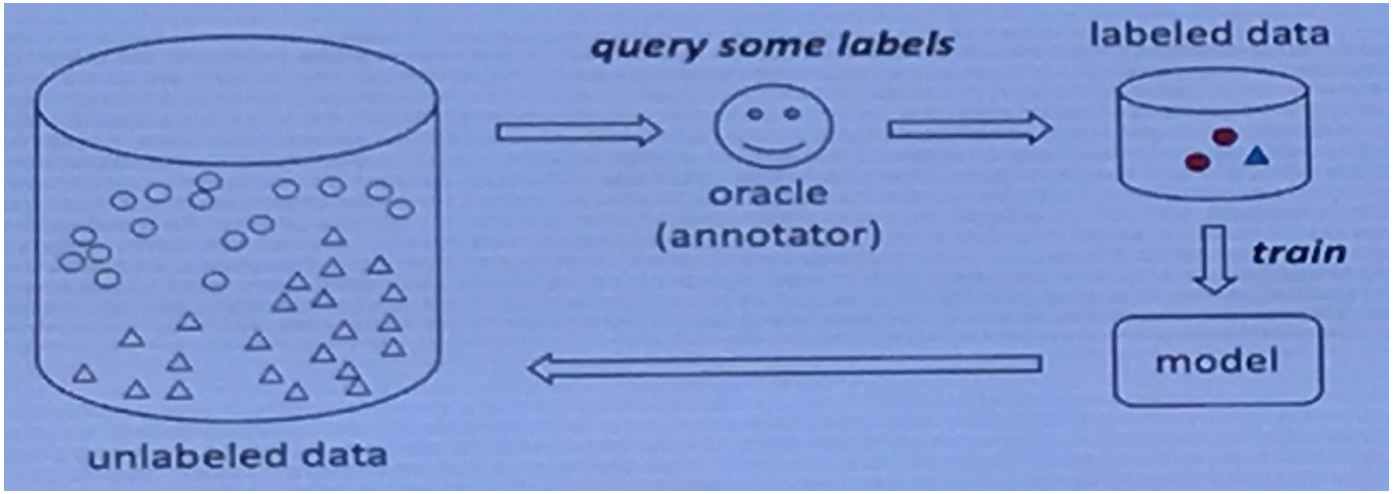
- Shengjun Huang, NHU
  - Active Learning
- Summary (个人观点，真伪自辩)

Weakly | Active Learning

- Motivation:
  - 大量的数据标记成本巨大，有一些甚至是不可得的
    - 比如医院的患者信息，异常检测的异常样本，几年才发生一次异常
  - 标记成本的判定非常复杂：**“Label 生而不平等”**
- Solution:
  - Active Learning：用最小的标记代价获得最大的performance
- Non-Active vs. Active Learning
  - 如下图

Weakly | Active Learning

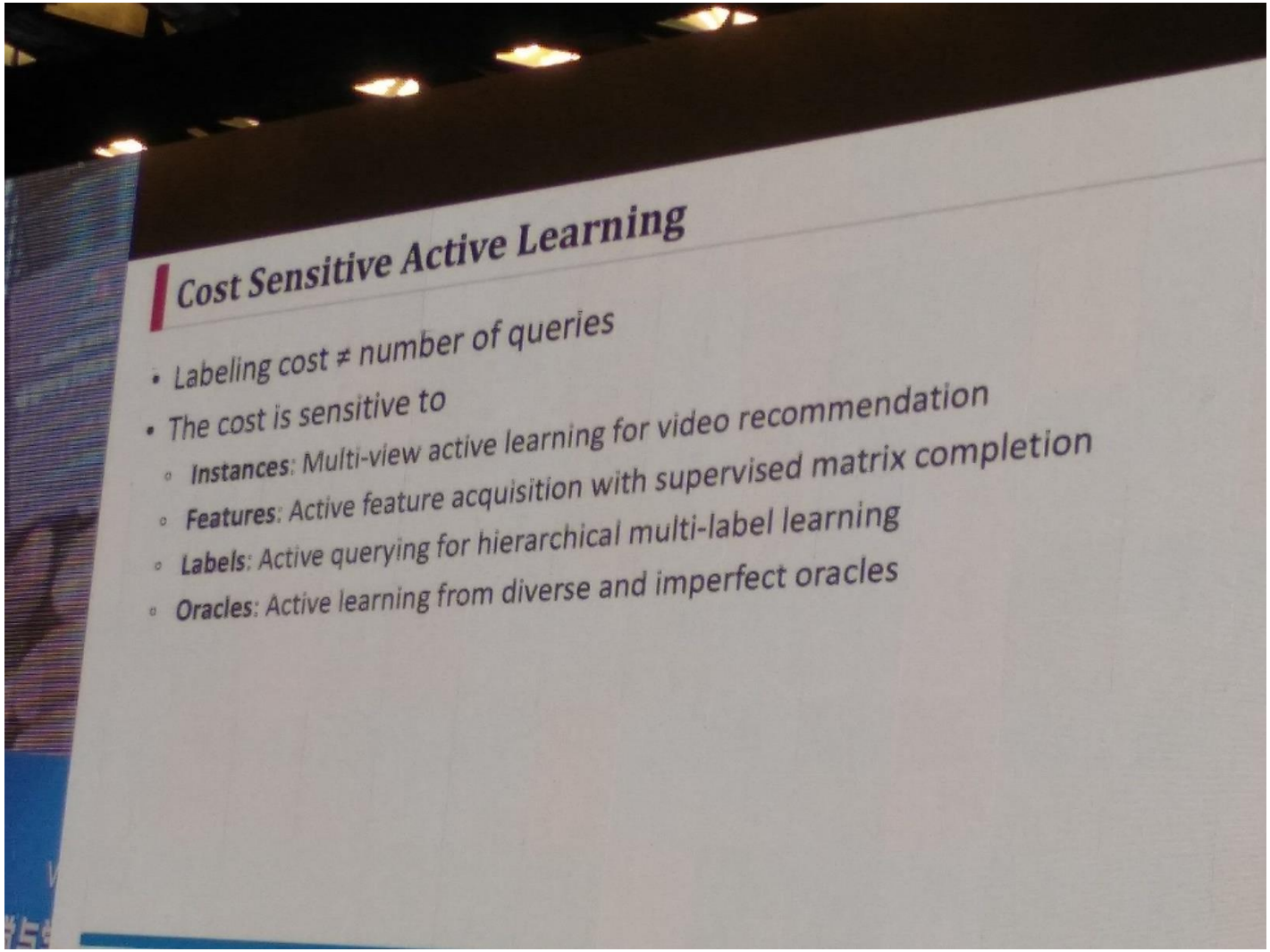




Weakly | Active Learning

- **Cost Sensitive** Active Learning
  - **goal**: train an effective model with **least** labeling cost
  - 细致定义 **least** 就是 Active Learning 的 核心
- The cost is not number of queries (too naïve too young),
  - It is about…
  - E.g. Multi-iew Active learning for Video Recommendation

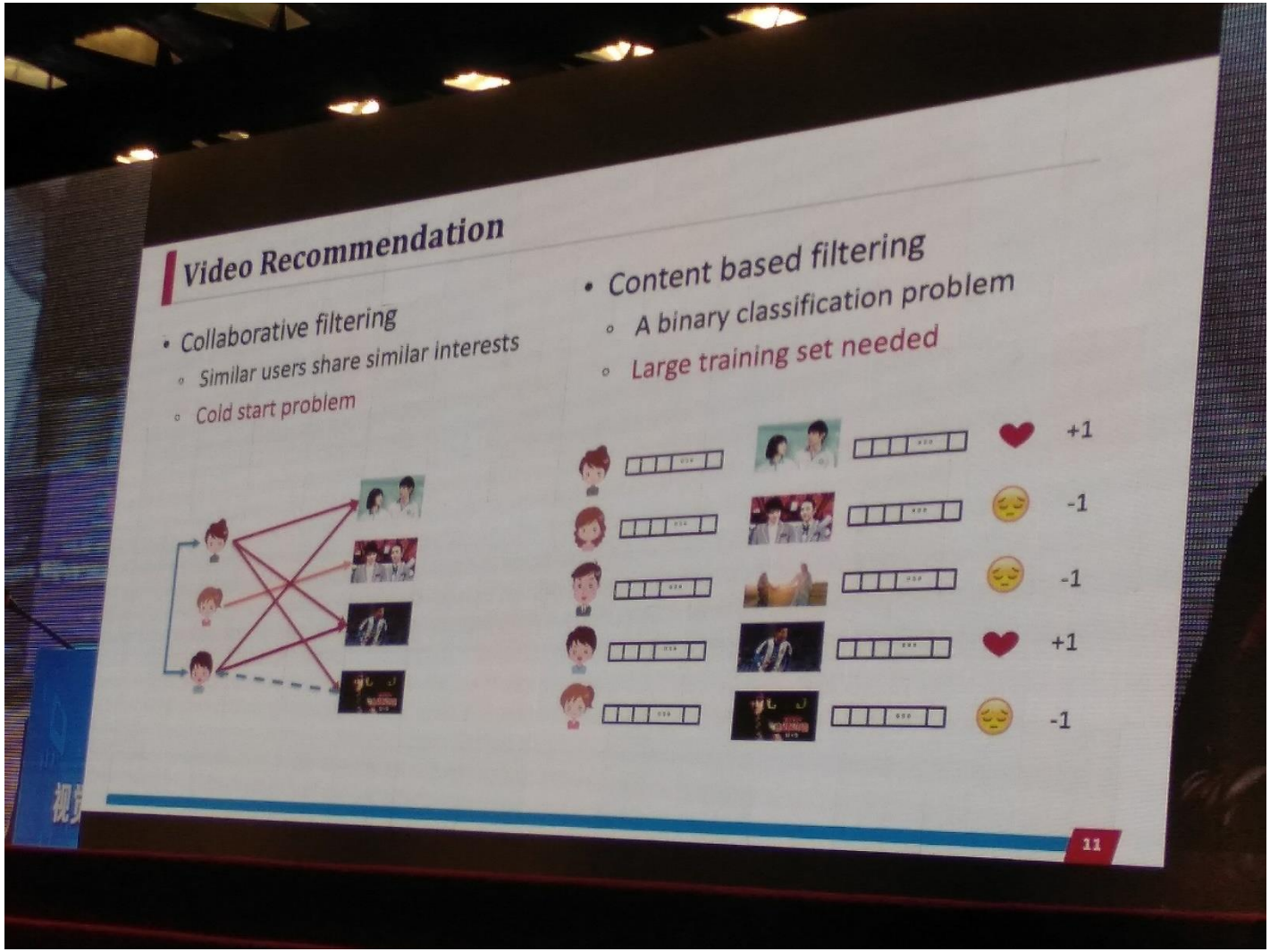
Weakly | Active Learning



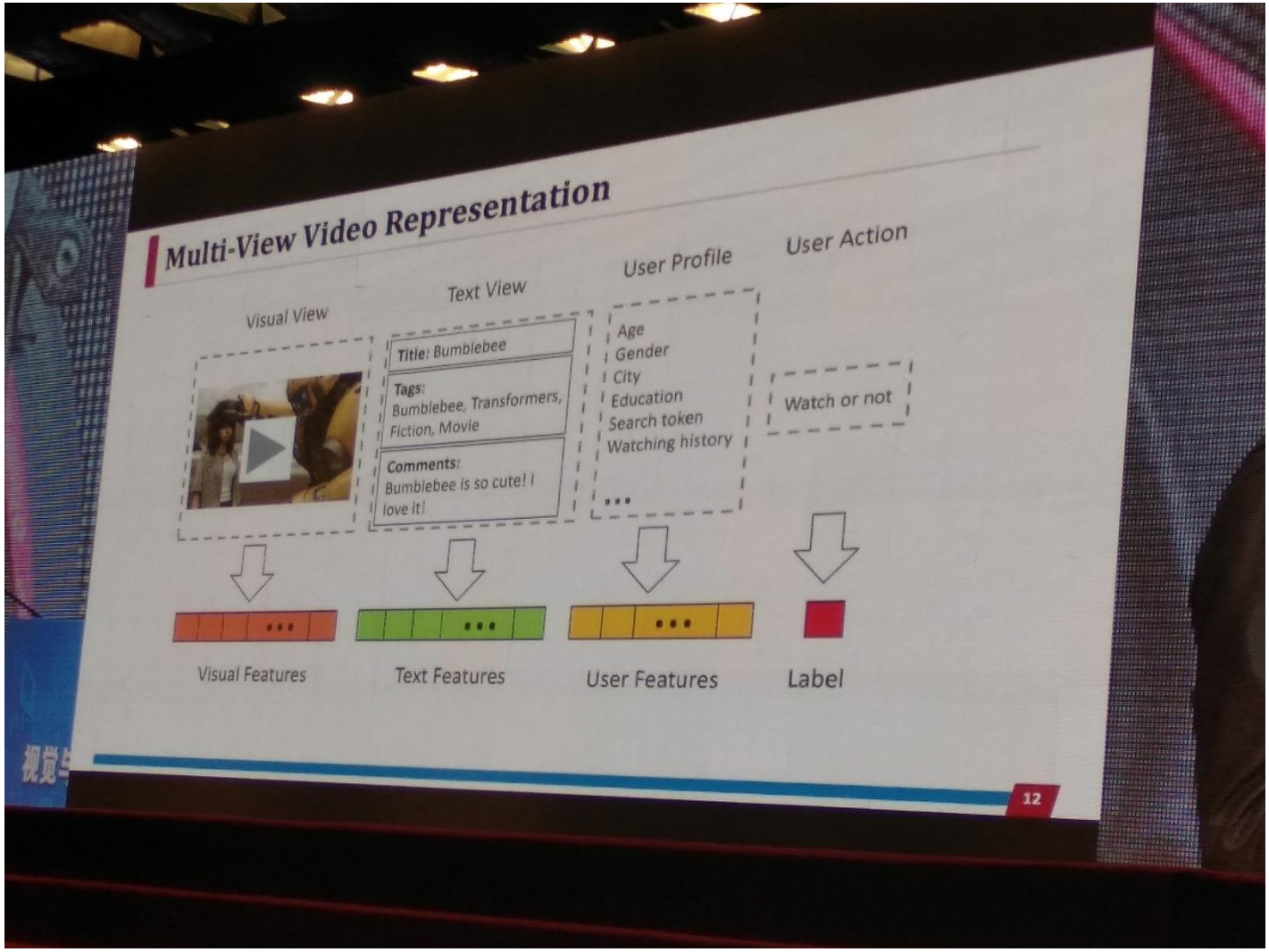
Weakly | Active Learning

- Instances
  - 核心：怎么选最好的instance？
- 视频推荐：
  - 协同过滤（冷门启动问题）、基于内容的过滤（需要大量数据训练）
- 多视角视频表示：
  - 视觉特征、文本特征、用户特征、标签
- Motivation：
  - 在视频推荐任务中，文本特征（即评论）获取需要很大代价，视觉特征不需要人力代价。
- Idea： Visual to text Mapping

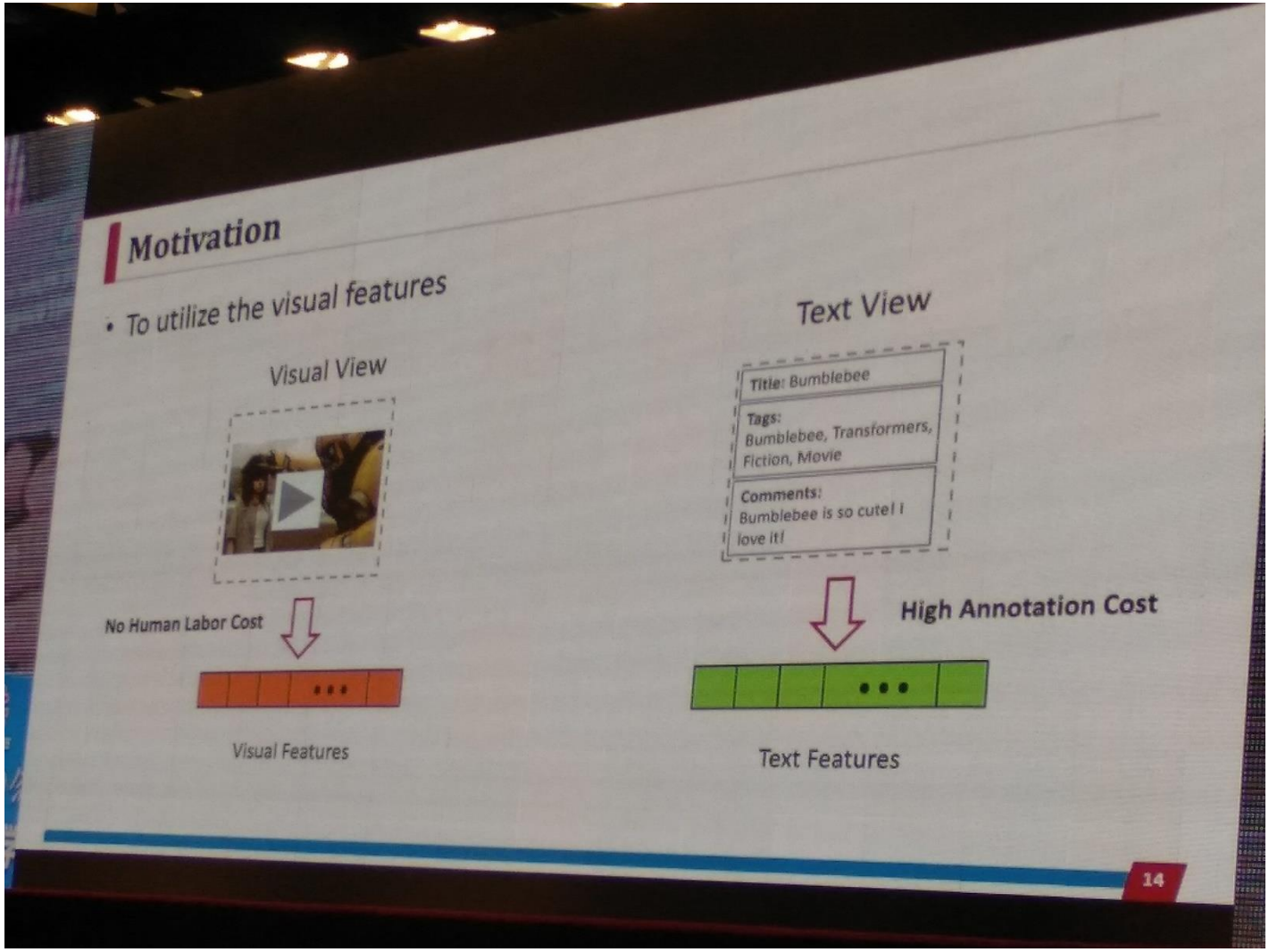
Weakly | Active Learning



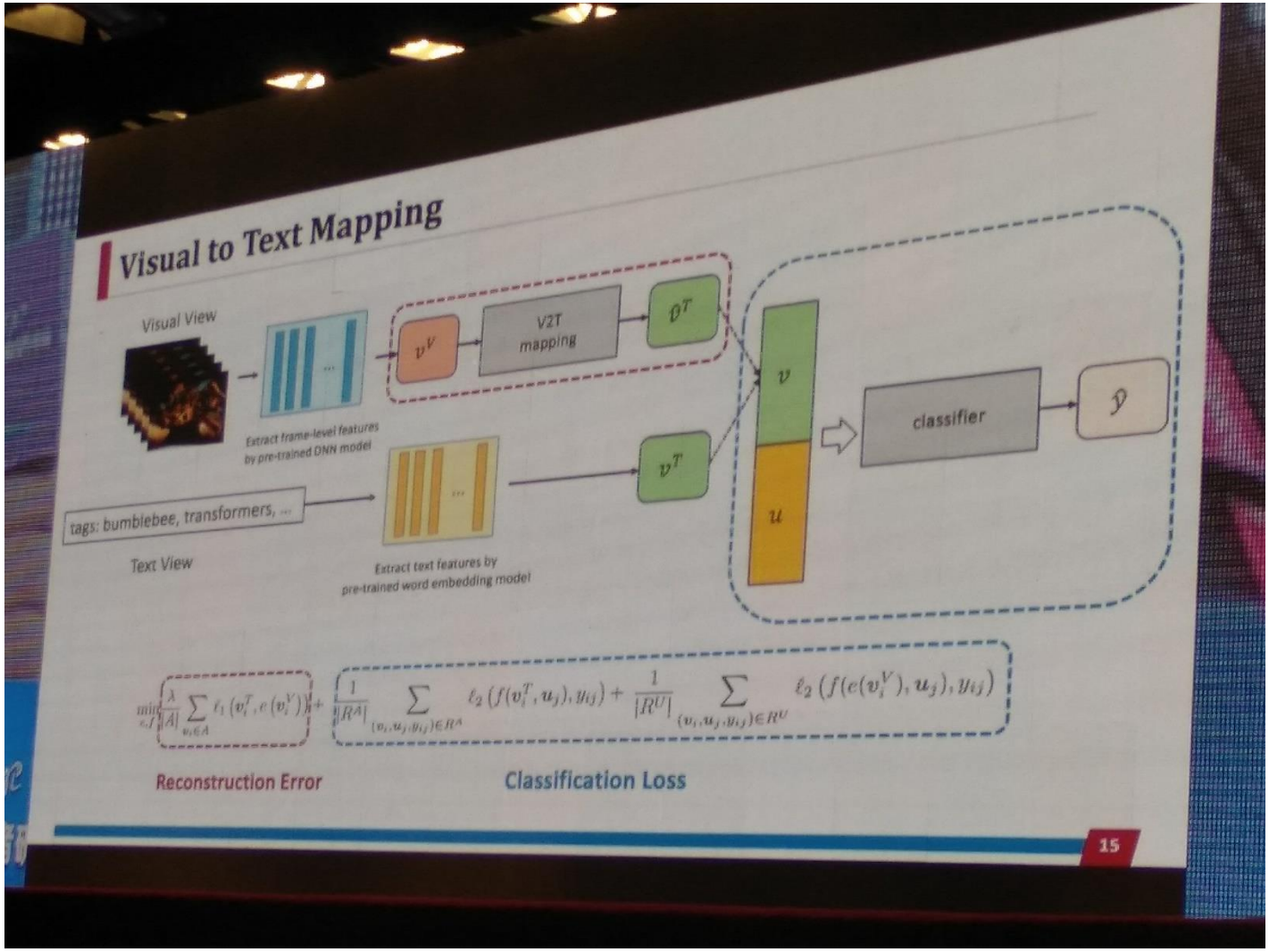
Weakly | Active Learning



Weakly | Active Learning



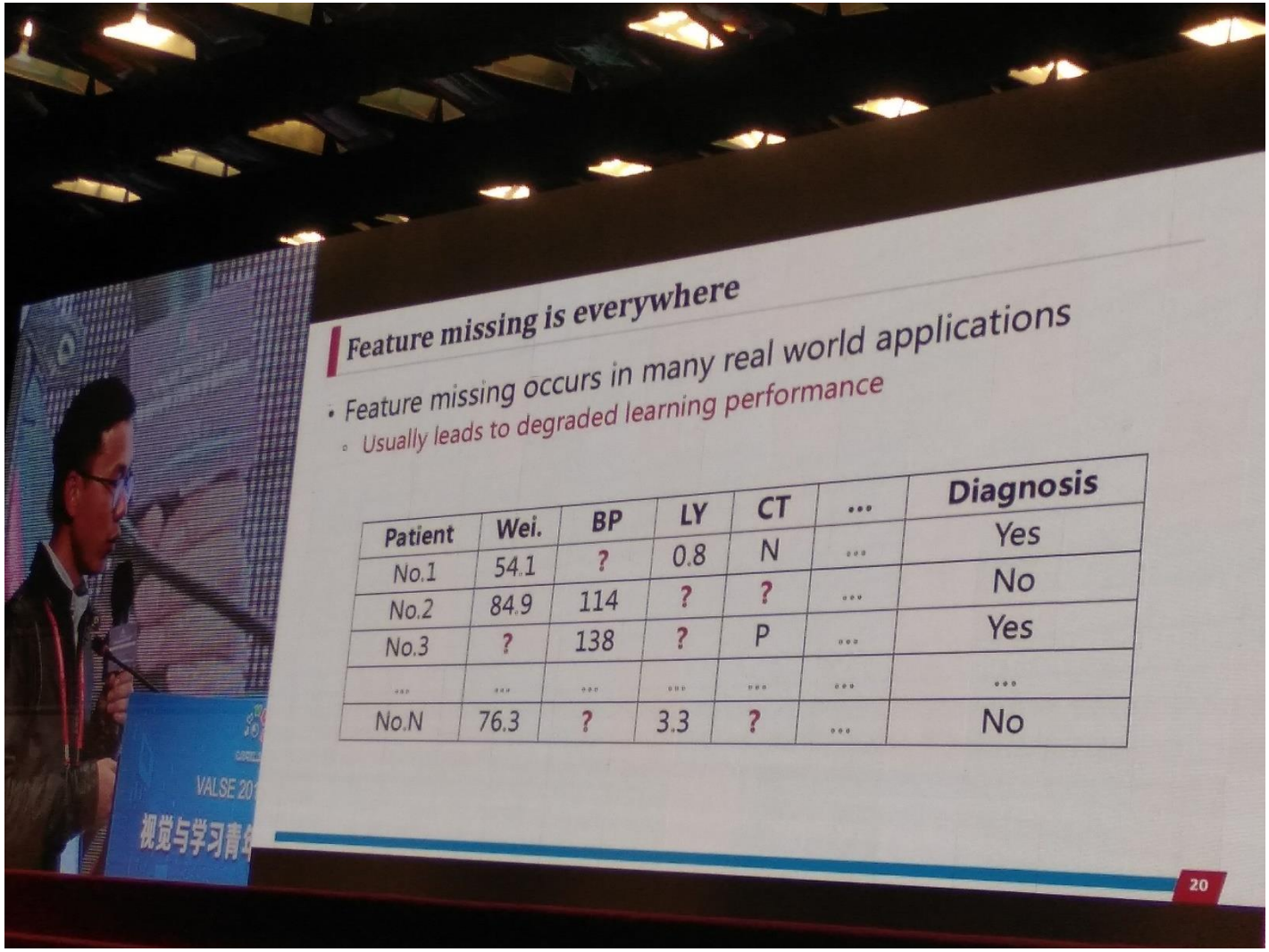
Weakly | Active Learning



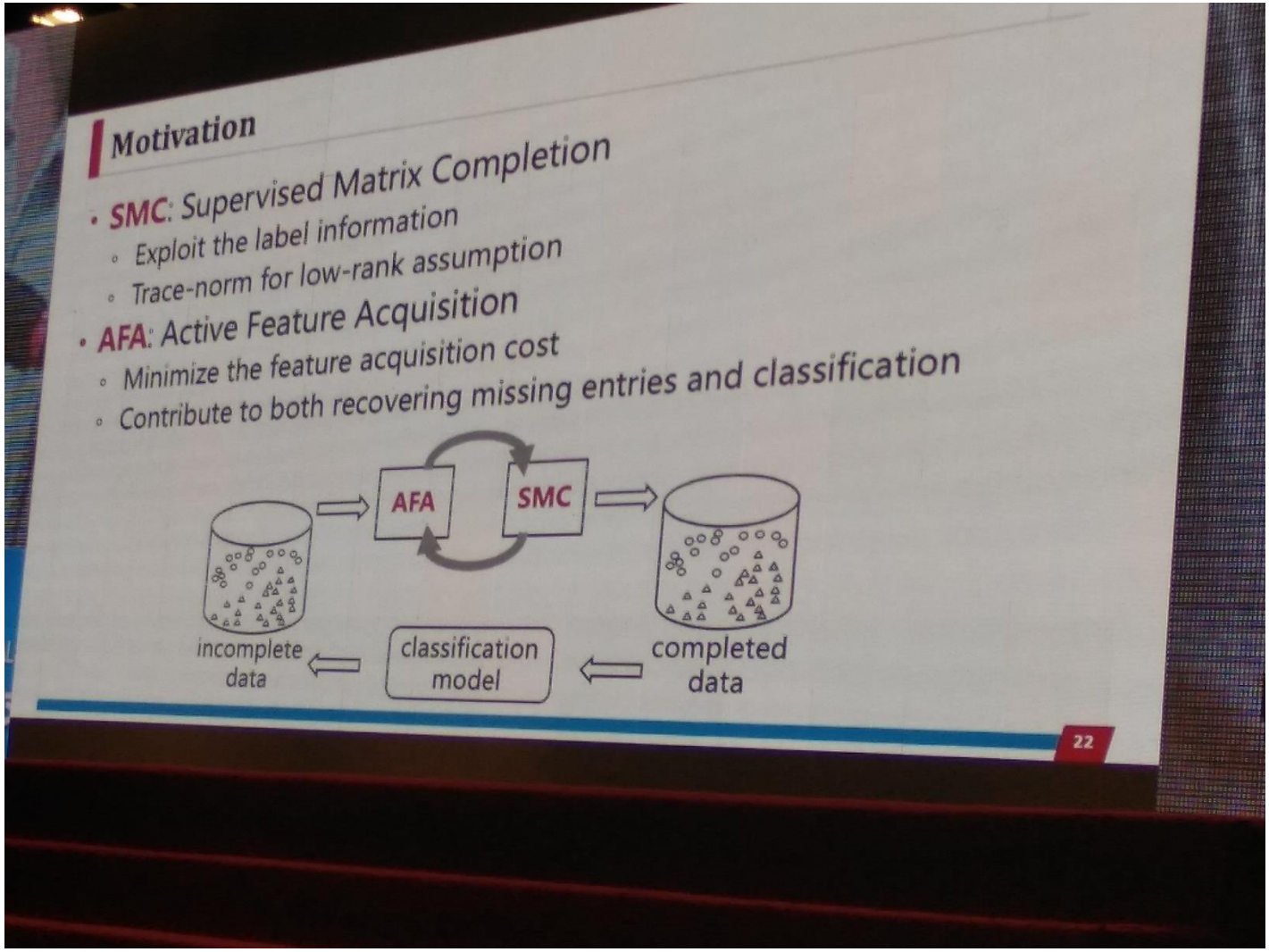
Weakly | Active Learning

- Features
  - 问题：现实应用中往往会出现特征丢失现象，通常导致学习性能下降
  - 核心：如何选更好的特征？

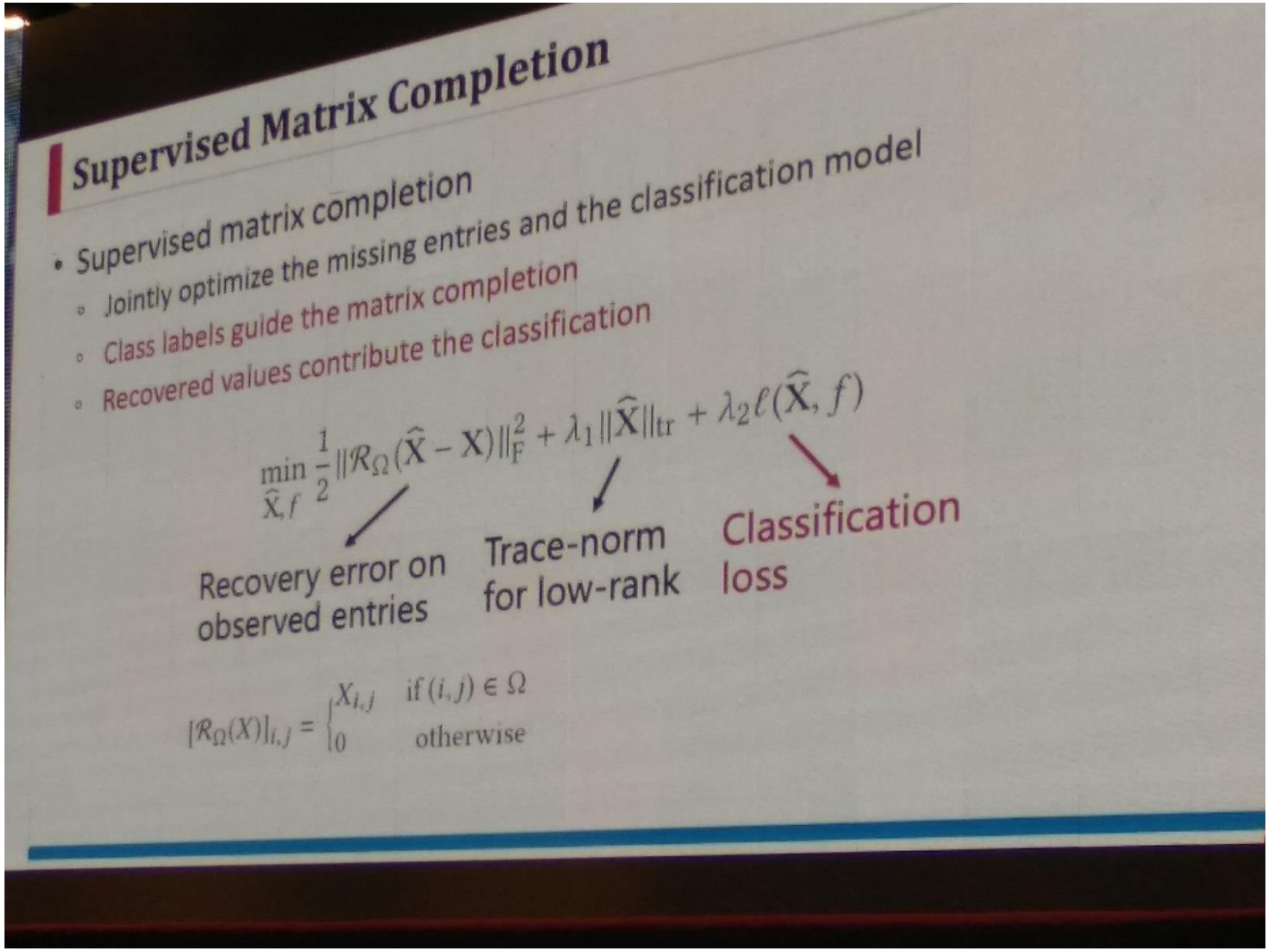
Weakly | Active Learning



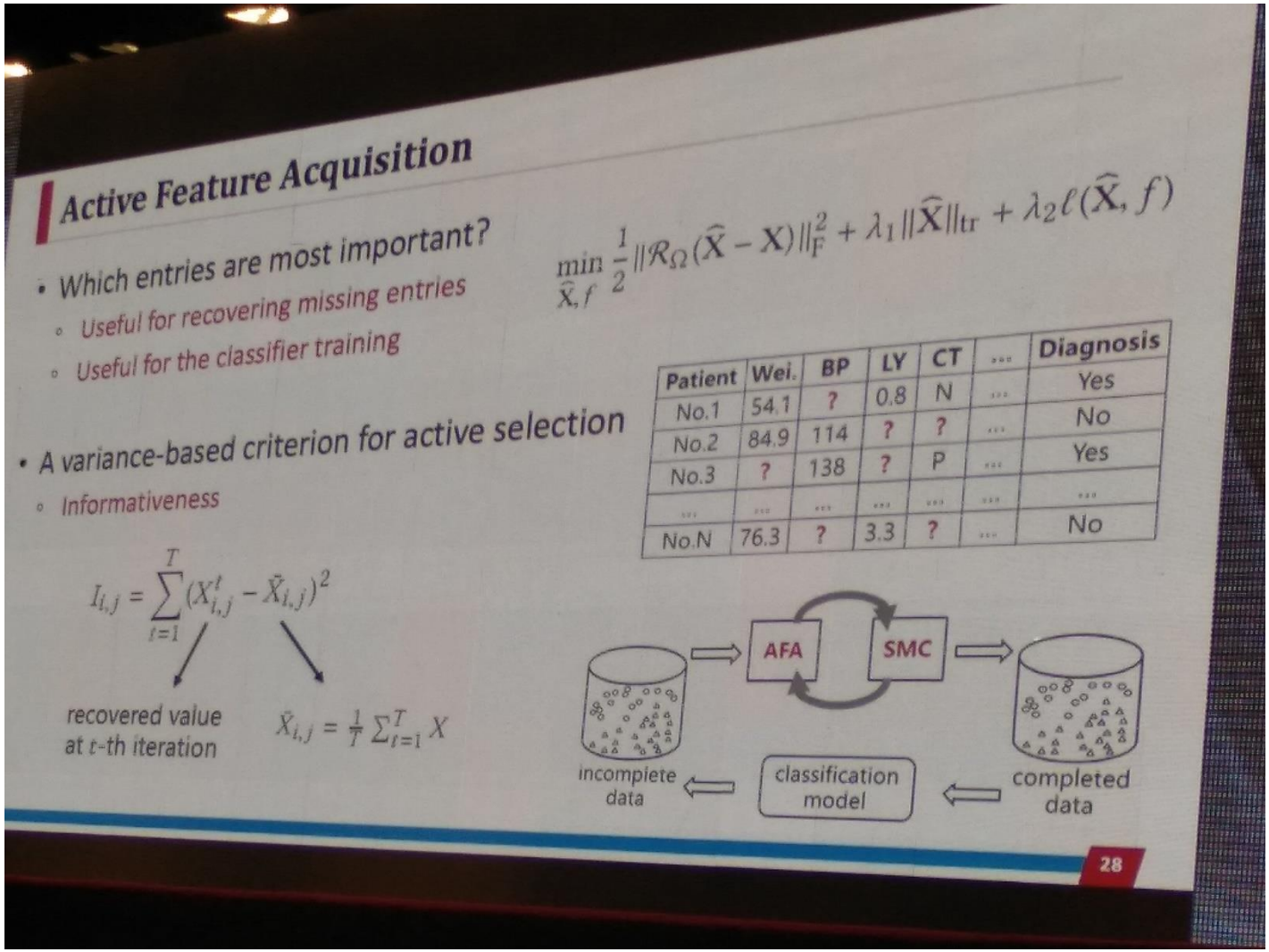
Weakly | Active Learning



Weakly | Active Learning



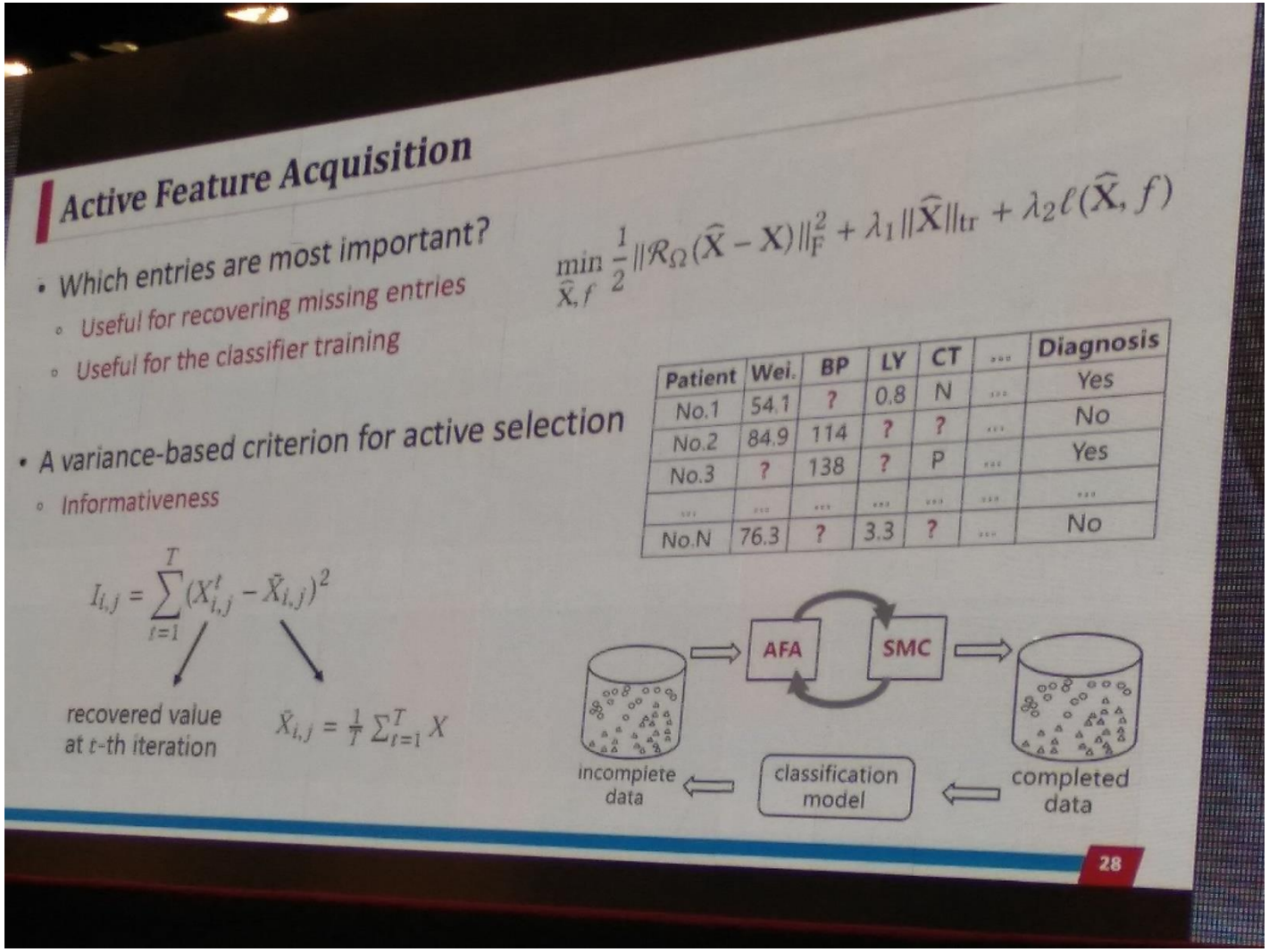
Weakly | Active Learning



Weakly | Active Learning

- Labels
  - 标签有层次结构，要平衡成本和信息

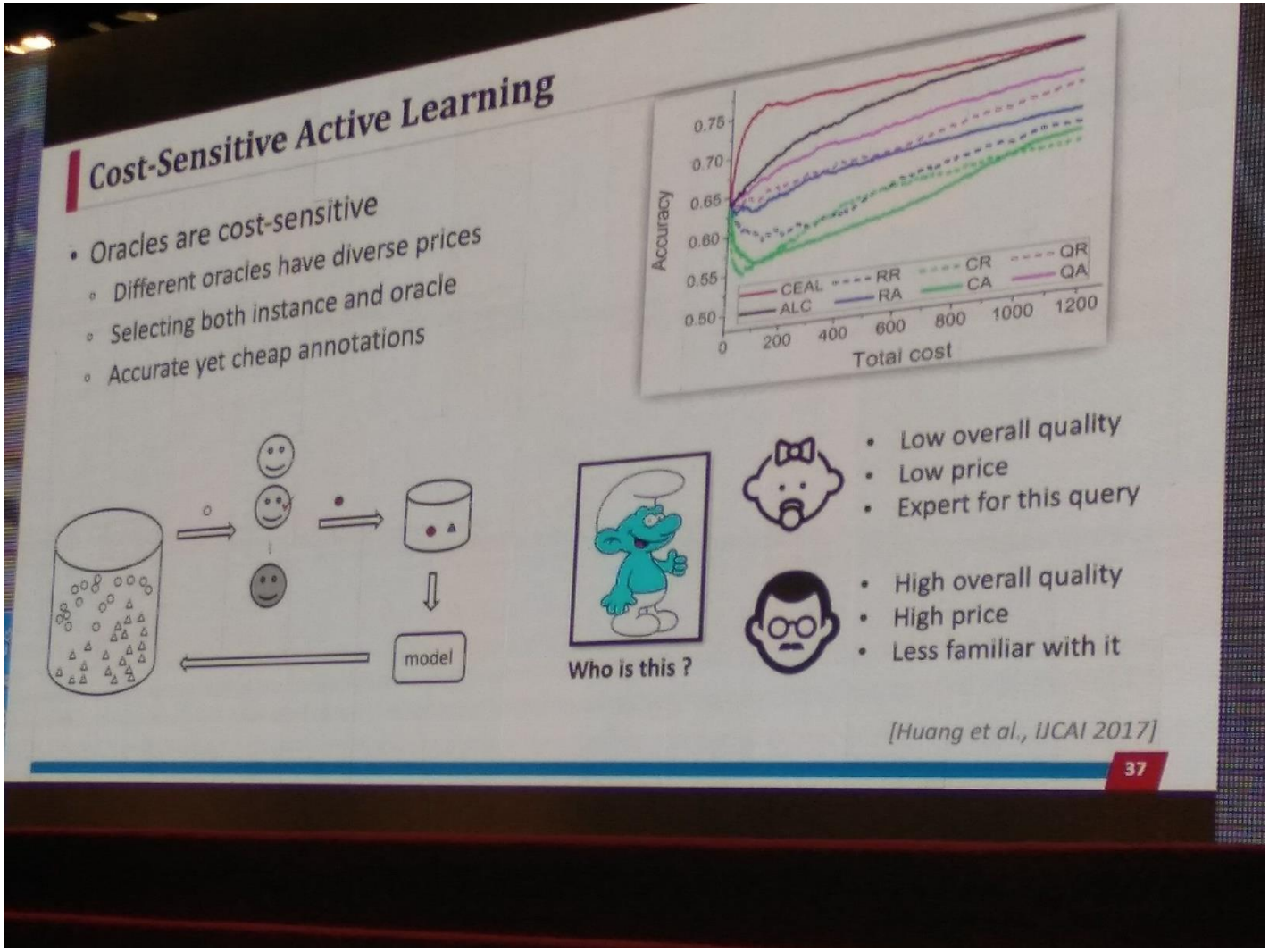
Weakly | Active Learning



Weakly | Active Learning

- Oracles
  - 不同的oracles有不同的价格

Weakly | Active Learning

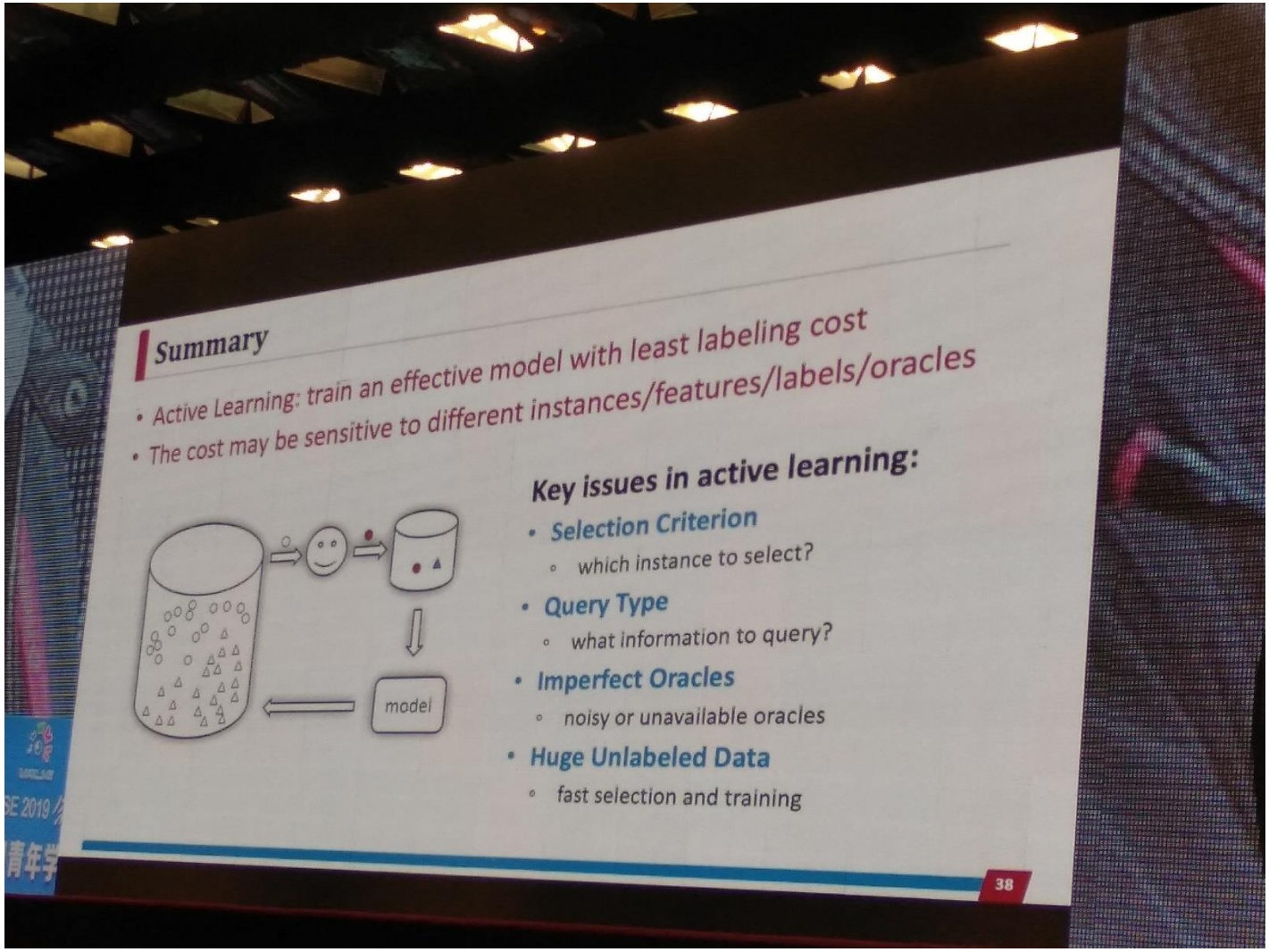


Outline

http://dmirlab.com

- About VALSE
- 弱监督视觉理解与主动学习
- 以人为中心的视觉理解
- 视觉中的知识推理
- CV+NLP 2.0
- Poster 选讲

Weakly | Active Learning



Weakly | Active Learning

- Shengjun Huang, NHU
  - Active Learning
- Summary (个人观点，真伪自辩)

Weakly | Active Learning

- 陈明明
  - **基元属性感知**：一种**特殊**的特征共享，挖掘各任务间的共享性，减少重复数据标记
- 叶齐祥：
  - **粗粒度标记**：只给目标物体上画一条线、打一个点、图像级别的标记，让模型自动学习完整的Label

Weakly | Active Learning

- Idea:
  - 卷积激励的**显著分布估计**: 对使原图产生激活大的区域进行分析
  - GAN生成边界周围的样本：得到**更细粒度的分类边界**
  - **多模态数据对齐**: 为了衡量图像和文本的相似度，将图像编码到文本的特征空间中，或者将文本编码到图像的特征空间中，以此衡量，**而不是**将二者编码到第三个特征空间中衡量
  - 设计**好的loss**是发表论文的好方法(尤其是人脸分析)

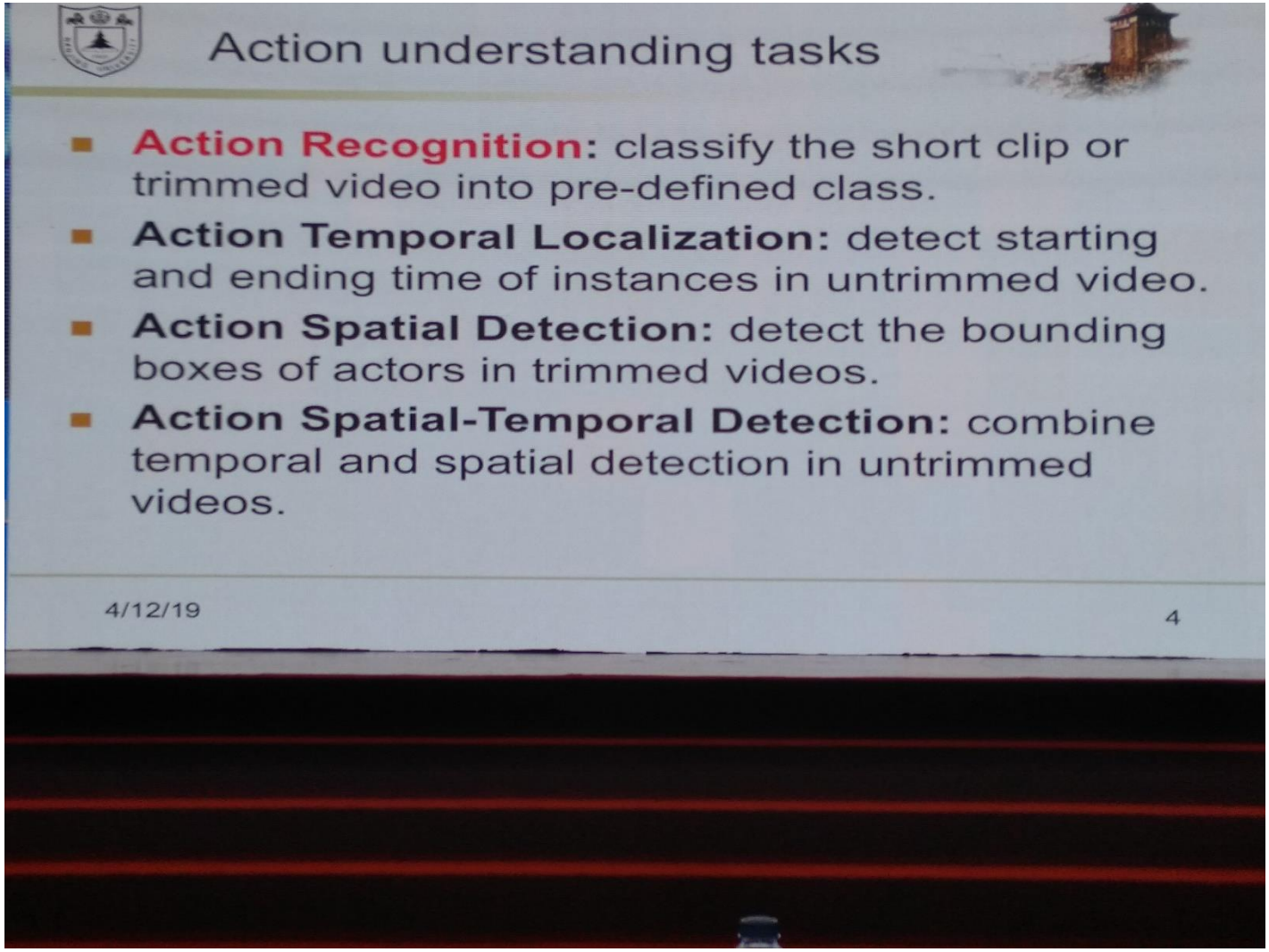
Outline

- About VALSE
- 弱监督视觉理解与主动学习
- 以人为中心的视觉理解
- 视觉中的知识推理
- CV+NLP 2.0
- Poster 选讲

Person

- 通用视频的时序建模与动作识别 (liming wang, NJU)
  - 通用视频的动作识别的深度网络全家福
  - Liming wang的三个工作

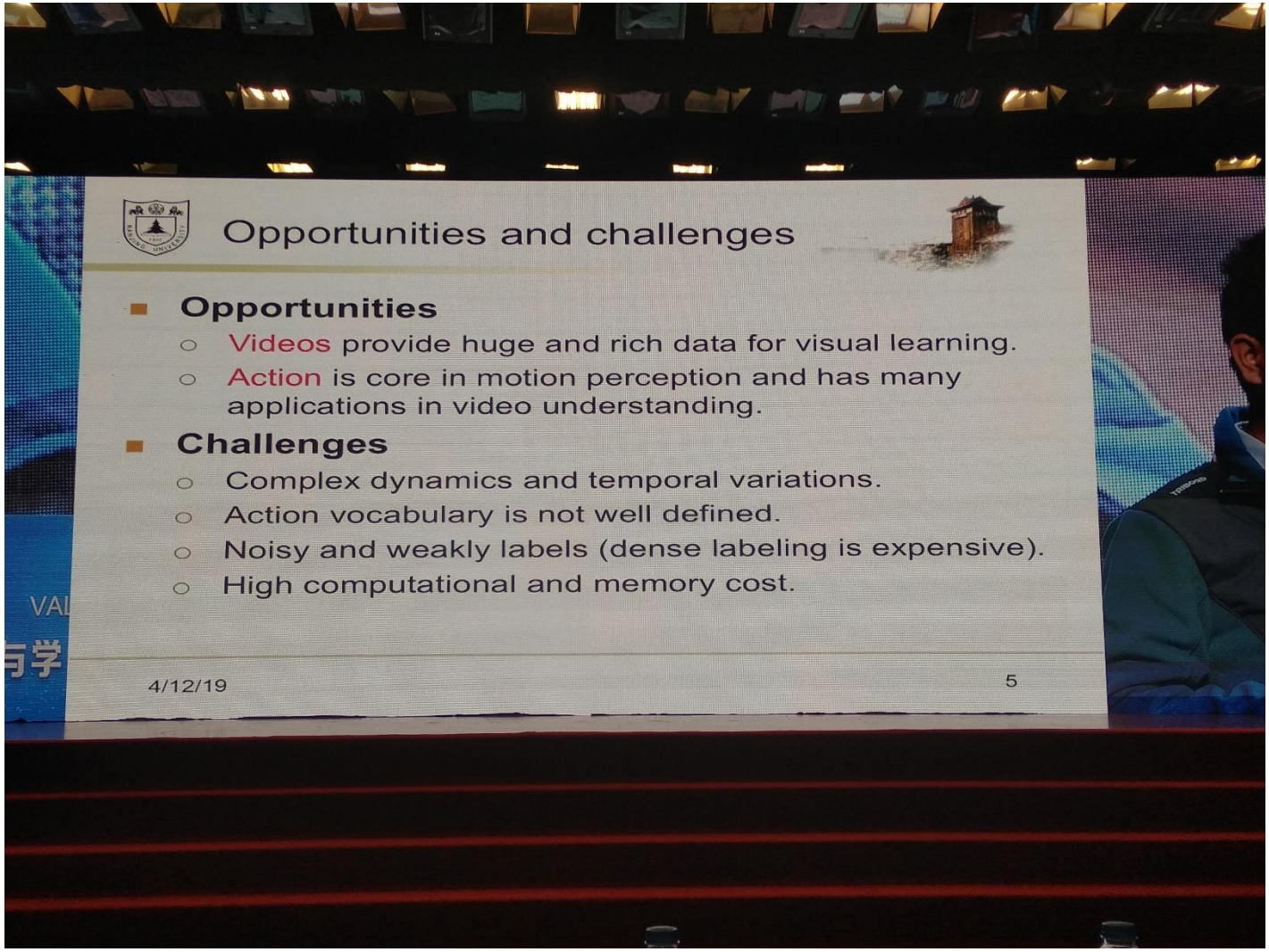
Person


Action understanding tasks
Action Recognition: classify the short clip or trimmed video into pre-defined class.
Action Temporal Localization: detect starting and ending time of instances in untrimmed video.
Action Spatial Detection: detect the bounding boxes of actors in trimmed videos.
Action Spatial-Temporal Detection: combine temporal and spatial detection in untrimmed videos.
4/12/19
4

Person


Action recognition in videos
1. Action recognition "in the lab": KTH, Weizmann etc.
2. Action recognition "in TV, Movies": UCF Sports, Holloywood etc.
3. Action recognition "in Web Videos": HMDB, UCF101, THUMOS, ActivityNet etc.
Haroon Idrees et al. The THUMOS Challenge on Action Recognition for Videos "in the Wild", in Computer Vision and Image Understan
(CVIU), 2017.
4/12/19
2

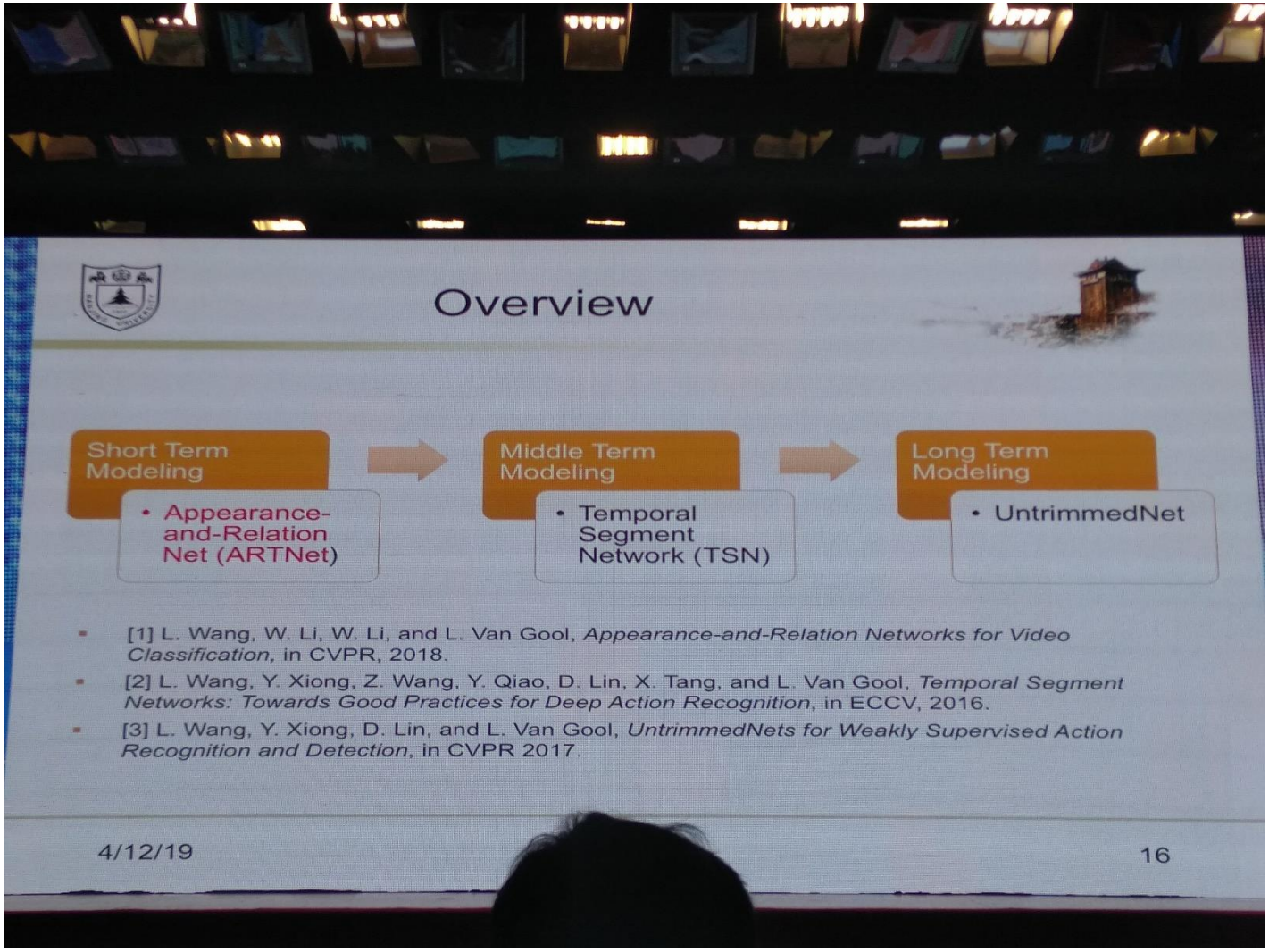
Person


Opportunities and challenges
Opportunities
Videos provide huge and rich data for visual learning.
Action is core in motion perception and has many applications in video understanding.
Challenges
Complex dynamics and temporal variations.
Action vocabulary is not well defined.
Noisy and weakly labels (dense labeling is expensive).
High computational and memory cost.
4/12/19
5

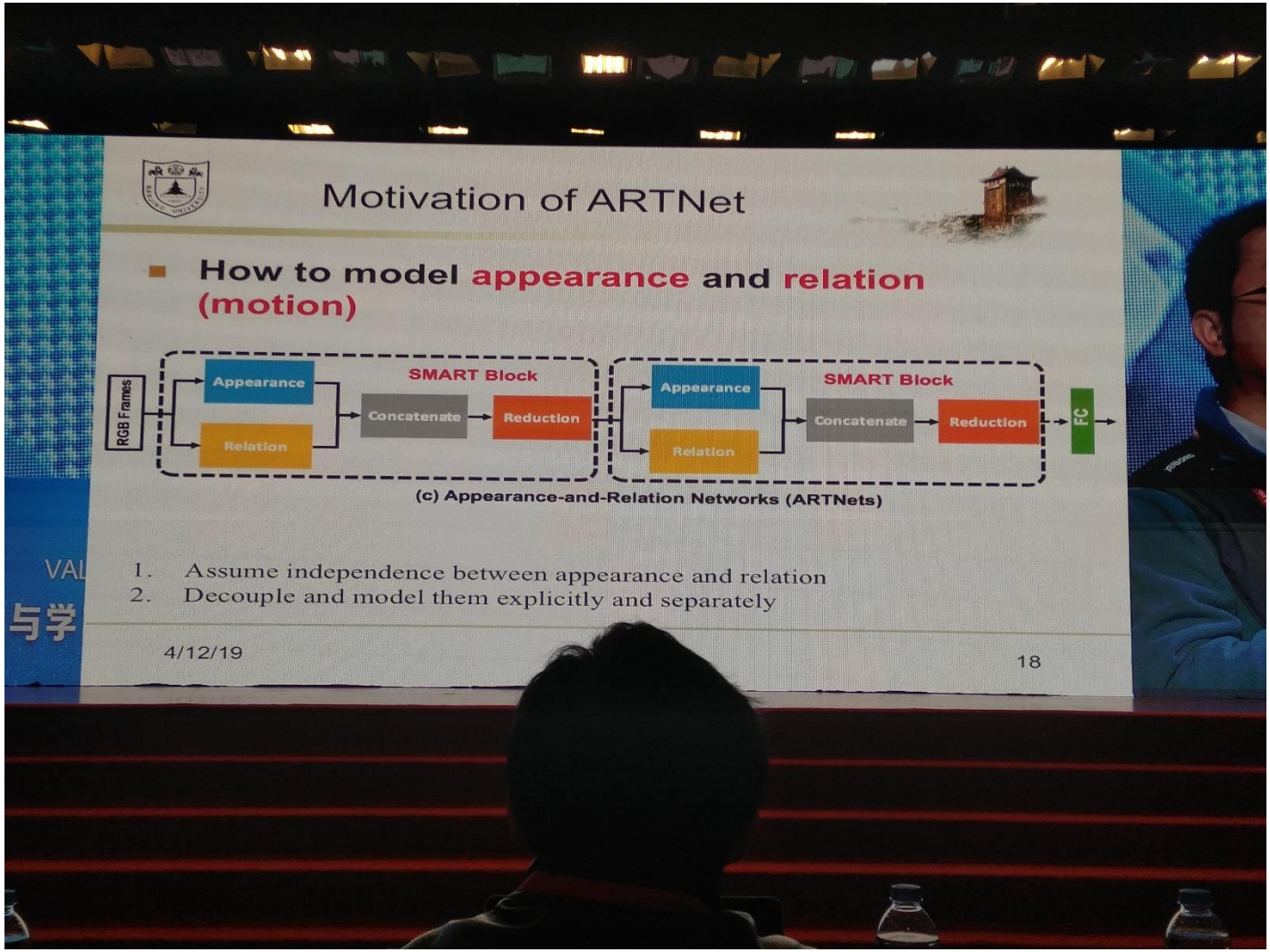
Person

- 代表性论文
  - Large-scale video classification with CNN (FeiFei Li, CVPR2014)
  - Two-Stream CNN for action recognition in videos (NIPS2014)
  - learning spatiotemporal features with 3D CNN (ICCV2015)
  - TDD (liming wang, CVPR2015)
  - Real-time action recognition with enhanced motion vector CNNs (CVPR2016)
  - Two Stream I3D (CVPR2017)
  - R(2+1)D (CVPR2018)
  - SlowFast Networks (kaiming he, CVPR2019)

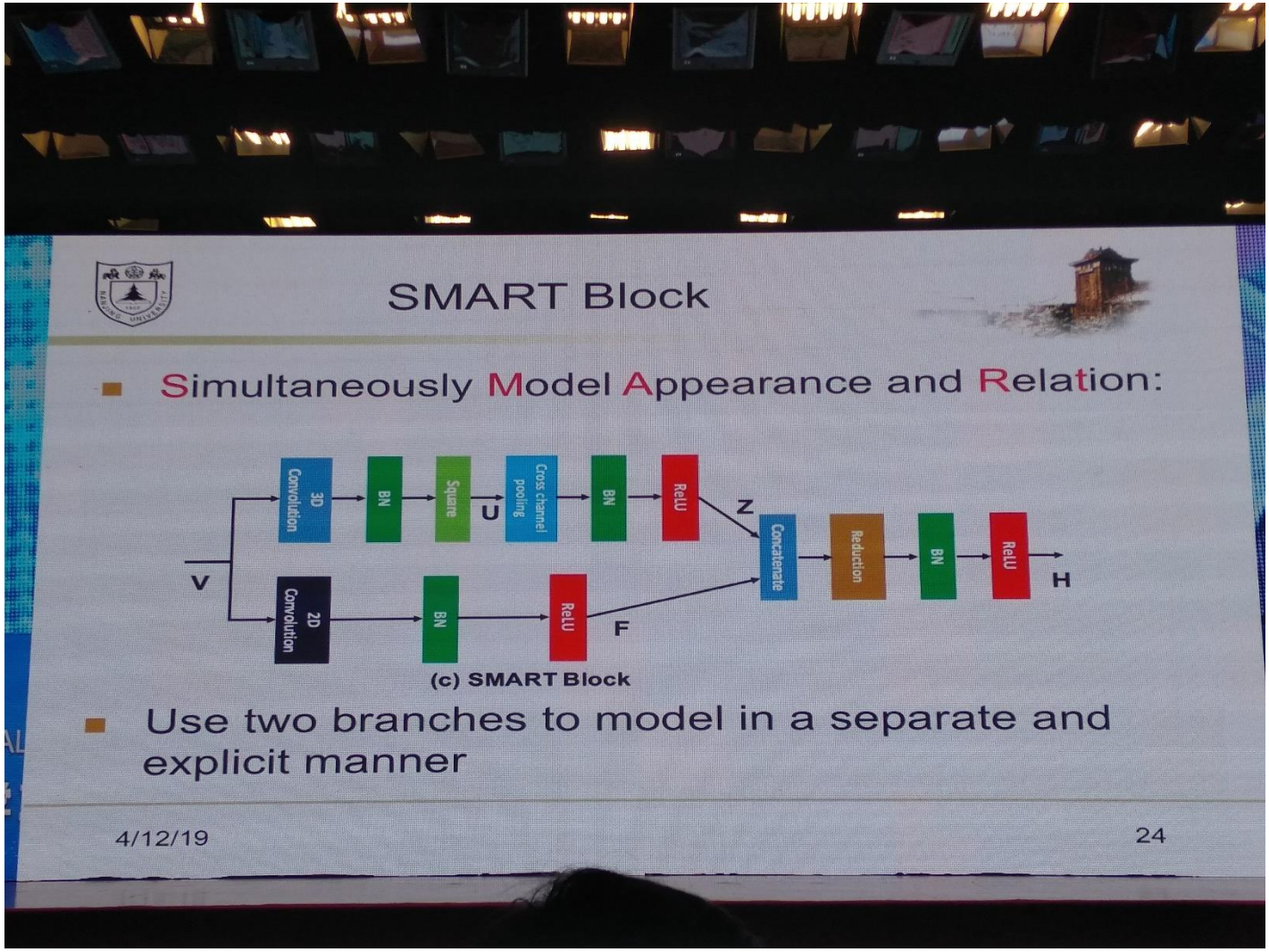
Person


Overview
Short Term Modeling
• Appearance-and-Relation Net (ARTNet)
Middle Term Modeling
• Temporal Segment Network (TSN)
Long Term Modeling
• UntrimmedNet
[1] L. Wang, W. Li, W. Li, and L. Van Gool, Appearance-and-Relation Networks for Video Classification, in CVPR, 2018.
[2] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, in ECCV, 2016.
[3] L. Wang, Y. Xiong, D. Lin, and L. Van Gool, UntrimmedNets for Weakly Supervised Action Recognition and Detection, in CVPR 2017.
4/12/19
16

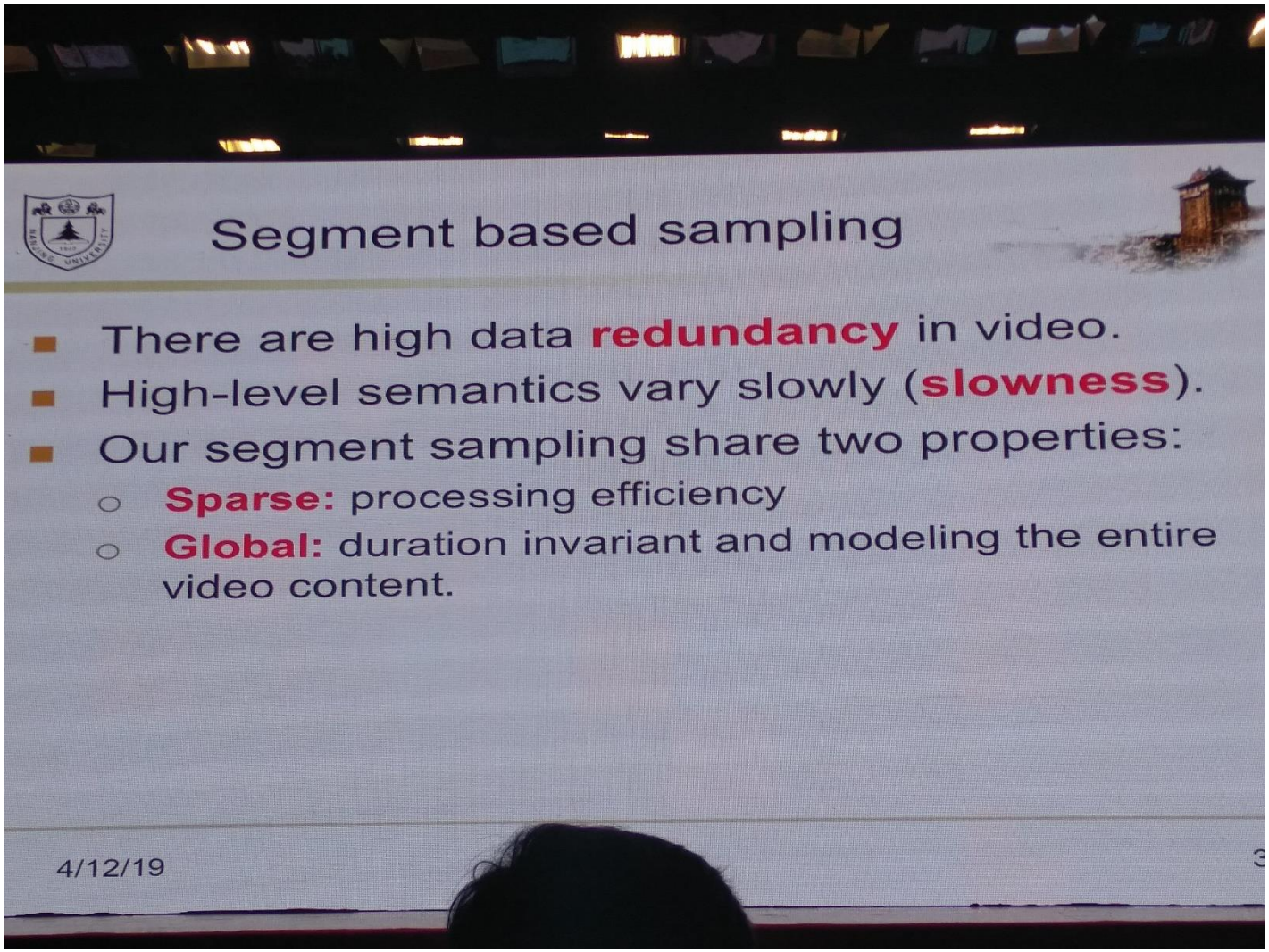
Person


Motivation of ARTNet
How to model appearance and relation (motion)
RGB Frames
Appearance
Relation
SMART Block
Concatenate
Reduction
Appearance
Relation
SMART Block
Concatenate
Reduction
FC
(c) Appearance-and-Relation Networks (ARTNets)
1. Assume independence between appearance and relation
2. Decouple and model them explicitly and separately
4/12/19
18

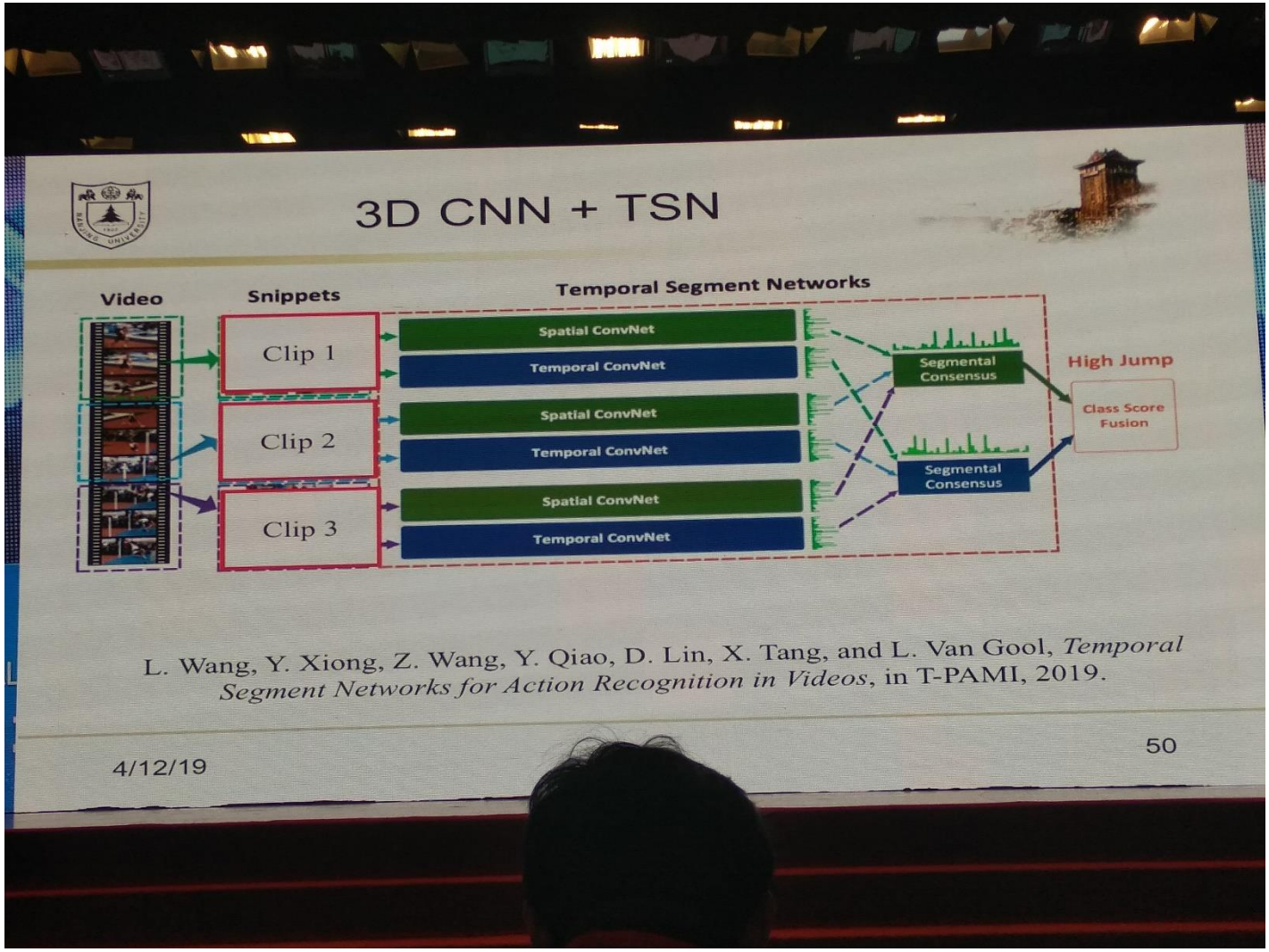
Person


SMART Block
Simultaneously Model Appearance and Relation:
3D Convolution
BN
Square
U
Cross channel pooling
BN
ReLU
Z
V
2D Convolution
BN
ReLU
F
Concatenate
Reduction
BN
ReLU
H
(c) SMART Block
Use two branches to model in a separate and explicit manner
4/12/19
24

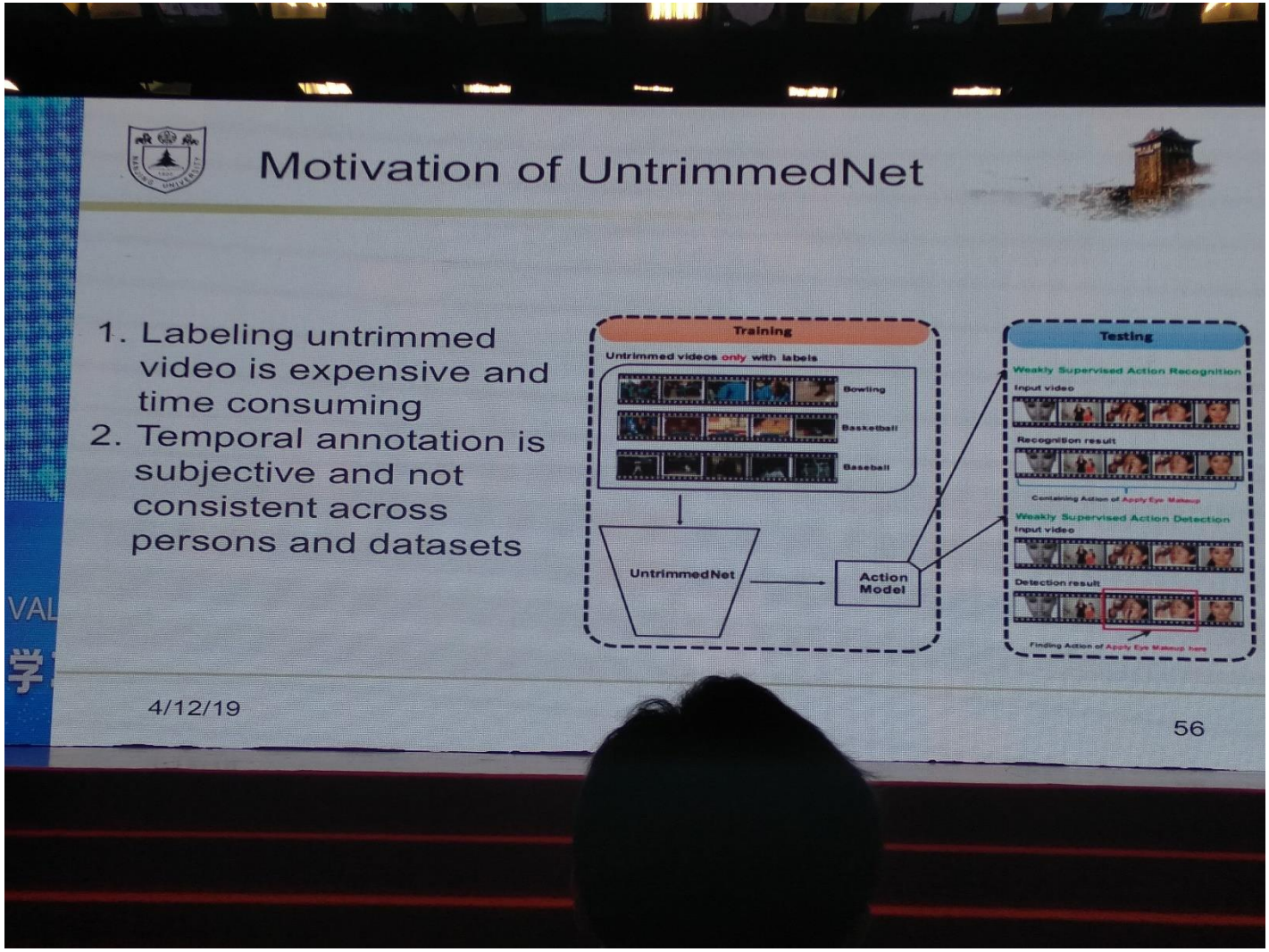
Person


Segment based sampling
There are high data redundancy in video.
High-level semantics vary slowly (slowness).
Our segment sampling share two properties:
Sparse: processing efficiency
Global: duration invariant and modeling the entire video content.
4/12/19

Person


3D CNN + TSN
Video
Snippets
Temporal Segment Networks
Clip 1
Clip 2
Clip 3
Spatial ConvNet
Temporal ConvNet
Spatial ConvNet
Temporal ConvNet
Spatial ConvNet
Temporal ConvNet
Segmental Consensus
Segmental Consensus
High Jump
Class Score Fusion
L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, Temporal Segment Networks for Action Recognition in Videos, in T-PAMI, 2019.
4/12/19
50

Person


Motivation of UntrimmedNet
1. Labeling untrimmed video is expensive and time consuming
2. Temporal annotation is subjective and not consistent across persons and datasets
Training
Untrimmed videos only with labels
Bowling
Basketball
Baseball
UntrimmedNet
Action Model
Testing
Weakly Supervised Action Recognition
Input video
Recognition result
Weakly Supervised Action Detection
Input video
Detection result
4/12/19
56

Outline

- About VALSE
- 弱监督视觉理解与主动学习
- 以人为中心的视觉理解
- 视觉中的知识推理
- CV+NLP 2.0
- Poster 选讲

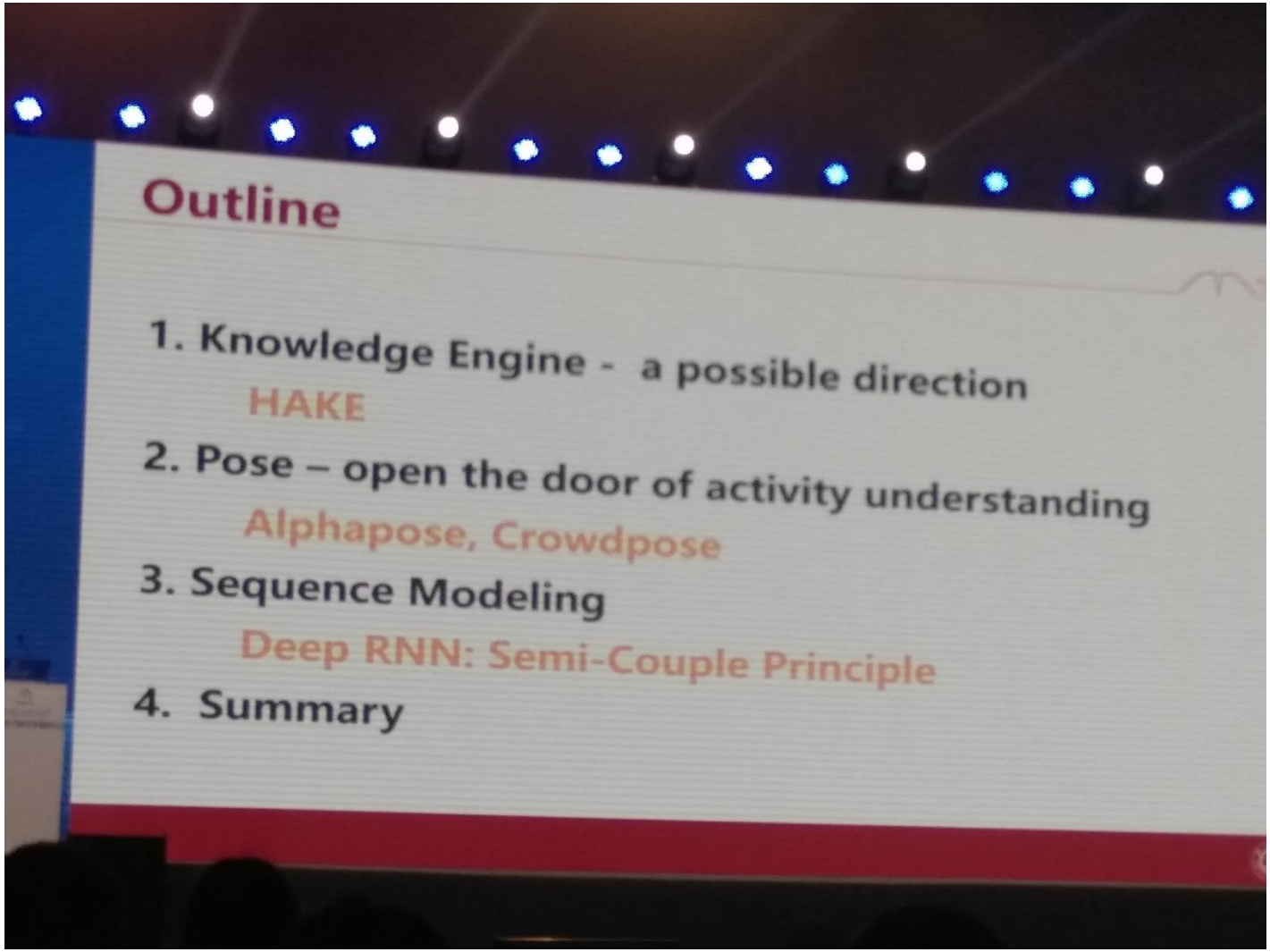
Knowledge | Resoning

- 基于知识驱动的行为理解(Ceiwu Lu, SJU)
- Towards X visual reasoning (hanwang zhang, NTU)

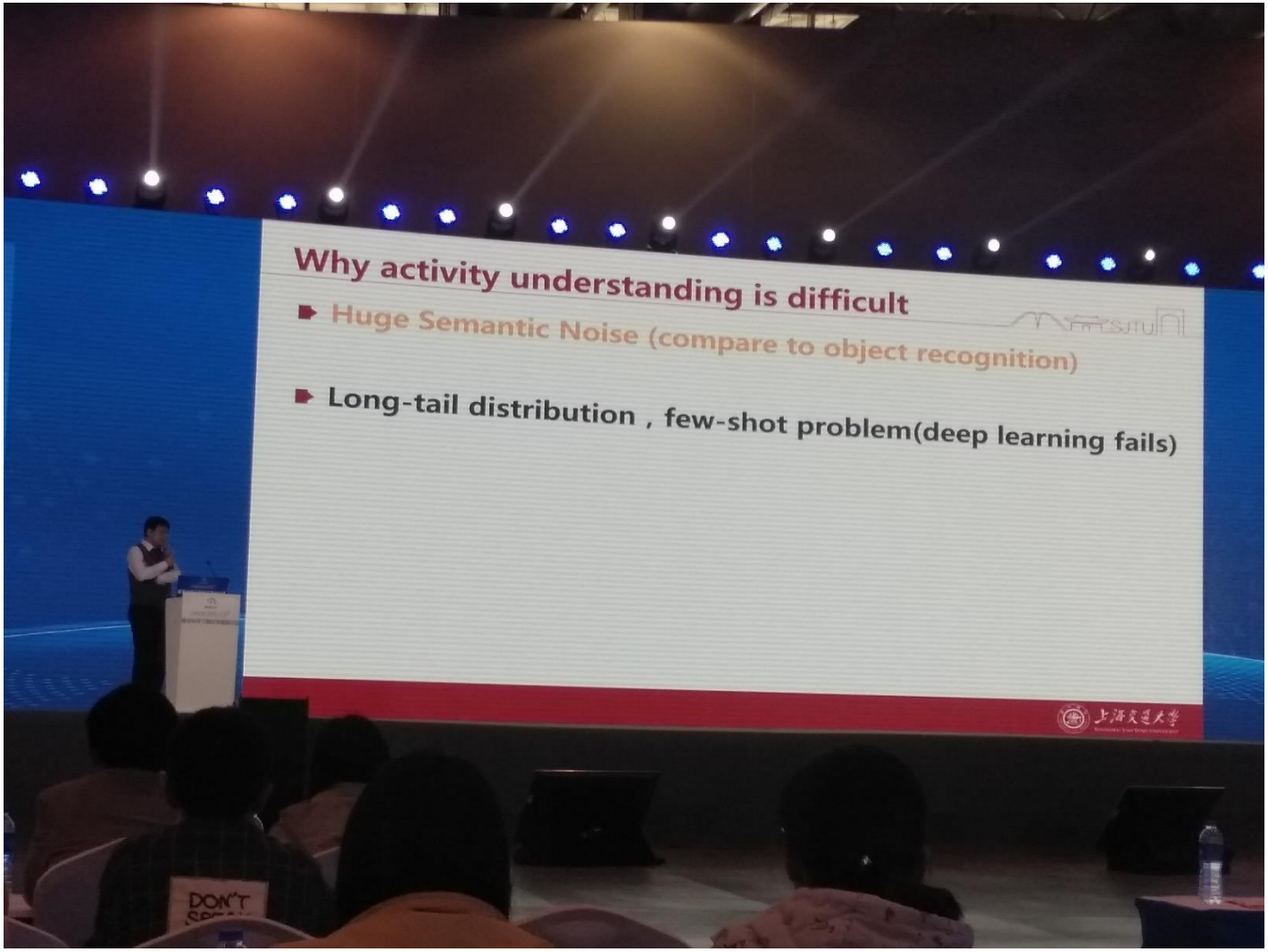
Knowledge | Resoning

- 基于知识驱动的行为理解(Ceiwu Lu, SJU)
- Towards X visual reasoning (hanwang zhang, NTU)

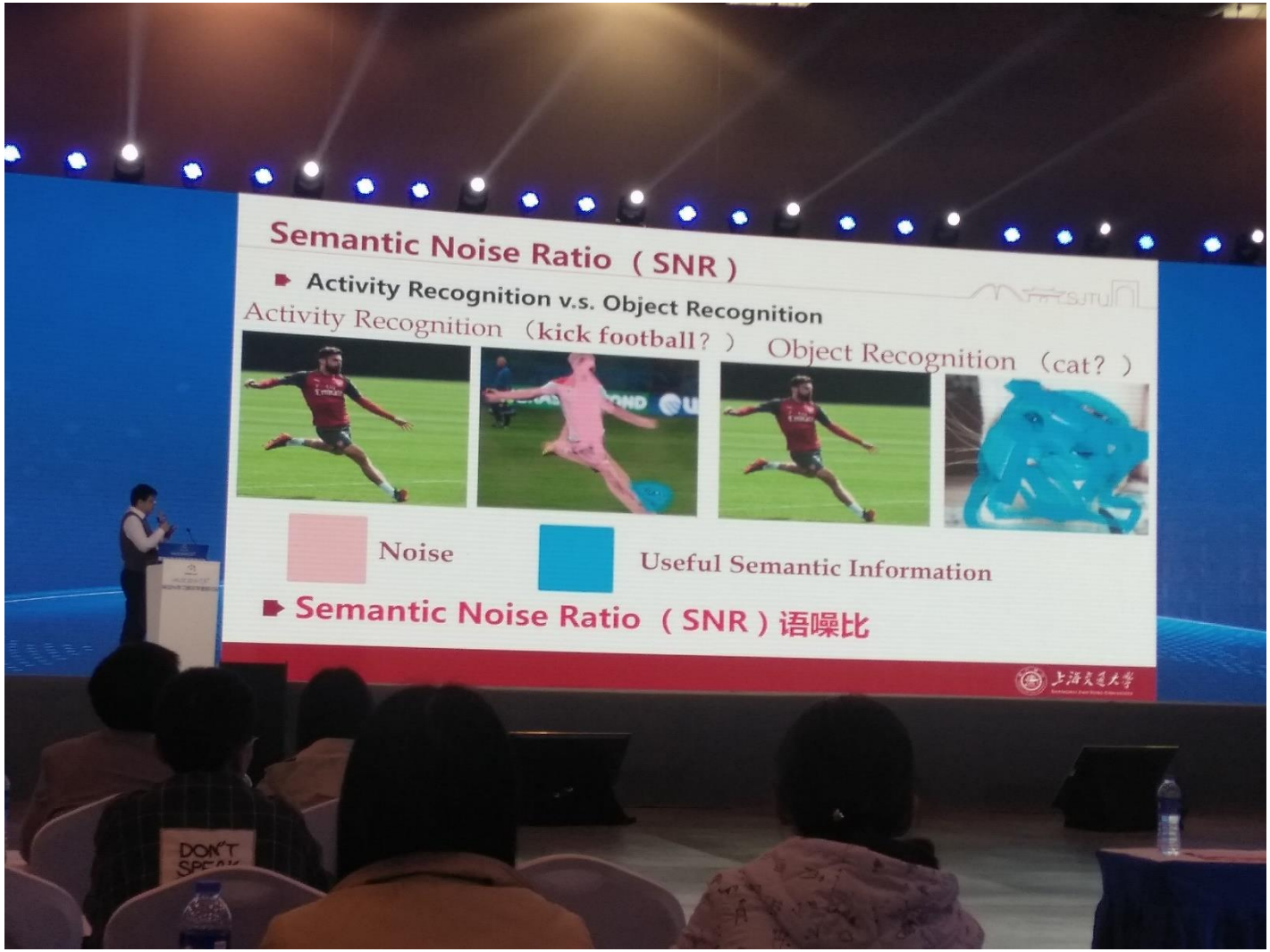
Knowledge | Resoning



Knowledge | Resoning

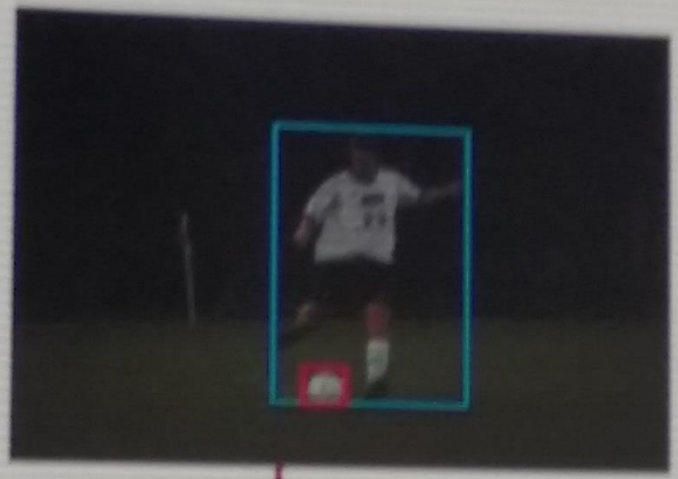


Knowledge | Resoning



Analogy: A Toy Model

Instance-level Understanding

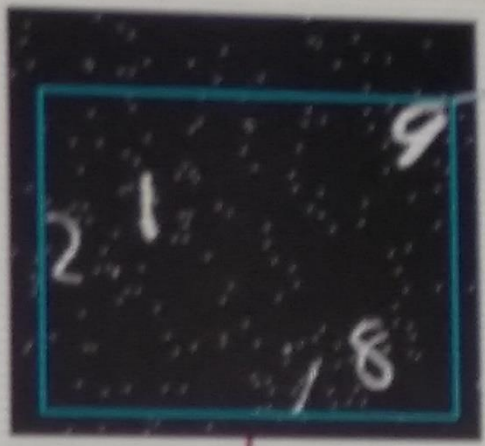


$f(I)$?

Labels:
person-kick-football

Analogy

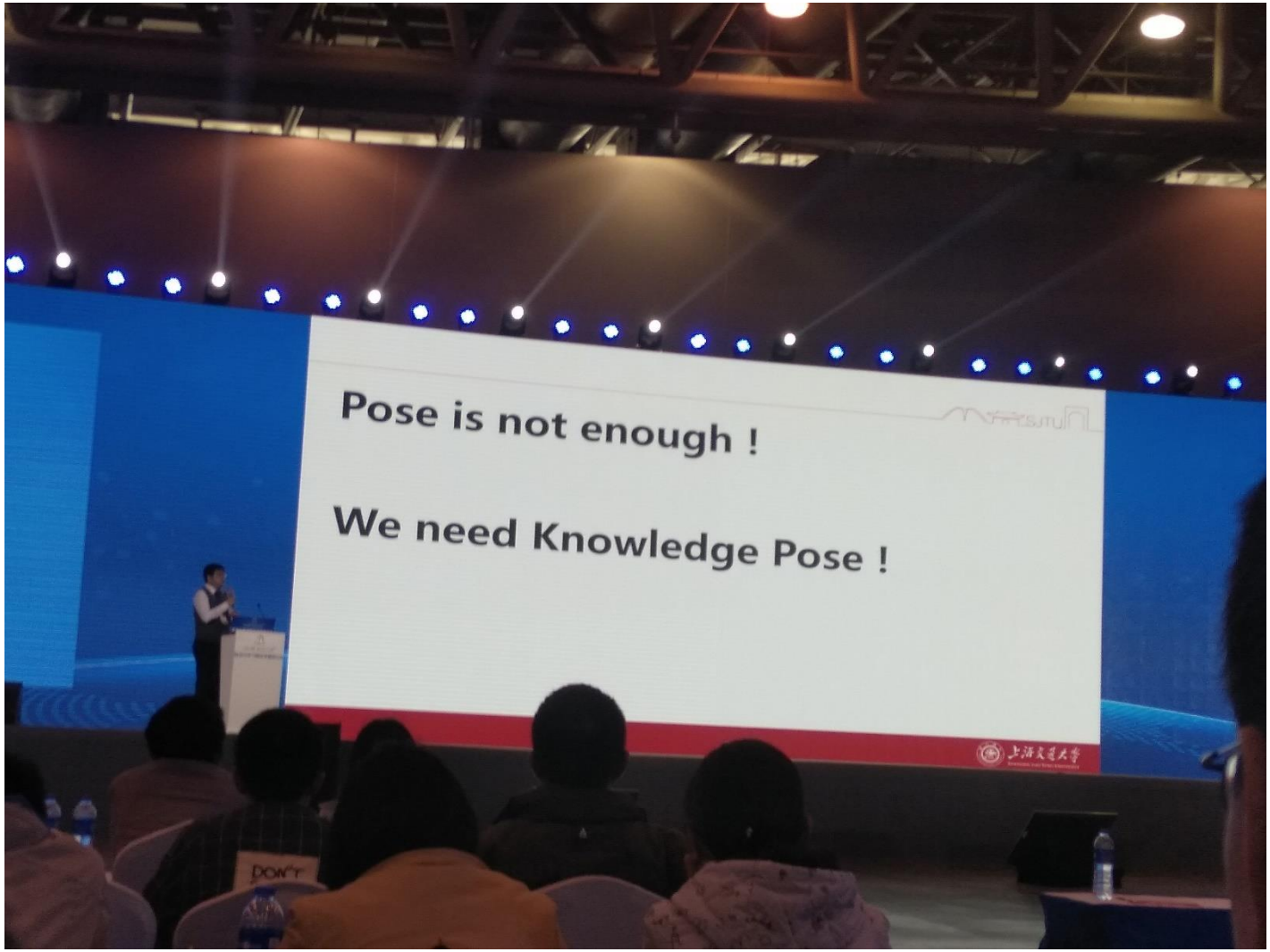
Assumption: $g(\cdot)$ is
"The sum of top-2 values, e.g. 9+8=17"
model → fitting → $g(\cdot)$



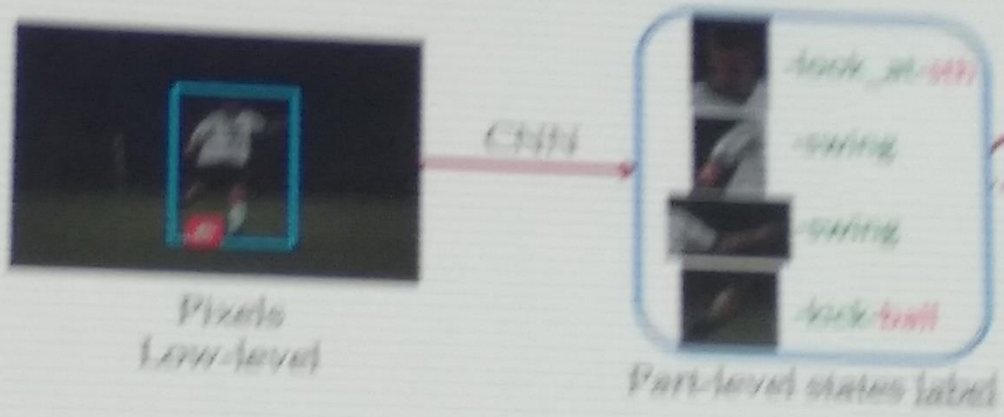
$g(x)$?

Labels:
17

Knowledge | Resoning

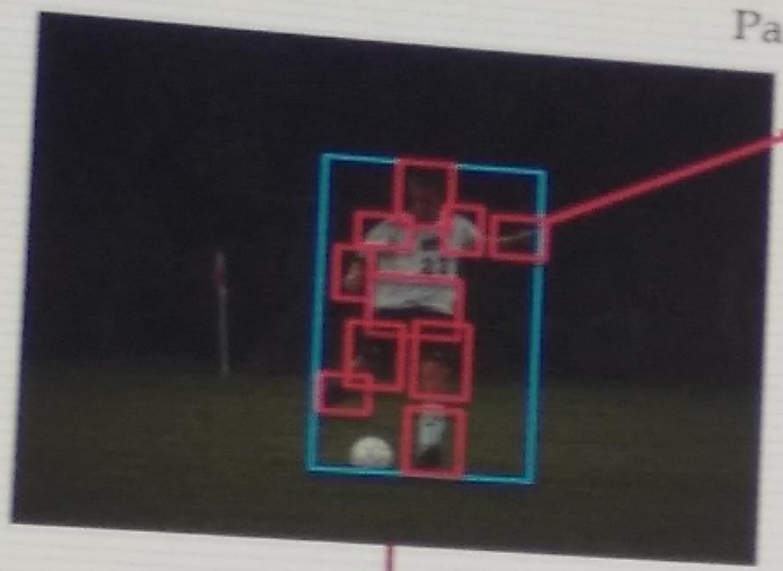


Two-stage Paradigm: Overview



nalogy: A Toy Model

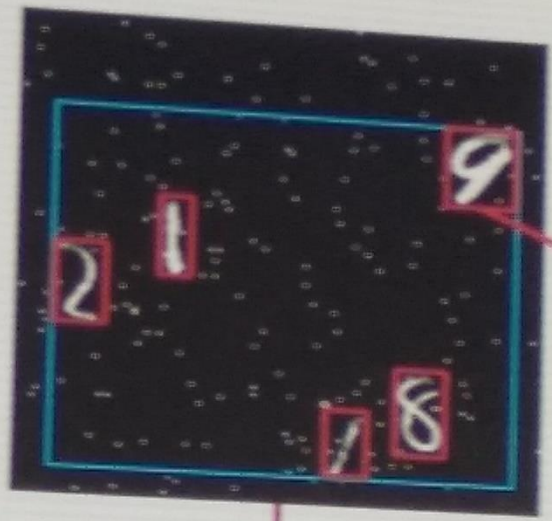
A analogy to human activity recognition:



"Part Bounding Box" from pose keypoints, Annotators → Part States

Part-level Understanding

Analogy



"Part Bound
Annotators

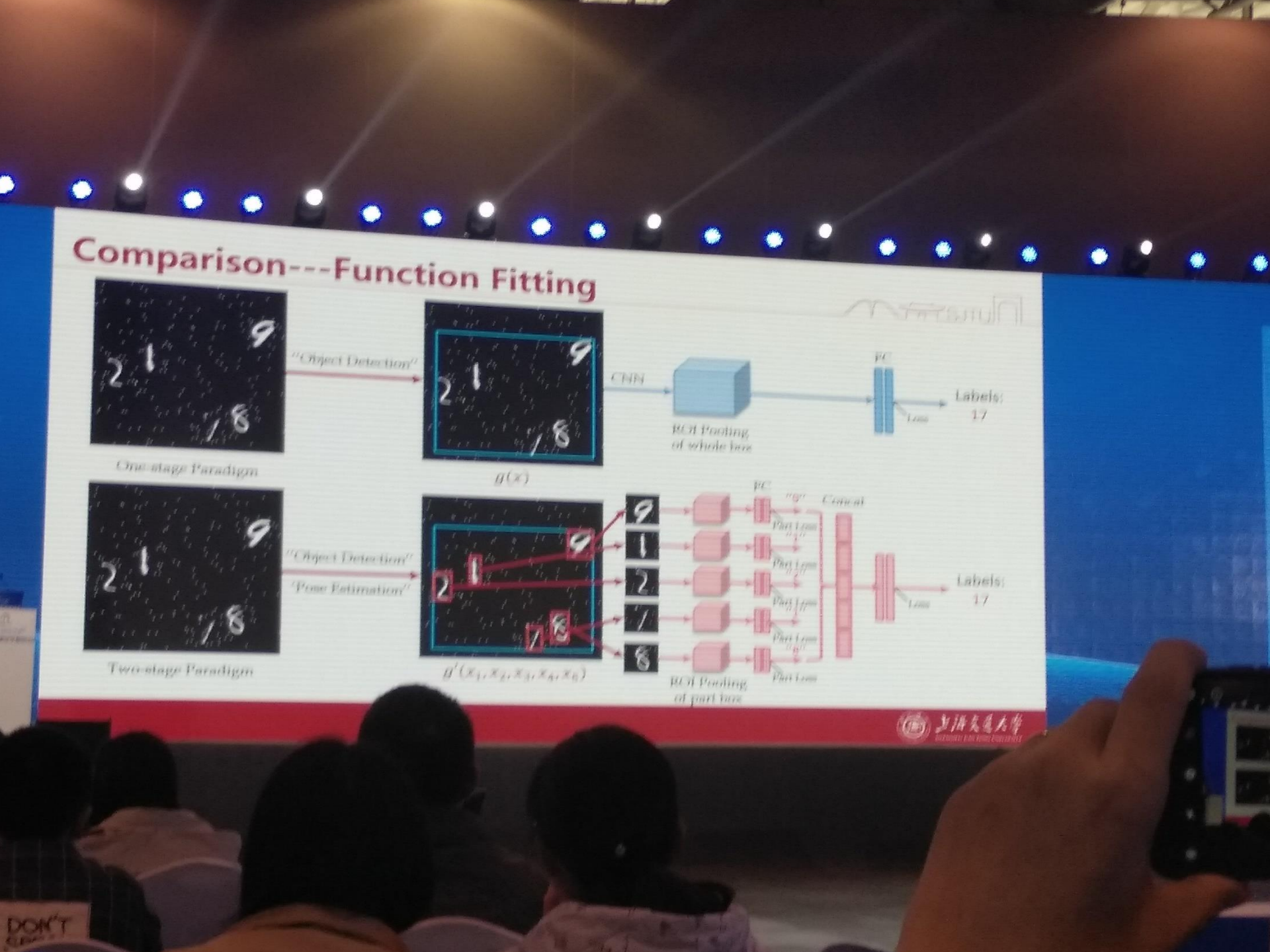
$f(I)$

Part State Labels:
head-look_at-sth right_arm-swing
left_arm-swing right_foot-

$g(x) \rightarrow g'(x_1, x_2, x_3, x_4, x_5)$

Number Labels:
9 8 1 1 2

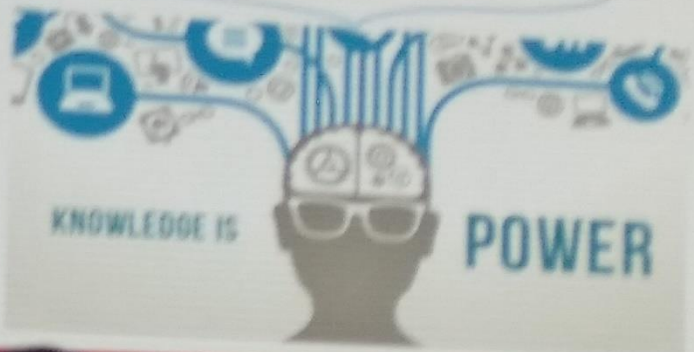
Labels:
17


Comparison---Function Fitting
"Object Detection"
CNN
FC
Labels:
17
RCR Pooling
of whole face
One-stage Paradigm
g(x)
"Object Detection"
"Pose Estimation"
FC
"9"
Concat
Labels:
17
Two-stage Paradigm
RCR Pooling
of part face

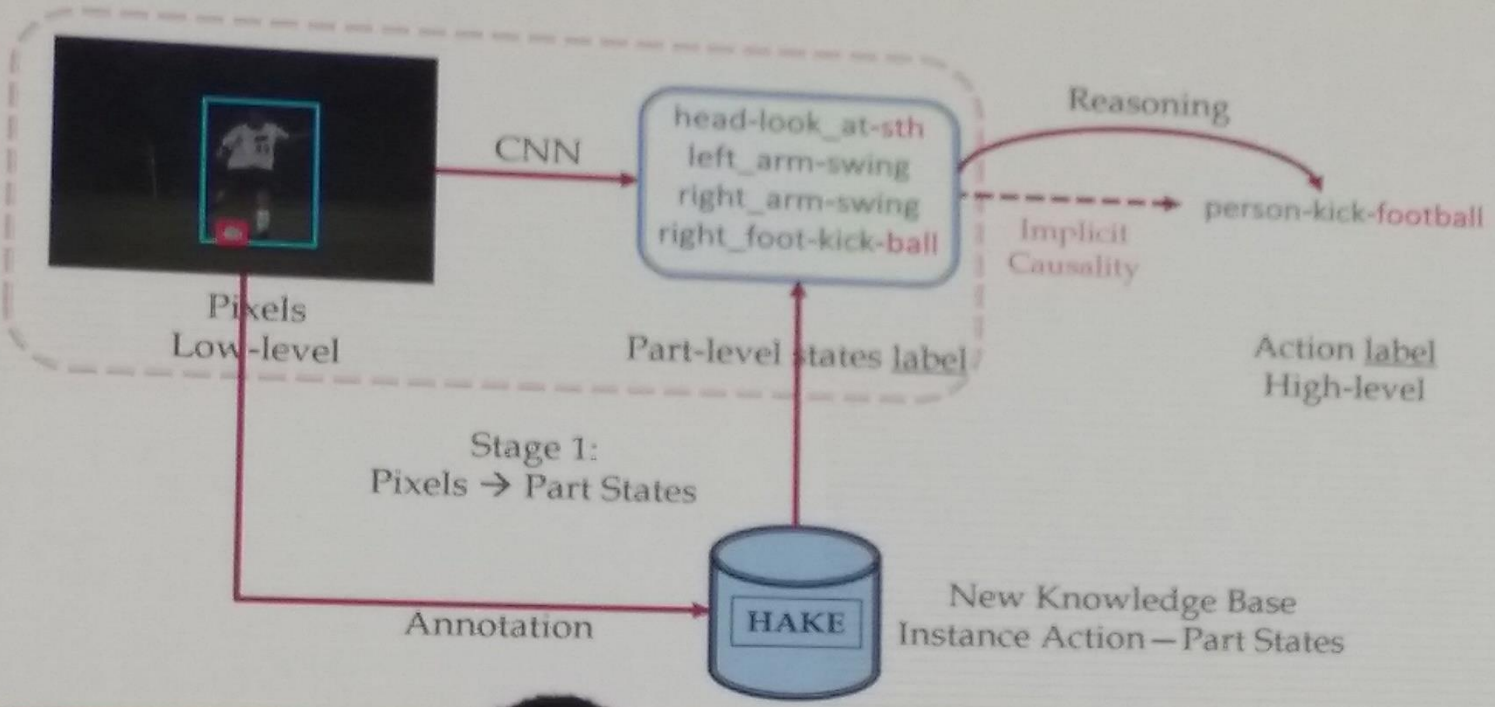
Human Activity Knowledge Engine

HAKE

To see the activity
To parse the activity
To understand the activity

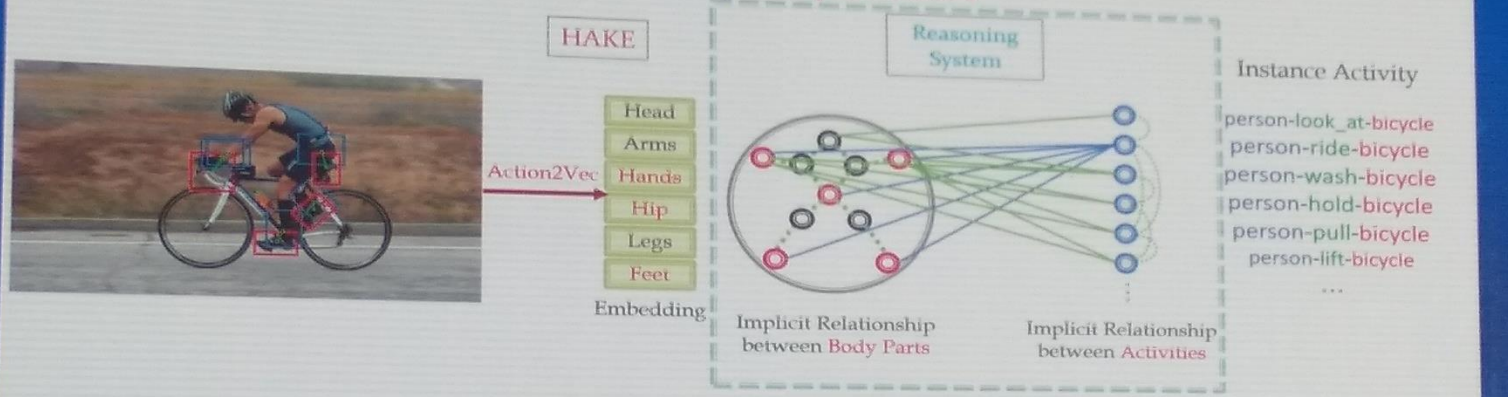


Knowledge Engine Construction



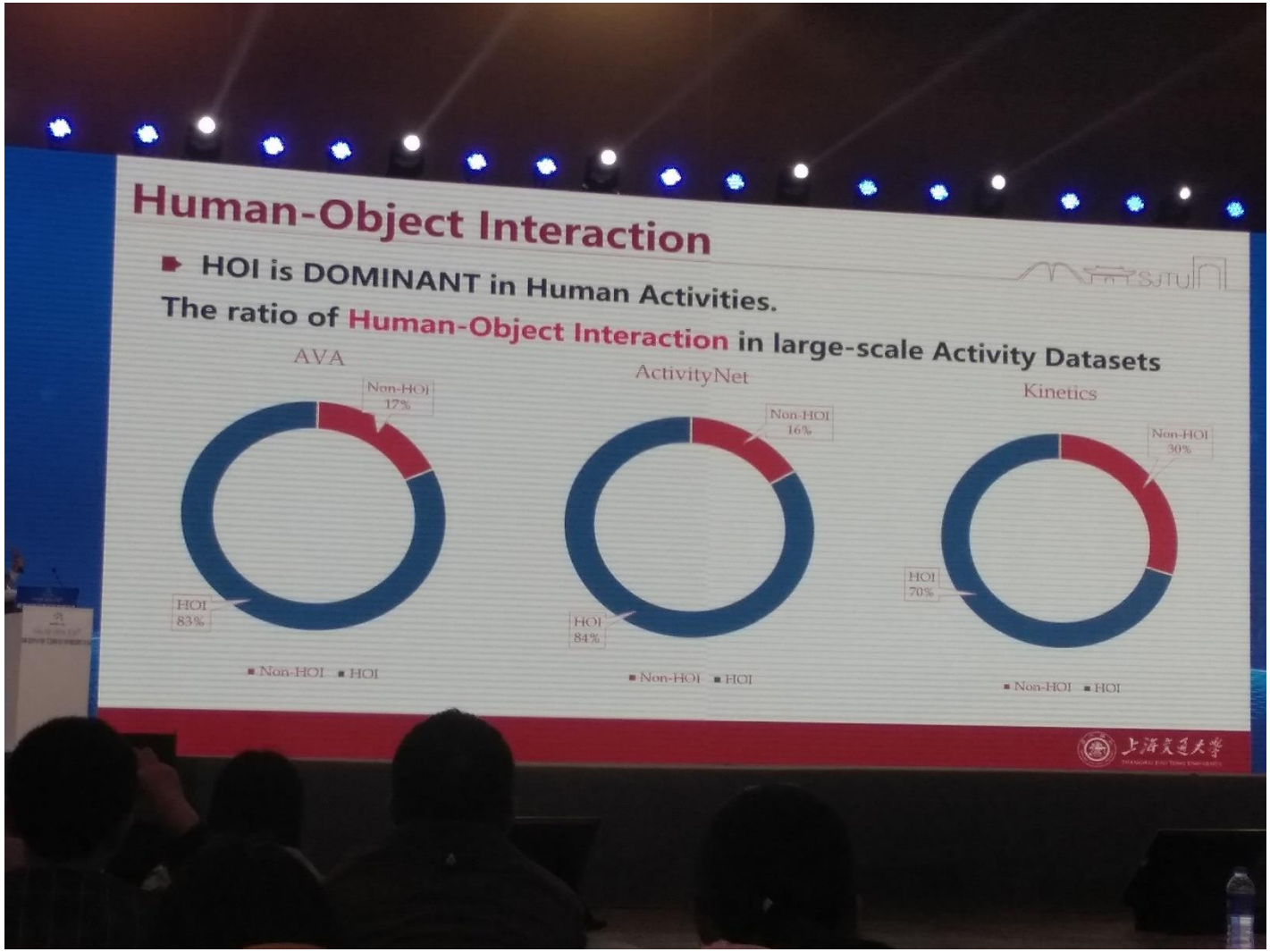
Reasoning via Part States

- Activity Knowledge based Reasoning
  - Implicit relationship between activity and part states



- Reasoning from Part States to Instance Activities $R(\{p_i\}_{i=1,2,\ldots,10}) = \{A_j\}_{j=N}$
- Taking GNNs, RNNs, FCs as $R(\bullet)$ all show prominent performance improvement

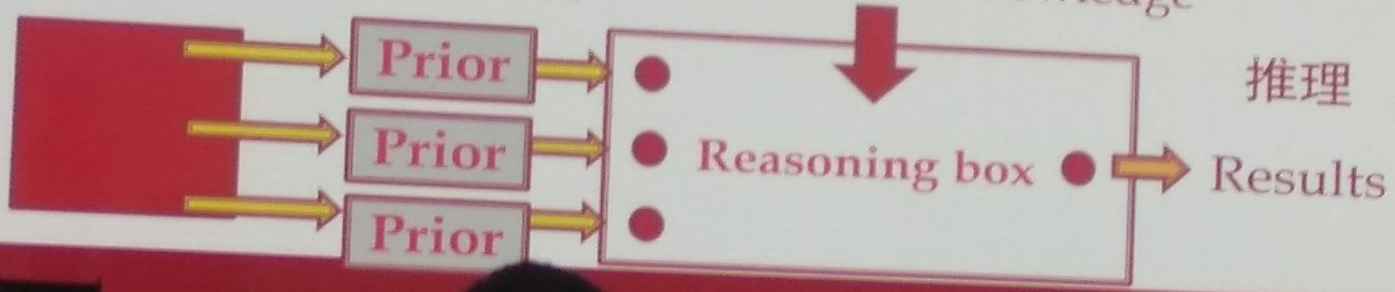
Knowledge | Resoning



Our Insight : Knowledge Equation

Prior-Reasoning

可被暴力解决



Our Insight : Knowledge Equation

Black box
Deep Learning

直觉
Results

Prior-Reasoning

可被暴力解决

Out-source Knowledge

Prior

Prior

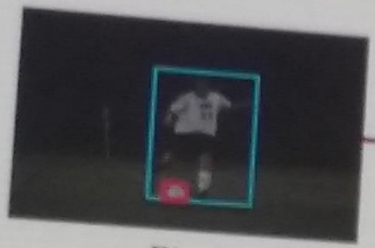
Prior

Reasoning box

推理
Results

Conclusion

- Activity data is Semantically NOISY!
- Knowledge at Body Part can help to denoise
- HAKE: Human Activity Knowledge Engine
- HAKE based Two-stage Paradigm:



Pixels
Low-level

CNN

head-look_at-sth
left_arm-swing
right_arm-swing
right_foot-kick-ball

Part-level states label

Reasoning

Implicit
Causality

person-kick-football

Action label
High-level

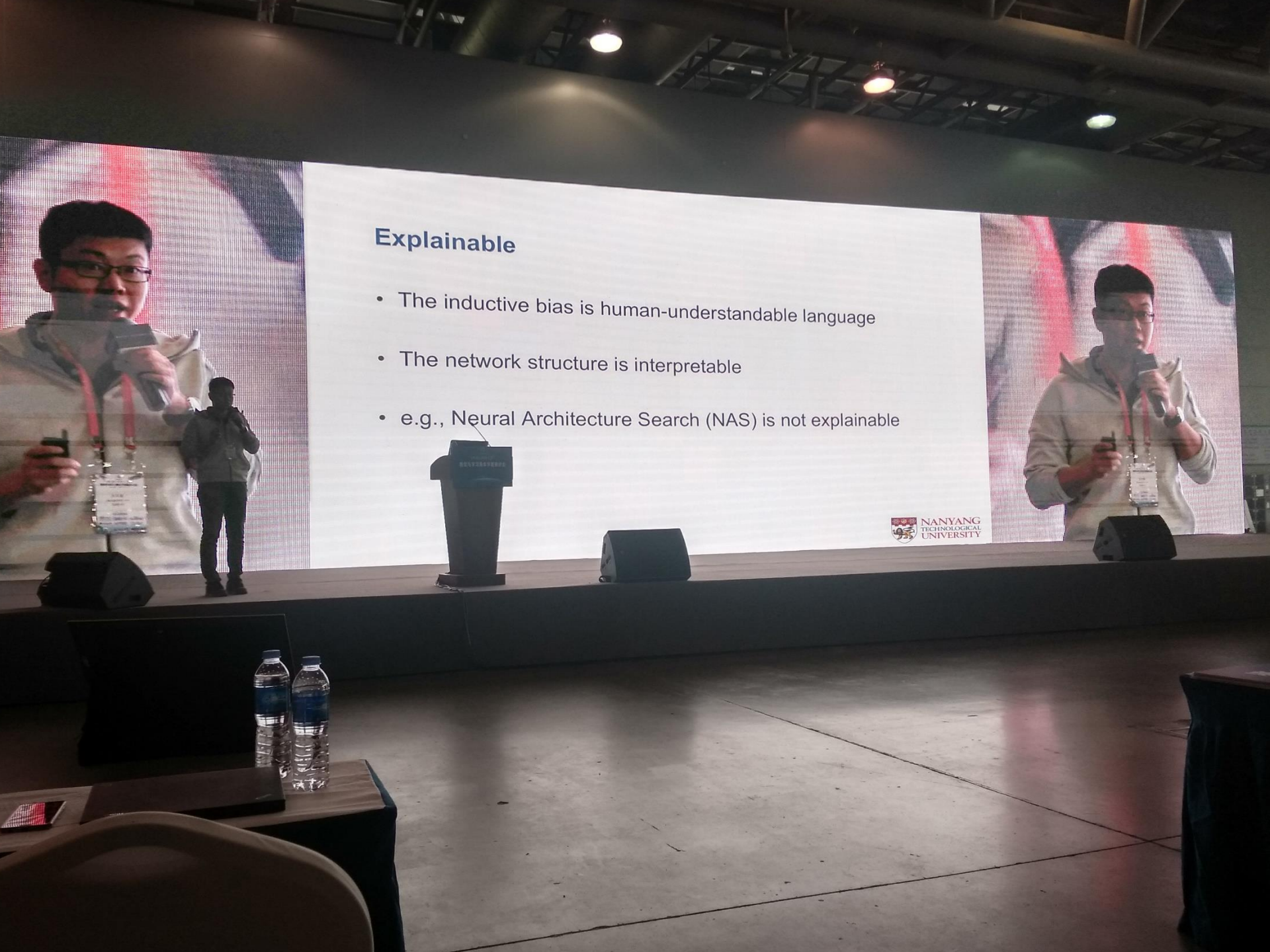
Stage 1:
Pixels → Part States

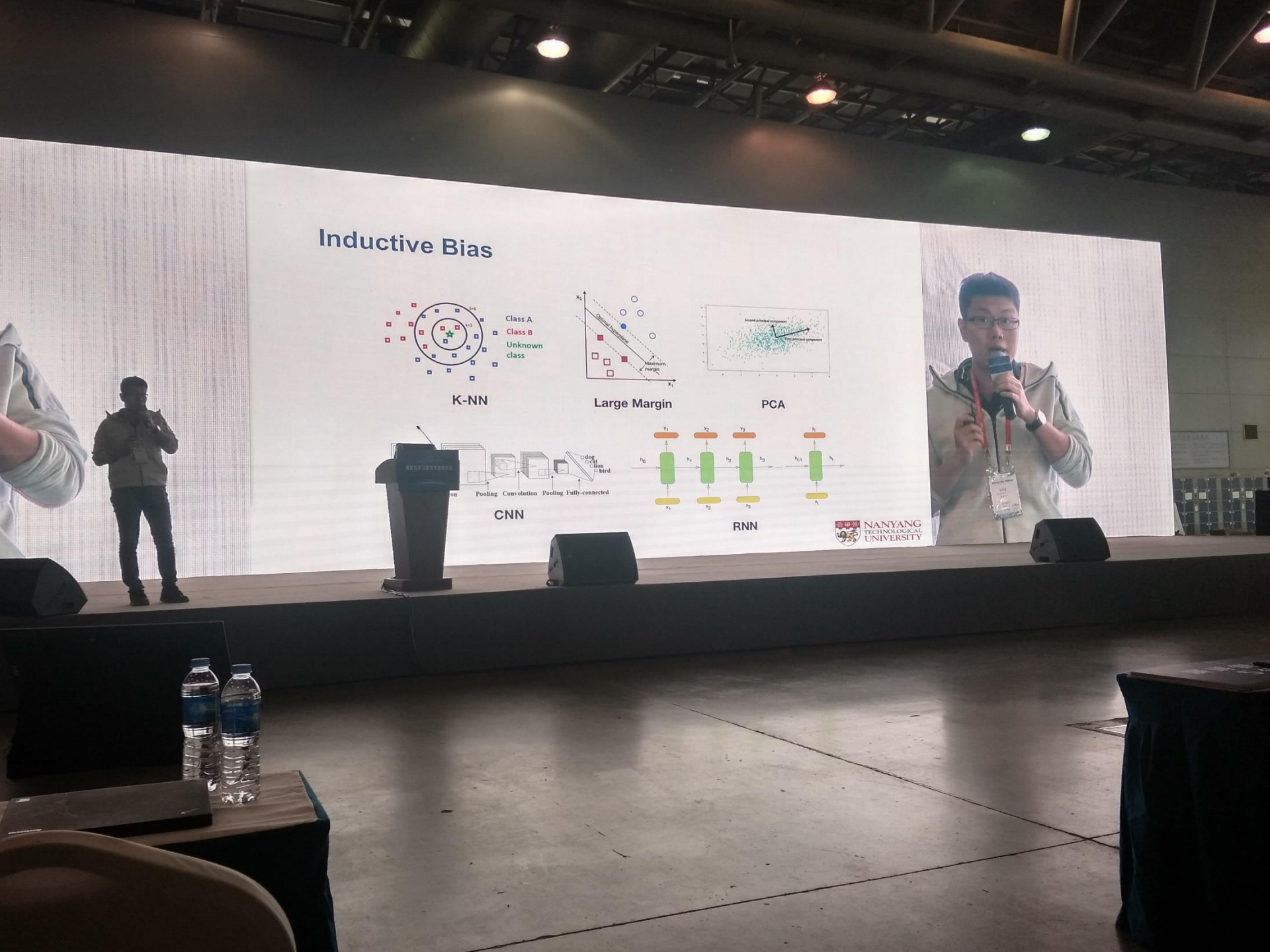
Stage 2:
Part States → Instance Action

Knowledge | Resoning

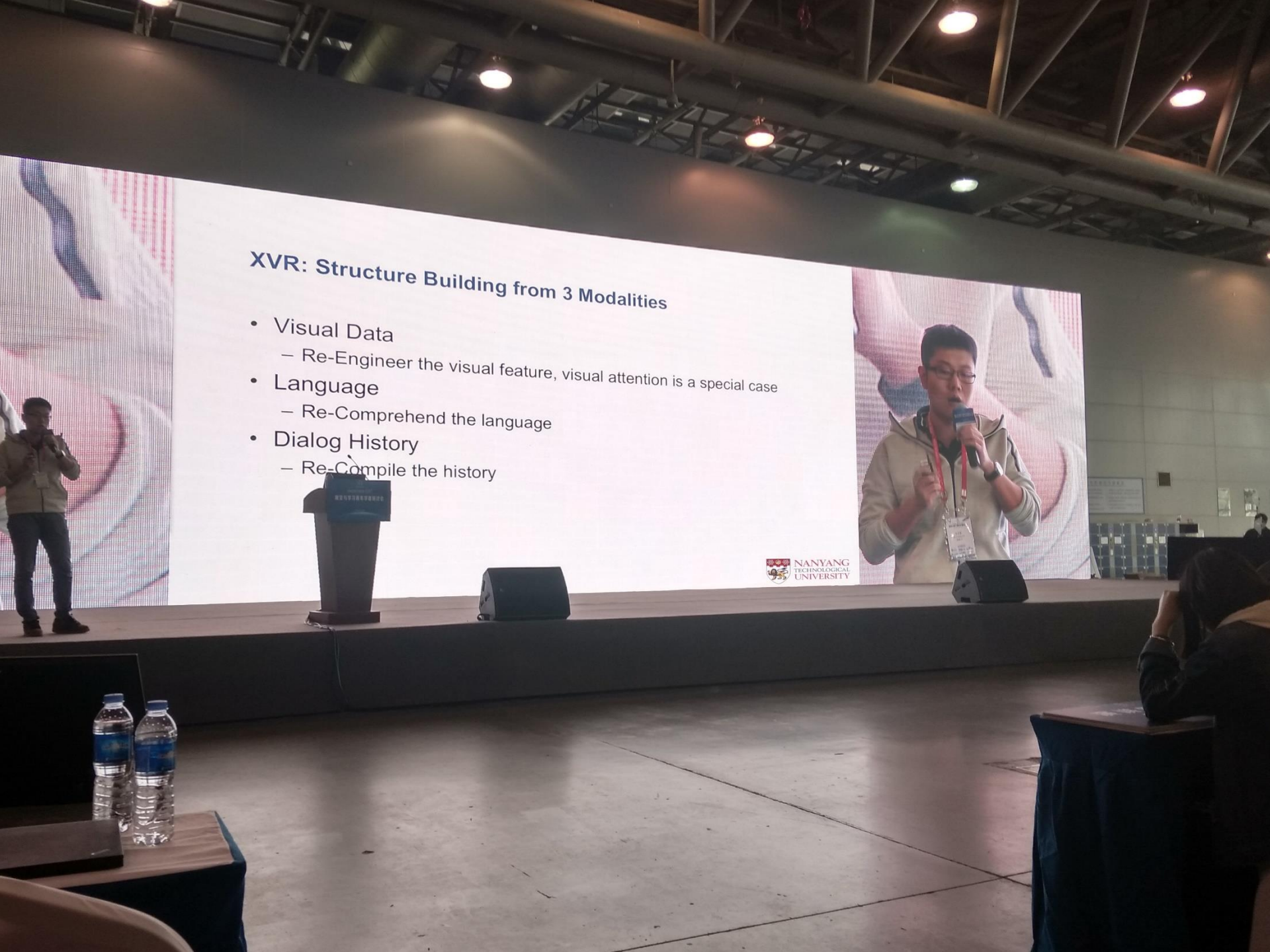
- 基于知识驱动的行为理解(Ceiwu Lu, SJU)
- Towards X visual reasoning (hanwang zhang, NTU)


X Visual Reasoning
(eXplainable & eXplicit)
Hanwang Zhang 张含望
MReaL Lab (mreallab.github.io)
hanwangzhang@ntu.edu.sg
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
School of Computer Science and Engineering
NANYANG TECHNOLOGICAL UNIVERSITY


Explainable
• The inductive bias is human-understandable language
• The network structure is interpretable
• e.g., Neural Architecture Search (NAS) is not explainable
NANYANG TECHNOLOGICAL UNIVERSITY


Inductive Bias
Class A
Class B
Unknown class
K-NN
Optimal hyperplane
Maximum margin
Large Margin
Second principal component
First principal component
PCA
dog
cat
lion
bird
Pooling
Convolution
Pooling
Fully-connected
CNN
RNN
NANYANG TECHNOLOGICAL UNIVERSITY

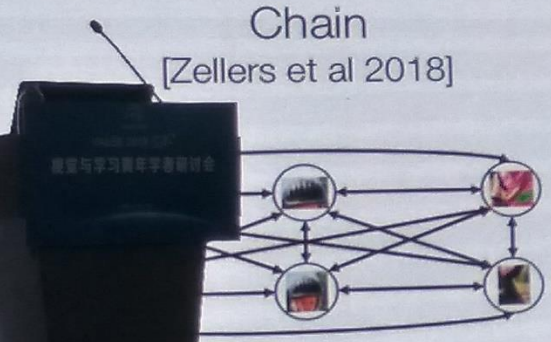

Inductive Bias for Visual Reasoning
• How to define the bias?
– Where is it from?
– What is it like?
• How to use the bias?
– Where to impose it?
– How to make it computationally feasible?
NANYANG
TECHNOLOGICAL
UNIVERSITY


XVR: Structure Building from 3 Modalities
• Visual Data
– Re-Engineer the visual feature, visual attention is a special case
• Language
– Re-Comprehend the language
• Dialog History
– Re-Compile the history
NANYANG TECHNOLOGICAL UNIVERSITY

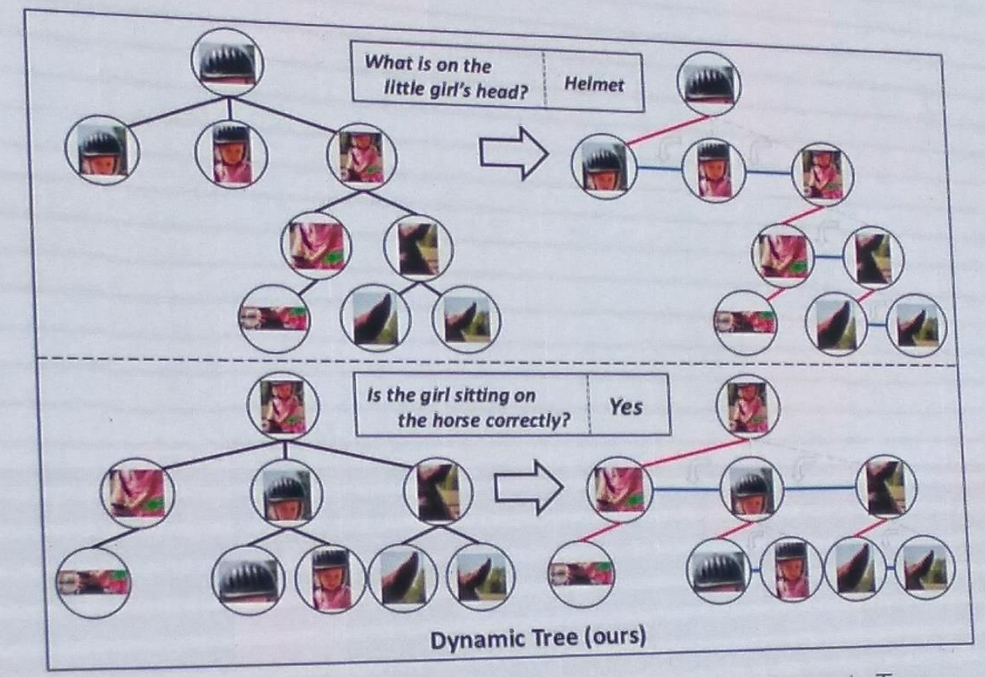
Tree

Independent Nodes
[Anderson et al 2018]

Chain
[Zellers et al 2018]



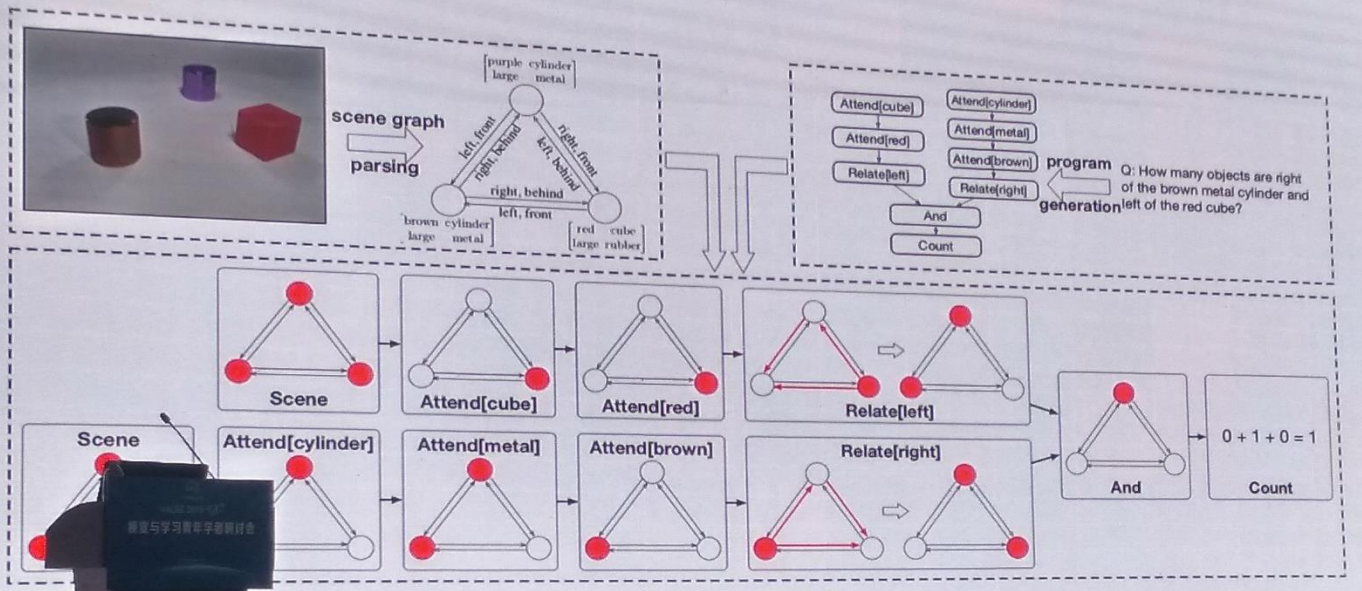
Graph
[Xu et al 2017]



Tang et al. **CVPR'19 oral**. *Learning to Compose Dynamic Tree Structures for Visual Contexts*

NANYANG TECHNOLOGICAL UNIVERSITY

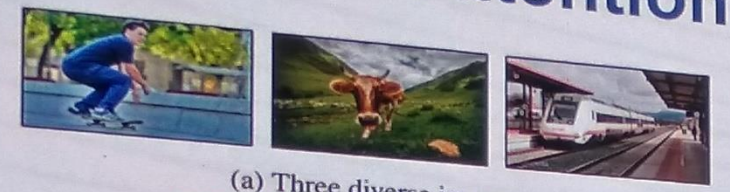
Scene Graph



Shi et al. **CVPR'19** *Explainable and Explicit Visual Reasoning over Scene Graphs*

NANYANG TECHNOLOGICAL UNIVERSITY

Modular Attention



(a) Three diverse images.

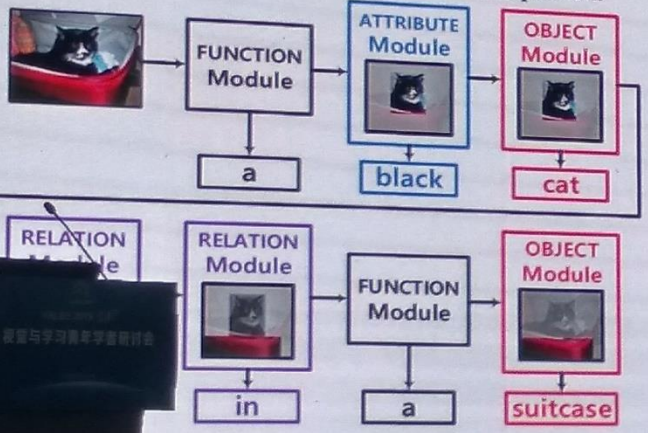
STH. DO STH. AT SOMEPLACE

A man plays a board on the road.

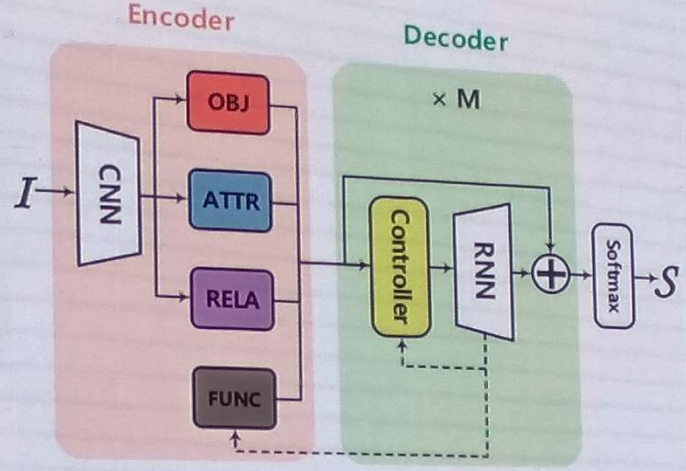
A cow is eating grass in a grassy hill

A train is coming into a platform

(b) Three captions with the same sentence pattern.



caption generation process of CNM.

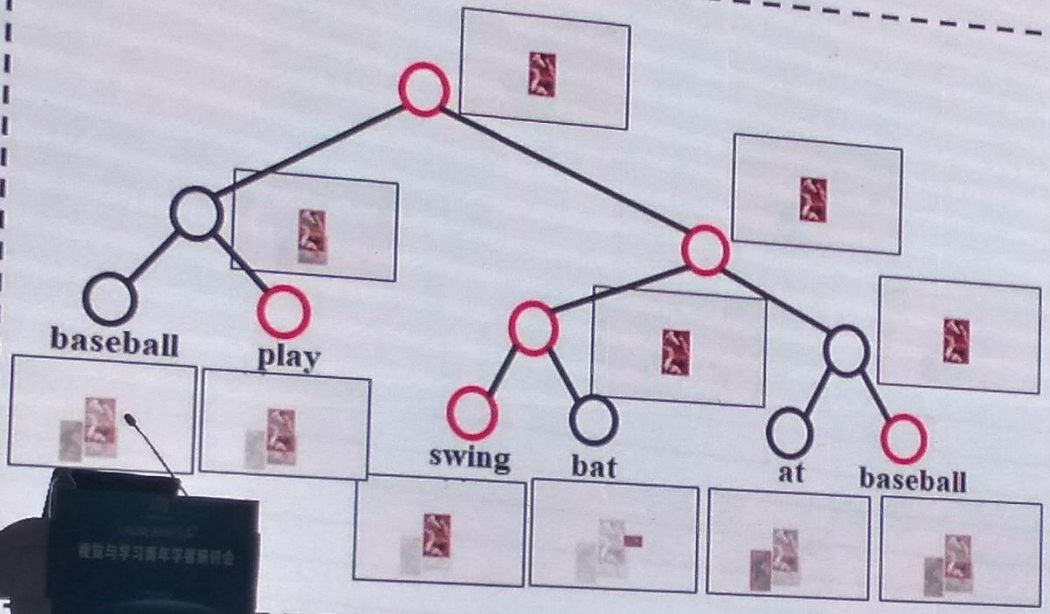


C: 127.9 Kaparthy, 126.0 server

ang et al. arxiv'19 *Learning to Collocate Neural Modules for Image Captioning*

NANYANG TECHNOLOGICAL UNIVERSITY

Constituency Parsing Tree

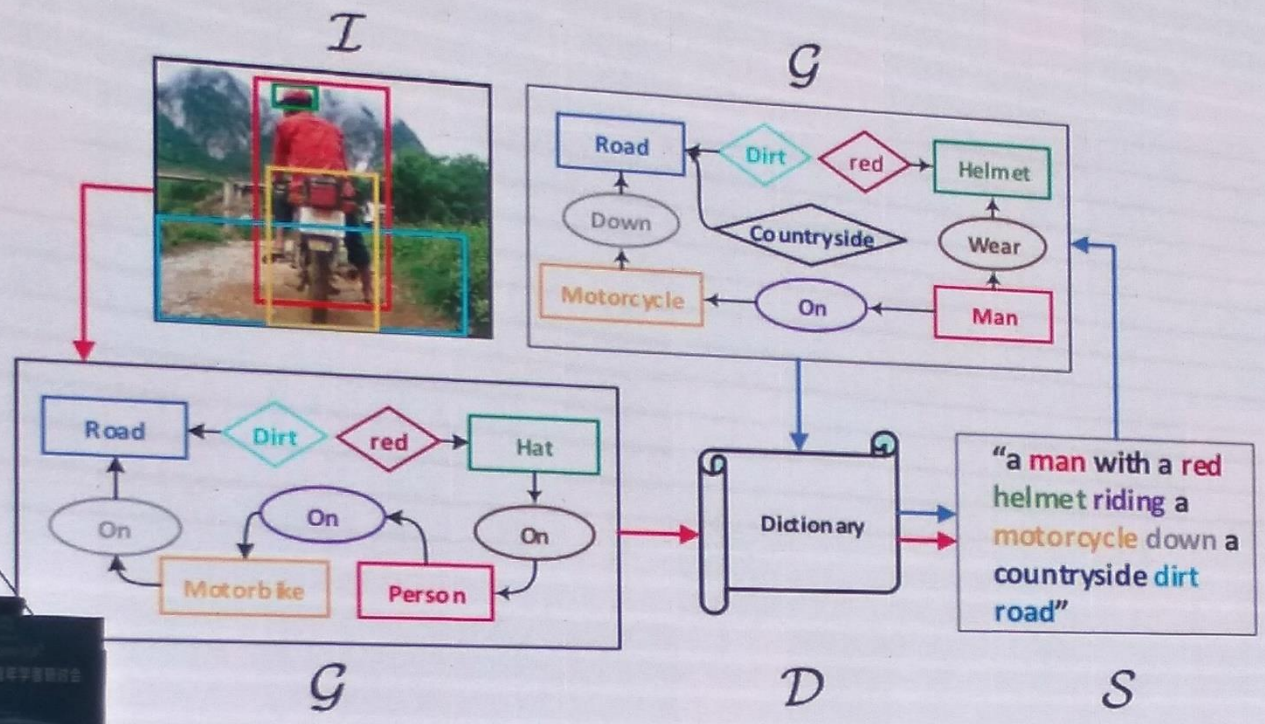


baseball player swinging bat at baseball

TPAMI'19 *Learning to Compose and Reason with Language Tree Structures for Visual Grounding*

NANYANG TECHNOLOGICAL UNIVERSITY

Graph

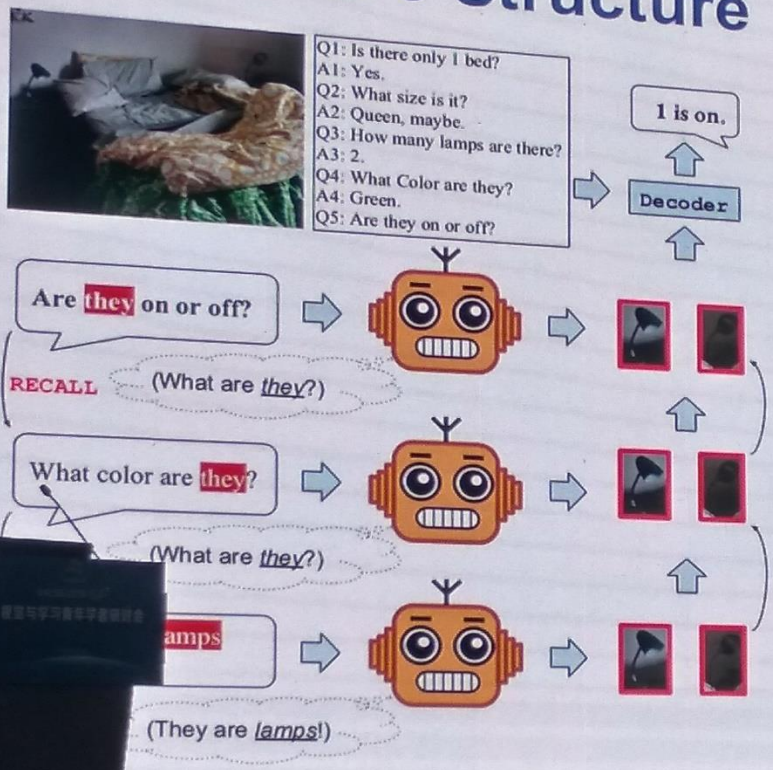


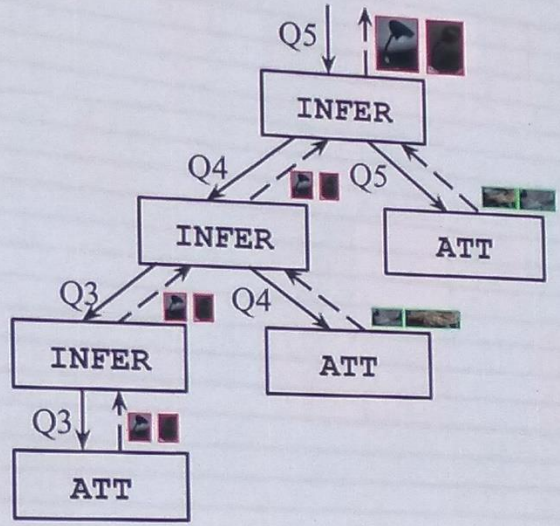
C: 129.1 Kaparthy, 126.5 server (code available)

Yang et al. **CVPR'19 oral** *Auto-Encoding Scene Graphs for Image Captioning*

NANYANG TECHNOLOGICAL UNIVERSITY

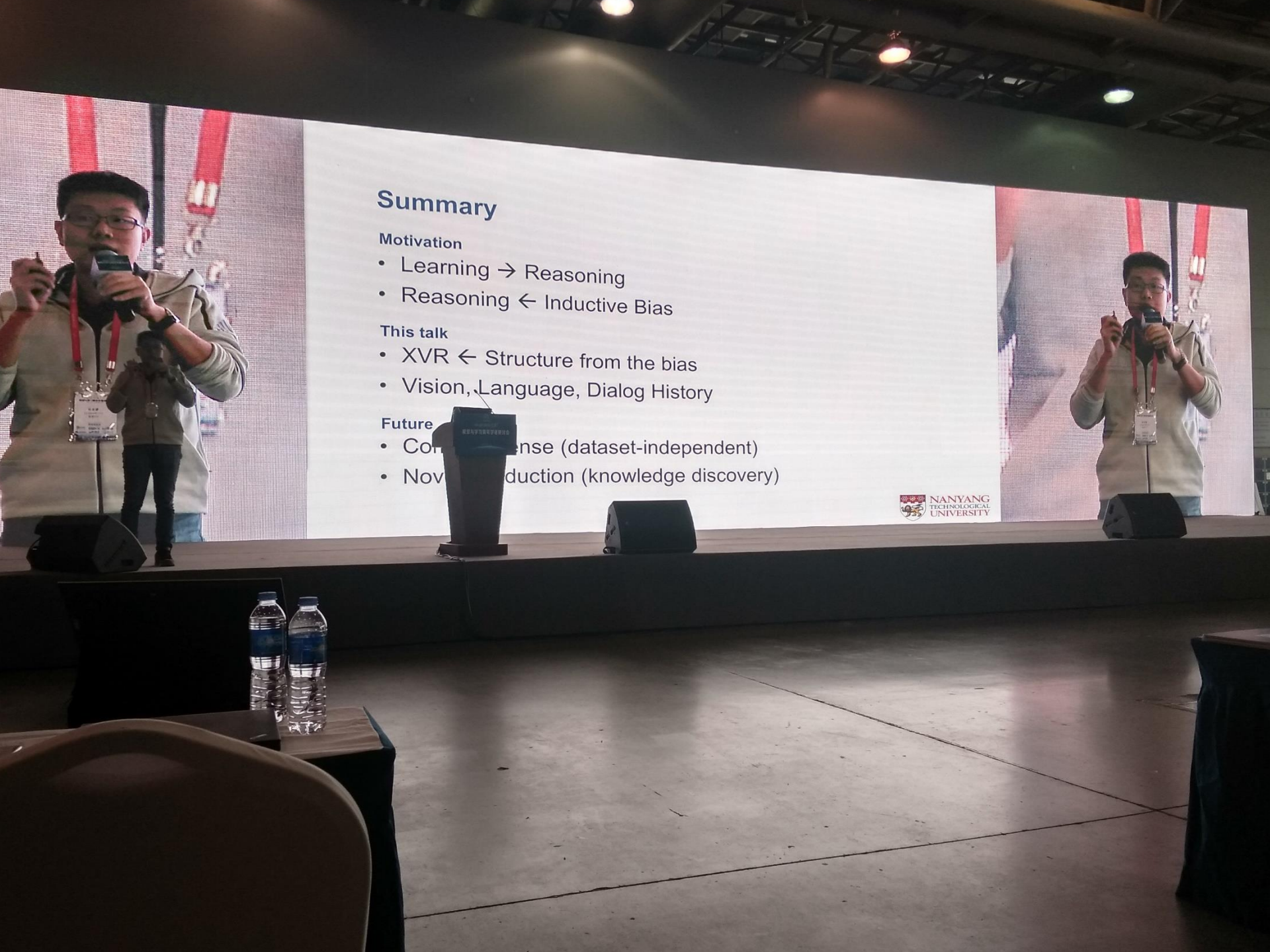
History-aware Structure





Niu et al. **CVPR'19 oral** Recursive Visual Attention in Visual Dialog

NANYANG TECHNOLOGICAL UNIVERSITY


Summary
Motivation
• Learning → Reasoning
• Reasoning ← Inductive Bias
This talk
• XVR ← Structure from the bias
• Vision, Language, Dialog History
Future
• Co nse (dataset-independent)
• Nov duction (knowledge discovery)
NANYANG
TECHNOLOGICAL
UNIVERSITY

Outline

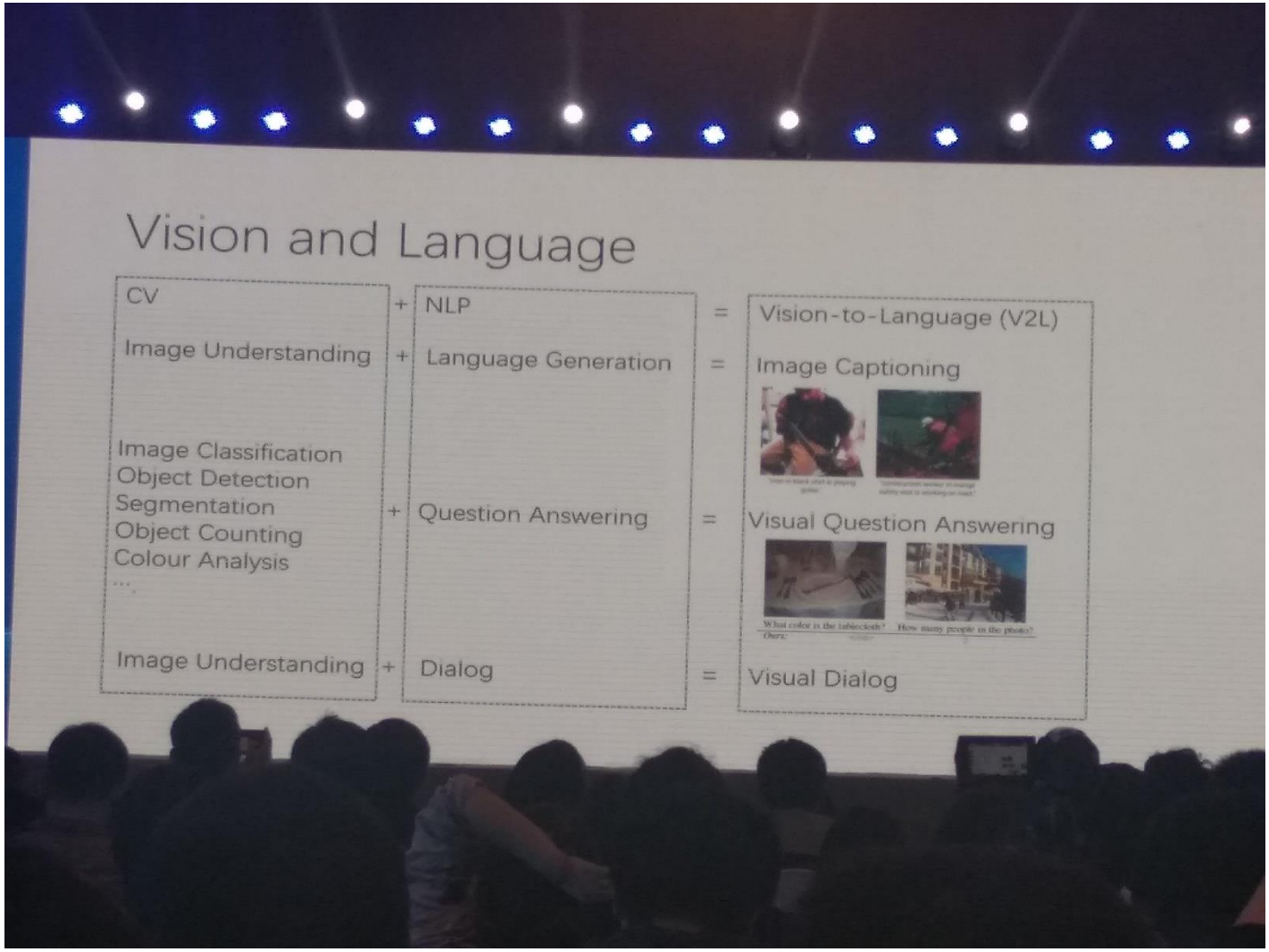
http://dmirlab.com

- About VALSE
- 弱监督视觉理解与主动学习
- 以人为中心的视觉理解
- 视觉中的知识推理
- CV+NLP 2.0
- Poster 选讲

CV+NLP

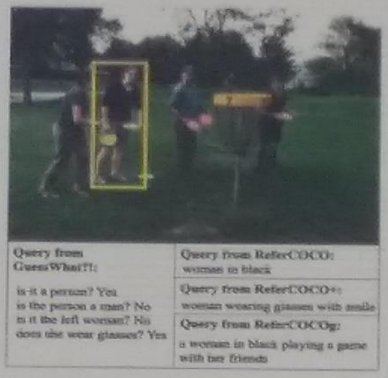

Vision-and-Language
2.0
Qi Wu (吴琦)
School of Computer Science
The University of Adelaide
Australia

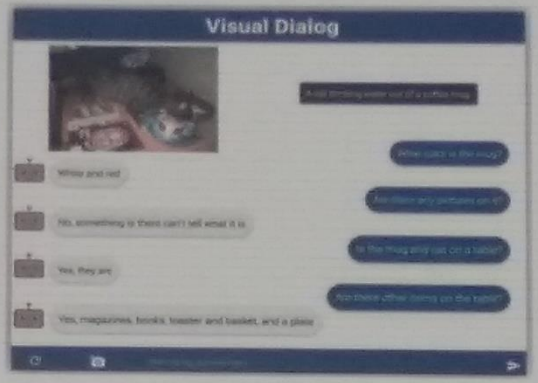
CV+NLP


Vision and Language
CV
+
NLP
=
Vision-to-Language (V2L)
Image Understanding
+
Language Generation
=
Image Captioning
Image Classification
Object Detection
Segmentation
Object Counting
Colour Analysis
...
+
Question Answering
=
Visual Question Answering
What color is the tablecloth?
How many people in the photo?
Image Understanding
+
Dialog
=
Visual Dialog

More Vision-and-Language Tasks

- Referring Expression
  - ReferIt Game, *EMNLP 2014*
  - RefCOCO, *ECCV 2016*
  - GuessWhat?!, *CVPR 2017*
- Visual Dialog
  - VisDial, *CVPR 2017*
  - Image Grounded Conversation, *ACL 2017*
  - Dialog-based Image Retrieval, *NIPS 2018*
- Text 2 image/video





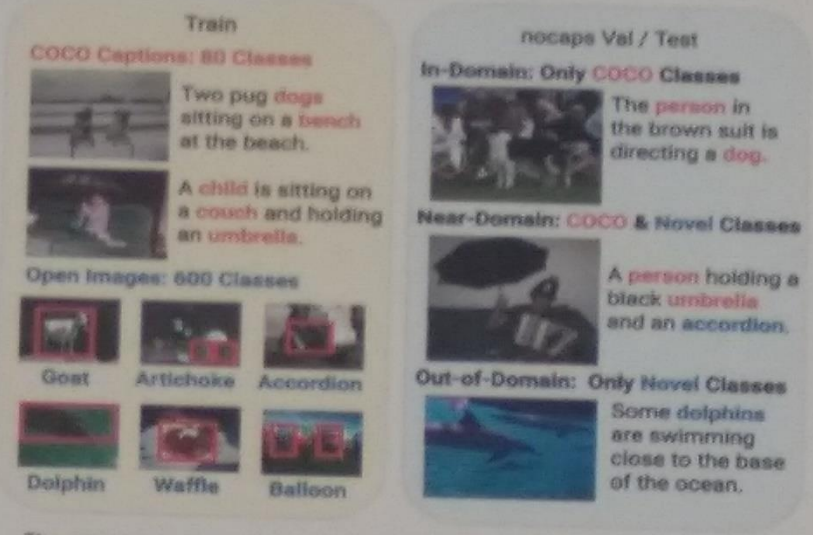
User1: My son is ahead and surprised!
User2: Did he end up winning the race?
User1: Yes he won, he can't believe it!

New Challenges – V&L 2.0

- Diversity and Controllable
  - Novel object captioning
  - Captioning with styles
- Reasoning
  - CLEVR dataset
  - GQA – CLEVR in the real world
  - Visual Commonsense Reasoning
- Embodied
  - Embodied VQA, Interactive QA
  - Language-guided visual navigation
  - Remote Embodied Referring Expression

Novel Object Captioning

- In-domain
  - objects have been described in the training data
- Near-domain
  - the most salient objects in the image are novel objects
- Out-of-domain
  - do not contain any COCO classes



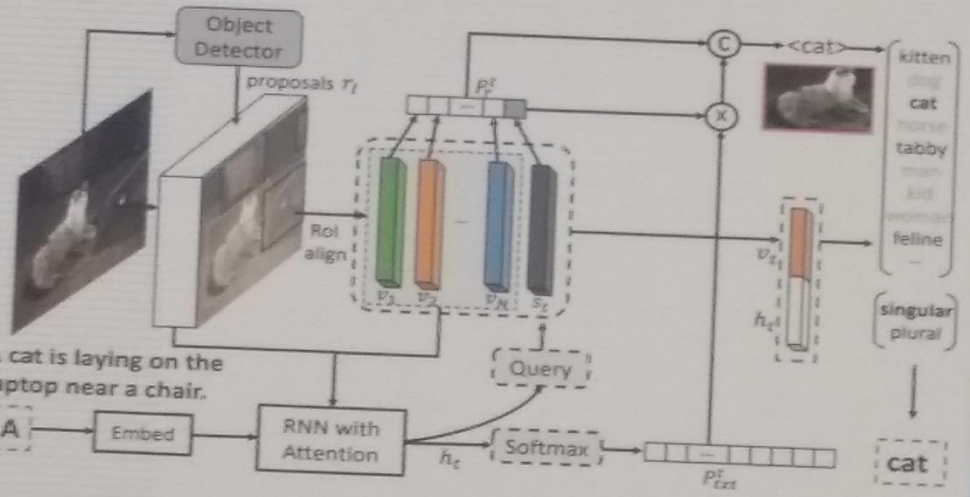
The **nocaps** benchmark for novel object captioning (at scale).

nocaps: novel object captioning at scale
Harsh Agrawal*, Karan Desai*, Xinlei Chen, [illegible] uv Batra, Devi Parikh, Stefan Lee, [illegible] Anderson, arXiv

Neural Baby Talk

'Template' generation
- at each time step, the model decides whether to generate a word from the textual vocabulary or generate a "visual" word.
- A <region-17> is sitting at a <region-123> with a <region-3>.

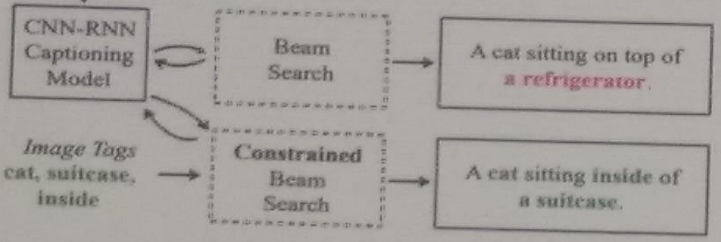
illing in the slots
- Classifies each of the indicated regions into object categories.
- A puppy is sitting at a table with a cake

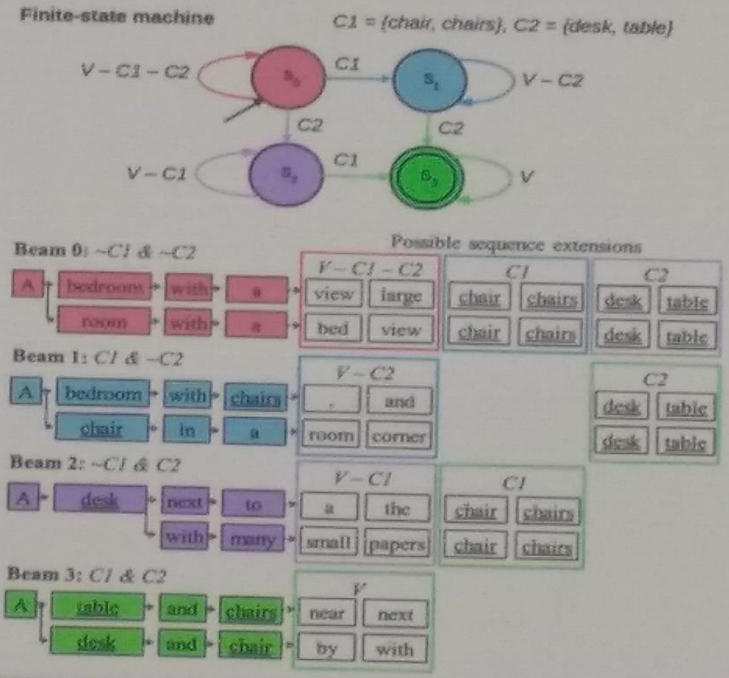


Lu, Jiasen, Jianwei Yang, Dhruv Batra, and Devi Parikh. "Neural baby talk." In *CVPR*. 2018.

Constrained beam search

Input image containing previously unseen object ('suitcase')





Anderson, Peter, Stephen Gould, and Mark Johnson. "Partially-Supervised Image Captioning." In NIPS. 2018.

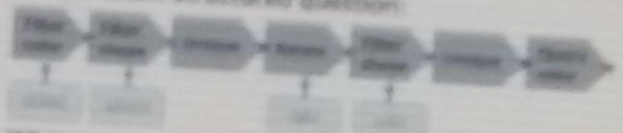
isual Reasoning - CLEVR

estions in CLEVR test various aspects of visual reasoning
uding **attribute identification**, **counting**, **comparison**, **spatial**
tionships, and **logical operations**.

equal number of large things and metal spheres?
the cylinder **that is left of** the brown metal thing **that is left of** the

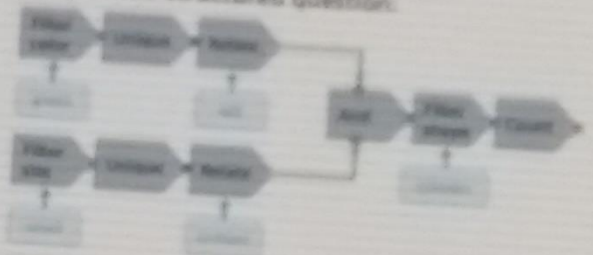
here with the **same size as** the metal cube; is it **made of the**
as the small red sphere?
bjects are **either** small cylinders or red things?

Sample chain-structured question:



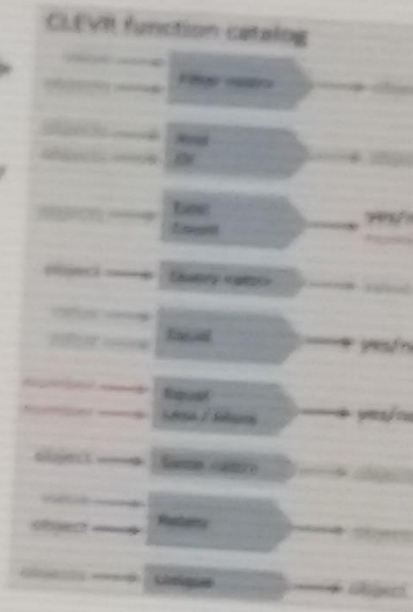
What color is the cube to the right of the yellow sphere?

Sample tree-structured question:



How many cylinders are in front of the tiny thing and on the left side of the green object?

CLEVR function catalog



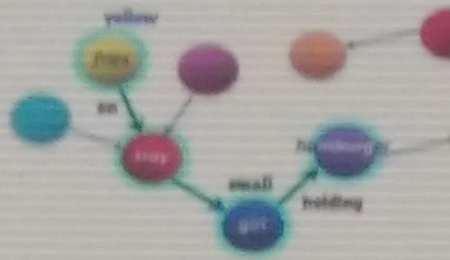
Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning." In *CVPR*. 2017.

GQA – Graph QA

Pattern: **What/Which** *(type) [do you think]* **(is)** *<objects, attrs> or <decoys>?*
Program: **Select:** <object> → **Choose** <type>:<attr>|<decoys>
Reference: *The food on the red object left of the small girl that is holding a hamburger*
Decoy: *brown*

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger → **Relate:** girl,holding → **Filter** size: small → **Relate:** object, left → **Filter** color: red → **Relate:** food,on → **Choose** color: yellow | brown



Graph Normalization

- Ontology construction
- Edge Pruning
- Object Augmentation
- Global Properties

Question Generation

- Patterns Collection
- Compositional References
- Decoys Selection
- Probabilistic Generation

Sampling and Balancing

- Distribution Balancing
- Type-Based Sampling
- Deduplication

Entailments Relations

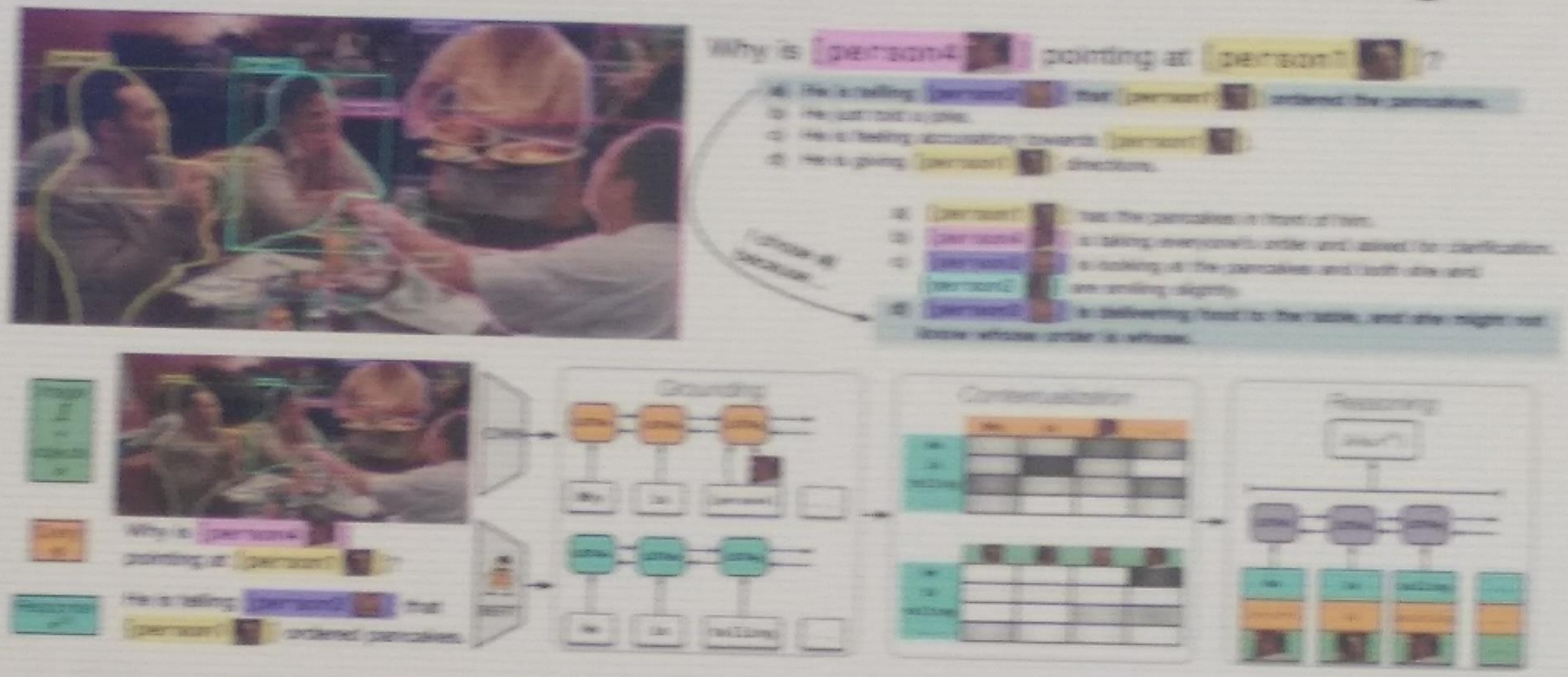
- Functional Programs
- Entailment Relations
- Recursive Reachability

New Metrics

- Consistency
- Validity & Plausibility
- Distribution
- Grounding

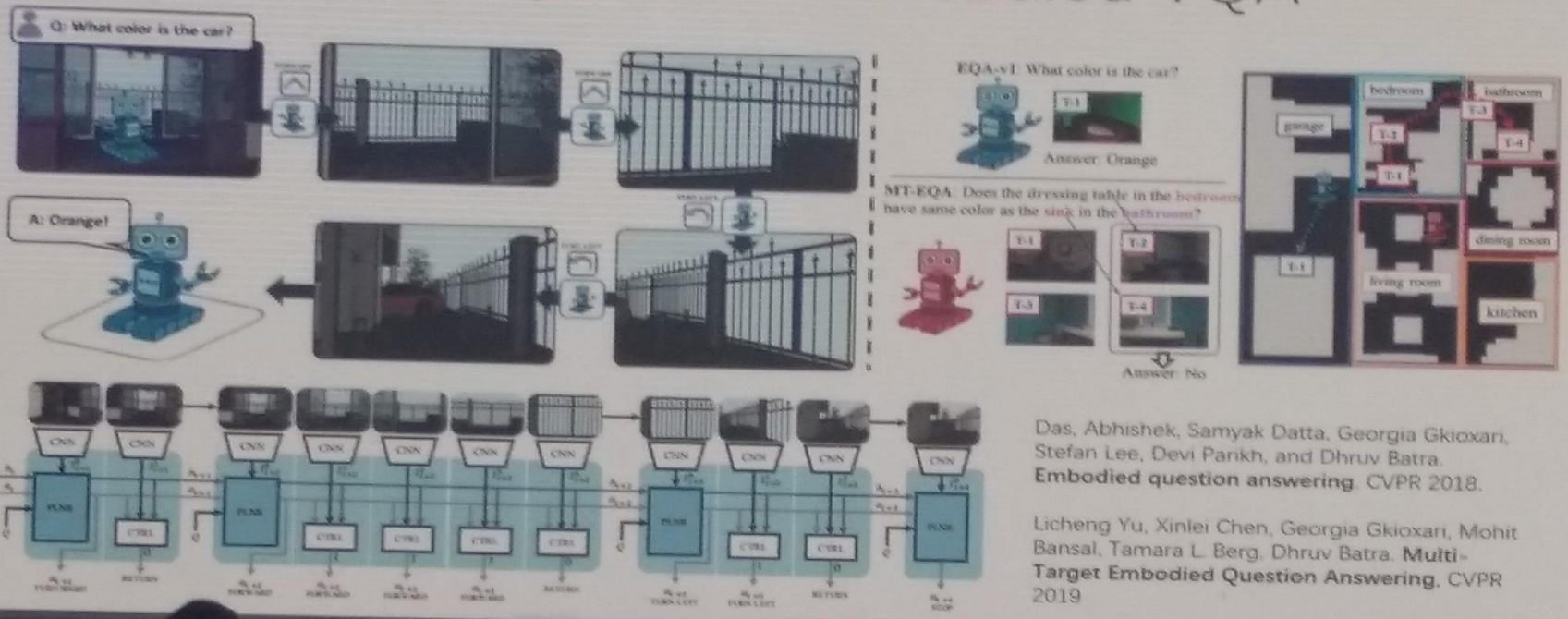
Hudson, Drew A., and Christopher D. Manning. "GQA: a new dataset for compositional question answering over real-world images." CVPR 2019

VCR – Visual Commonsense Reasoning



lers, Rowan, Yonatan Bisk, Ali Farhadi, and Yejin Choi. "From Recognition to Cognition: Visual Commonsense Reasoning." CVPR 2019

Embodied VQA & MT-Embodied VQA

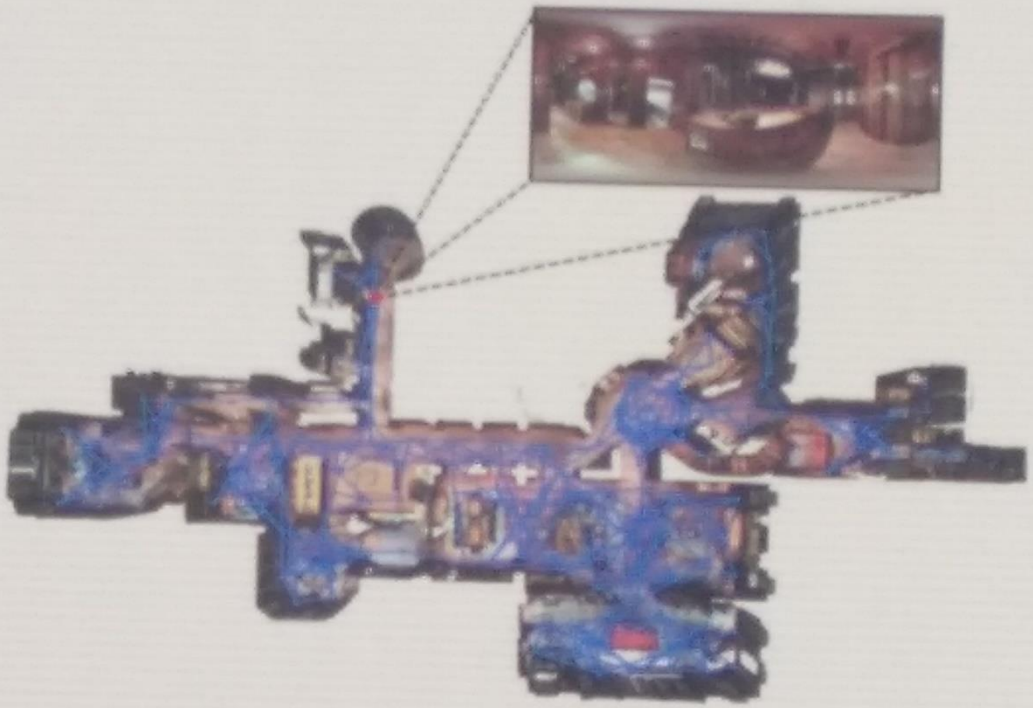


Das, Abhishek, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. **Embodied question answering**. CVPR 2018.

Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L. Berg, Dhruv Batra. **Multi-Target Embodied Question Answering**. CVPR 2019

n-and-Language Navigation

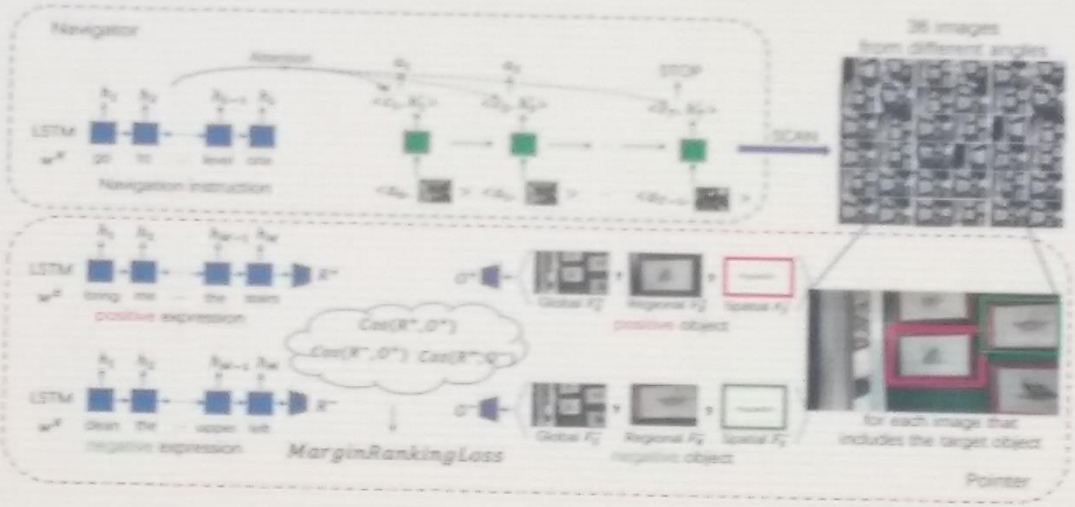
om. Turn left and exit the room
on the left. Wait there.



Anderson, Peter, **Qi Wu**, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments." In *CVPR* 2018.

Remote Embodied Referring Expression

ruction: Go to the stairs on level one and bring me the
om picture that is next to the top of the stairs.



Conclusions

- Limited improvements on classical image captioning and VQA
- Face to the real challenges!
  - Diversity
  - Controllable
  - Reasoning
  - Deploy the model in the real environment
  - Ask, Answer and Act
- Work on our **embodied visual-navigation & referring expression** tasks

- Two Fully-Funded International PhD positions!

Outline

- About VALSE
- 弱监督视觉理解与主动学习
- 以人为中心的视觉理解
- 视觉中的知识推理
- CV+NLP 2.0
- Poster 选讲

Poster

- IndRNN: Independently Recurrent Neural Networks (CVPR2018)
- FishNet: A Versatile Backbone for Image, Region, and Pixel Level Prediction (NIPS2018)
- HRNet: Deep High-Resolution Representation Learning for Human Pose Estimation (CVPR 2019, MSRA)

Poster

- GLCN: semi-supervised learning with graph learning-convolutional networks (CVPR2019)
- Data Representation and Learning with graph diffusion-embedding networks (CVPR 2019)
- multi-label image recognition with GCN (CVPR2019)
- two-stream adaptive GCN for skeleton-based action recognition (CVPR2019)

VALSE 2019
April 11 – 14, 2019

Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN

Shuai Li*†, Wanqing Li*, Chris Cook*, Ce Zhu†, Yanbo Gao†
*University of Wollongong, †University of Electronic Science and Technology of China

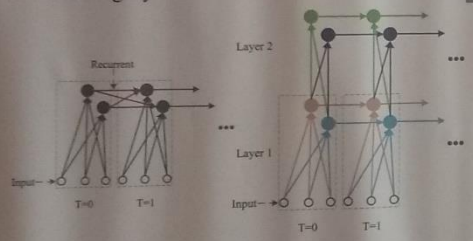
Problem

- **For RNN**: Gradient vanishing and exploding (**Short**)
- Repeated use of recurrent weight matrix
- **For LSTM**: Gradient decay over layers (**Shallow**)
- Usage of tanh and sigmoid functions

Common component: $\mathbf{U}\mathbf{h}_{t-1}$

Method: IndRNN

- $\boldsymbol{h}_t = \sigma(\boldsymbol{W}\boldsymbol{x}_t + \boldsymbol{u}\odot\boldsymbol{h}_{t-1} + \boldsymbol{b})$
- Independently Recurrent
  - For n-th neuron: $h_{n,t} = \sigma(\boldsymbol{w}_n\boldsymbol{x}_t + u_n h_{n,t-1} + b_n)$
- Exploration of cross-channel information in the following layers



Gradient Backpropagation Through Time (BPTT)

$$\frac{\partial J_n}{\partial h_{n,t}} = \frac{\partial J_n}{\partial h_{n,T}}\frac{\partial h_{n,T}}{\partial h_{n,t}} = \frac{\partial J_n}{\partial h_{n,T}} u_n^{T-t}\prod_{k=t}^{T-1}\sigma'_{n,k+1}u_n = \frac{\partial J_n}{\partial h_{n,T}} u_n^{T-t}\prod_{k=t}^{T-1}\sigma'_{n,k+1}$$

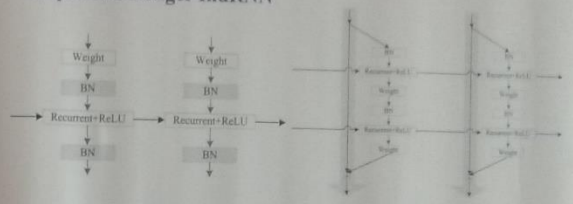
Gradient BPTT	
IndRNN	RNN
$\frac{\partial J_n}{\partial h_{n,T}} u_n^{T-t}\prod_{k=t}^{T-1}\sigma'_{n,k+1}$	$\frac{\partial J}{\partial h_T}\prod_{k=t}^{T-1}\sigma'_{n,k+1}U^T$
Scalar multiplication	Matrix multiplication
Robust training	Unstable (when U ill-conditioned)
Easy to regulate	Hard to control

Memory and recurrent weight

Memory	Gradient	Recurrent weight
Long-term	effective after T-t steps: $\frac{\partial J_n}{\partial h_{n,t}} > \varepsilon$	$\lvert u_n\rvert \in [\sqrt[(T-t)]{\frac{\varepsilon}{\prod_{k=t}^{T-1}\sigma'_{n,k+1}}}, +\infty)$
Short-term	reduced after T-t steps: $\frac{\partial J_n}{\partial h_{n,t}} \le \varepsilon$	$\lvert u_n\rvert \in [0, \sqrt[(T-t)]{\frac{\varepsilon}{\prod_{k=t}^{T-1}\sigma'_{n,k+1}}}]$
	not exploding: $\frac{\partial J_n}{\partial h_{n,t}} < \gamma$	$\lvert u_n\rvert \in [0, \sqrt[(T-t)]{\frac{\gamma}{\prod_{k=t}^{T-1}\sigma'_{n,k+1}}})$

- To keep long and short-term memory: $\lvert u_n\rvert \in [0, \sqrt[(T-t)]{\frac{\gamma}{\prod_{k=t}^{T-1}\sigma'_{n,k+1}}})$
- **Robust with ReLU**: $\lvert u_n\rvert \in [0, \sqrt[(T-t)]{\gamma})$
- **Easier to interpret**: Neurons independent, weight for range of memory

Deeper and Longer IndRNN

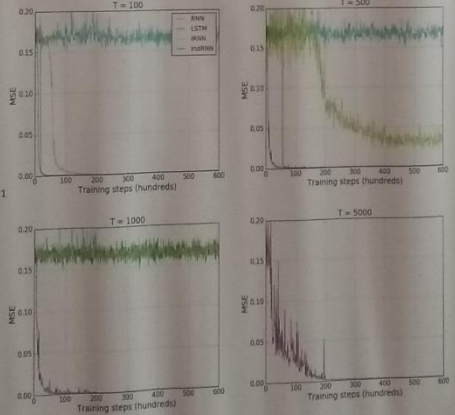


- Convolutional extension of IndRNN
  - $\boldsymbol{h}_t = \sigma(\boldsymbol{W} * \boldsymbol{x}_t + \boldsymbol{u}\odot\boldsymbol{h}_{t-1} + b)$
- **New perspective of RNN**
  - Independently aggregating spatial patterns (through $\boldsymbol{W}$) over time (through $\boldsymbol{u}$)

Results

1, Adding problem

- **Able to keep long-term memory(over 5000 steps)**



2, Sequential MNIST Classification

Table 1. Results (in terms of error rate (%)) for the sequential MNIST and permuted MNIST.

	MNIST	pMNIST
IRNN [26]	5.0	18
uRNN [1]	4.9	8.6
RNN-path [33]	3.1	-
LSTM [1]	1.8	12
LSTM+Recurrent dropout [38]	-	7.5
LSTM+Recurrent batchnorm [7]	-	4.6
LSTM+Zoneout [23]	-	6.9
LSTM+Recurrent batchnorm+Zoneout	-	4.1
IndRNN (6 layers)	1.0	4.0

3, Language Modelling

- **Very deep networks (21 layers)**

Table 2. Results of PTB-c for our proposed IndRNN model in comparison with results reported in the literature,in terms of BPC.

	Test
RNN-tanh [24]	1.55
RNN-relu [33]	1.55
RNN-TRec [24]	1.48
HF-MRNN [32]	1.42
RNN-path [33]	1.47
LSTM [23]	1.36
LSTM+Recurrent dropout [38]	1.32
LSTM+Recurrent batchnorm [7]	1.32
HyperLSTM + LN [11]	1.25
Hierarchical Multiscale LSTM + LN [6]	1.24
LSTM+Zoneout [23]	1.27
IndRNN (6 layers)	1.26
IndRNN (21 layers)	1.21

4, Skeleton based Action Recognition

- **Significant performance improvement**

Table 3. Results of all skeleton based methods on NTU RGB+D dataset.

Method	CS	CV
Deep learning on Lie Group [15]	61.37%	66.95%
JTM+CNN [42]	73.40%	75.20%
Res-TCN [21]	74.30%	83.10%
SkeletonNet(CNN) [19]	75.94%	81.16%
JDM+CNN [29]	76.20%	82.30%
Clips+CNN+MTLN [20]	79.57%	84.83%
Enhanced Visualization+CNN [31]	80.03%	87.21%
1 Layer RNN [39]	56.02%	60.24%
2 Layer RNN [39]	56.29%	64.09%
1 Layer LSTM [39]	59.14%	66.81%
2 Layer LSTM [39]	60.09%	67.29%
1 Layer PLSTM [39]	62.05%	69.40%
2 Layer PLSTM [39]	62.93%	70.27%
JL_d+RNN [45]	70.26%	82.39%
STA-LSTM [40]	73.40%	81.20%
ST-LSTM + Trust Gate [30]	69.20%	77.70%
Pose conditioned STA-LSTM[2]	77.10%	84.50%
IndRNN (4 layers)	78.58%	83.75%
IndRNN (6 layers)	81.80%	87.97%

Any questions, comments or suggestions are welcome!

Email: shuaili@uestc.edu.cn

Code: https://github.com/Sunnydreamrain/IndRNN_Theano_Lasagne
https://github.com/Sunnydreamrain/IndRNN_pytorch

roposed IndRNN

radient backpropagation through time (BPTT)

- For n-th neuron: $h_{n,t} = \sigma(w_n x_t + u_n h_{n,t-1} + b_n)$
- $\frac{\partial J_n}{\partial h_{n,t}} = \frac{\partial J_n}{\partial h_{n,T}} \frac{\partial h_{n,T}}{\partial h_{n,t}} = \frac{\partial J_n}{\partial h_{n,T}} \prod_{k=t}^{T-1} \sigma'_{n,k+1} u_n = \frac{\partial J_n}{\partial h_{n,T}} u_n^{T-t} \prod_{k=t}^{T-1} \sigma'_{n,k+1}$

To keep long and short-term memory: $|u_n| \in [0, \sqrt[(T-t)]{\frac{\gamma}{\prod_{k=t}^{T-1} \sigma'_{n,k+1}}})$

Easily solving the gradient exploding and vanishing problem

Robust with ReLU: $|u_n| \in [0, \sqrt[(T-t)]{\gamma})$

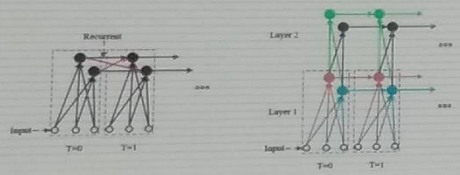
Easier to interpret: Neurons independent, weight for range of memory

New perspective of RNN

- Independently aggregating spatial patterns (through $\boldsymbol{W}$) over time (through $\boldsymbol{u}$)

Proposed Independently Recurrent Neural Networks (IndRNN)

- Form
  - $\boldsymbol{h}_t = \sigma(\boldsymbol{W}\boldsymbol{x}_t + \boldsymbol{u} \odot \boldsymbol{h}_{t-1} + \boldsymbol{b})$
  - Using element-wise product to process recurrent input
- Independently recurrent in one layer
  - For n-th neuron: $h_{n,t} = \sigma(\boldsymbol{w}_n x_t + u_n h_{n,t-1} + b_n)$
- Cross-channel information explored in the following layers



FishNet

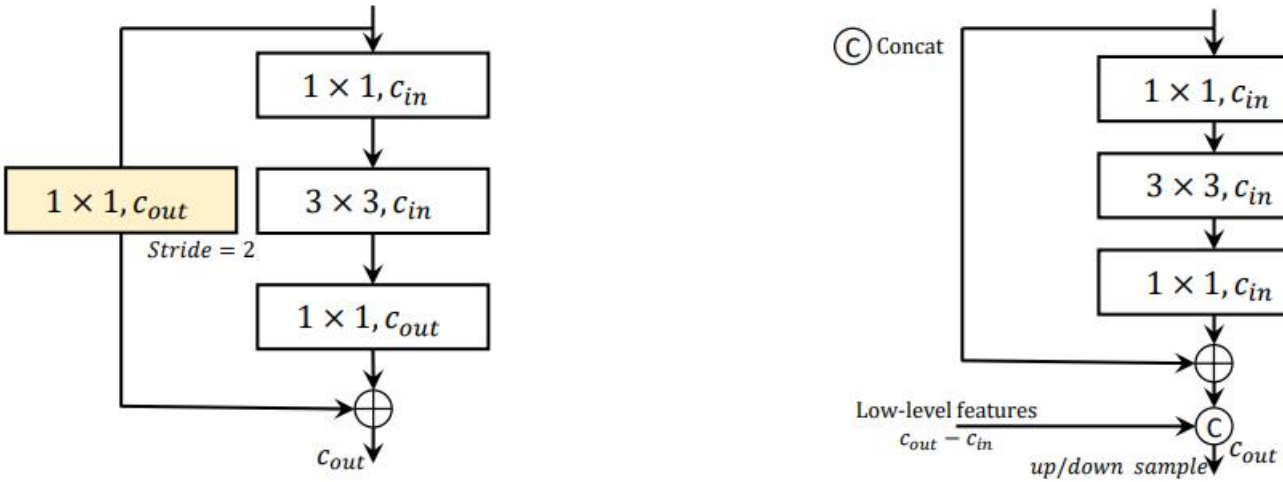


Figure 1: The up/down-sampling block for ResNet (*left*), and FishNet (*right*). The 1×1 convolution layer in yellow indicates the *Isolated convolution (I-conv, see Section 2)*, which makes the direct BP incapable and degrades the gradient from the output to shallow layers.

FishNet: A Versatile Backbone for Image, Region, and Pixel Level Prediction (NIPS2018)

FishNet

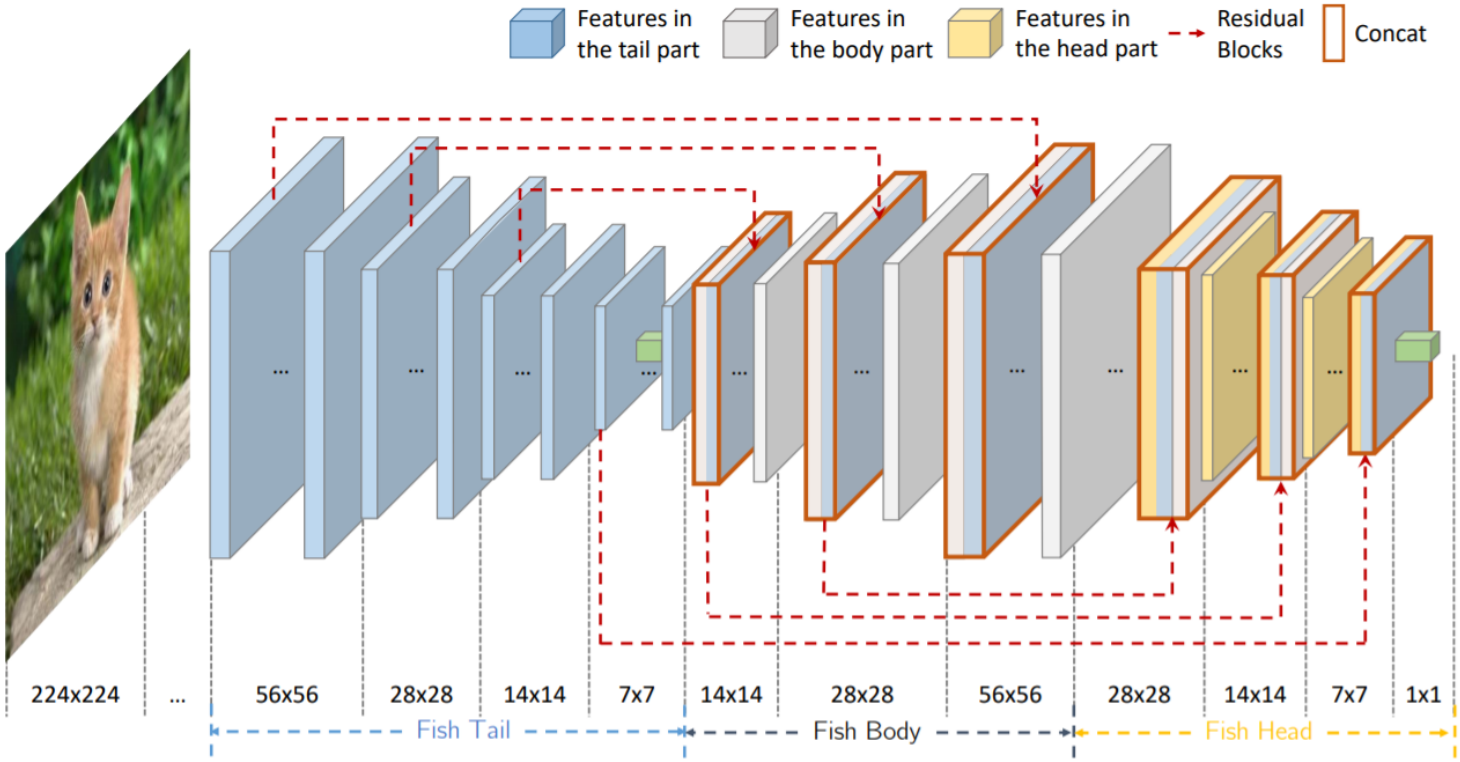
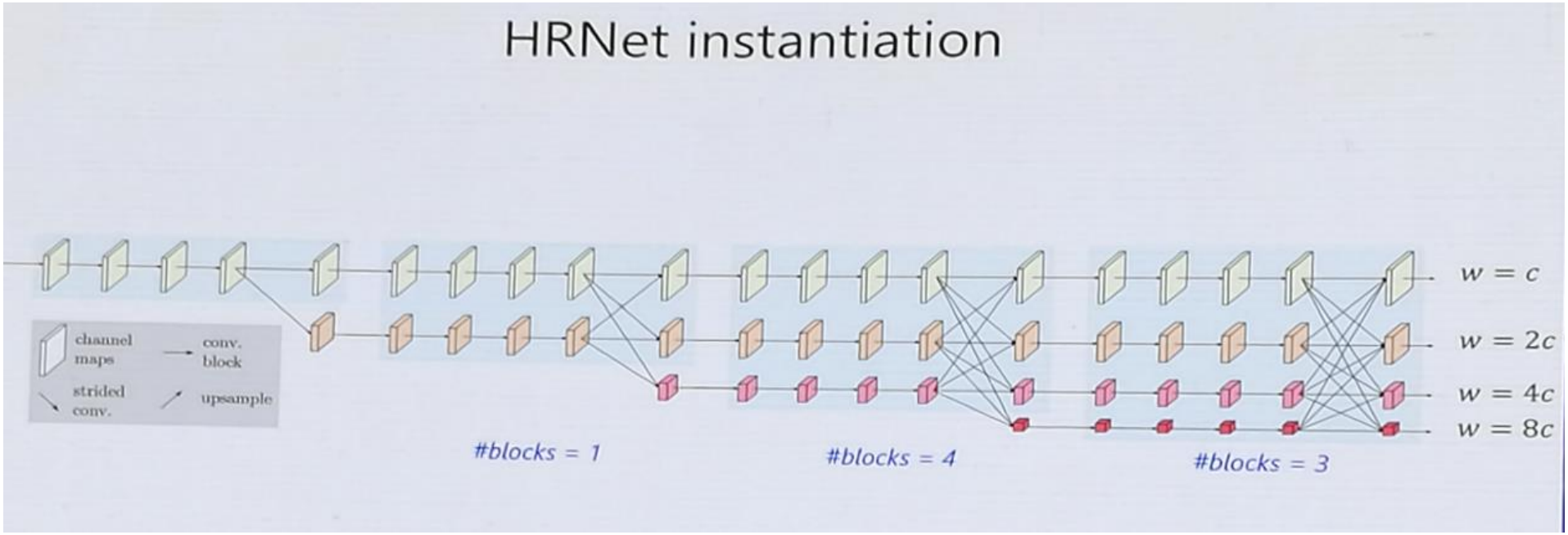


Figure 2: Overview of the FishNet. It has three parts. *Tail* uses existing works to obtain deep low-resolution features from the input image. *Body* obtains high-resolution features of high-level semantic information. *Head* preserves and refines the features from the three parts.

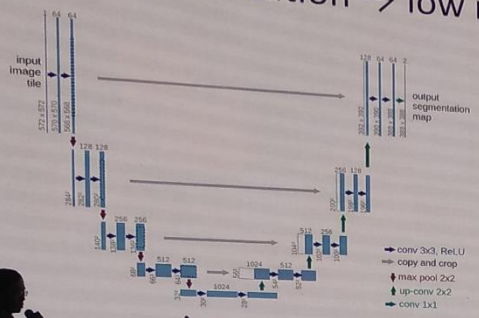
FishNet: A Versatile Backbone for Image, Region, and Pixel Level Prediction (NIPS2018)

HRNet

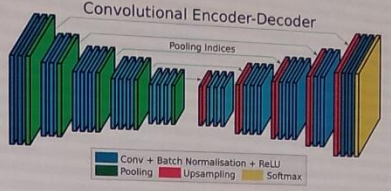


HRNet: Deep High-Resolution Representation Learning for Human Pose Estimation (CVPR 2019, MSRA)

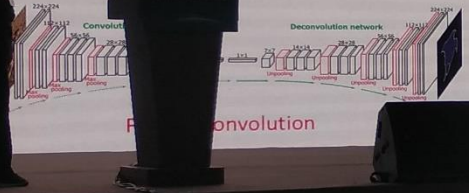
High resolution → low resolution → high resolution



U-Net



SegNet



Hourglass

6

VALSE 2019
April 11 – 14, 2019

Semi-supervised Learning with Graph Learning-Convolutional Networks

Author's Name/s: Bo jiang, Ziyan Zhang Address/es: Hefei, China

Introduction

Motivation

1. In many applications, the graph data are not available, and it is difficult to evaluate whether the graphs obtained from domain knowledge or established by human are optimal for semi-supervised learning in graph CNNs.
2. In other applications, the graph structure of data are available from domain knowledge, such as chemical molecules, social networks etc. But it may have some noise.

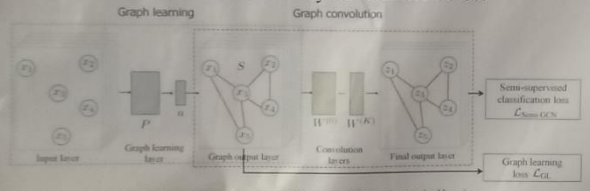


Figure 1. Architecture of the proposed GLCN network for semi-supervised learning.

Graph learning architecture

Given an input $X = (x_1, x_2 \cdots x_n) \in \mathbb{R}^{n \times p}$, we aim to seek a nonnegative function $S_{ij} = g(x_i, x_j)$ that represents the pairwise relationship between data x_i and x_j. We implement $g(x_i, x_j)$ via a single-layer neural network, which is parameterized by a weight vector $a = (a_1, a_2, \cdots a_p)^T \in \mathbb{R}^{p \times 1}$. Formally, we learn a graph S as

$$S_{ij} = g(x_i, x_j) = \frac{\exp(\text{ReLU}(a^T|x_i - x_j|))}{\sum_{i=1}^{n} \exp(\text{ReLU}(a^T|x_i - x_j|))} \quad (4)$$

$$\sum_{j=1}^{n} S_{ij} = 1, S_{ij} \geq 0 \quad (5)$$

We optimize the optimal weight vector a by minimizing the following loss function,

$$\mathcal{L}_{\text{GL}} = \sum_{i,j=1}^{n} \|x_i - x_j\|_2^2 S_{ij} + \gamma \|S\|_F^2 \quad (6)$$

Introduction

For some problems, when an initial graph A is available, we can incorporate it in our graph learning as

$$S_{ij} = g(x_i, x_j) = \frac{A_{ij} \exp(\text{ReLU}(a^T|x_i - x_j|))}{\sum_{j=1}^{n} A_{ij} \exp(\text{ReLU}(a^T|x_i - x_j|))} \quad (7)$$

We can also incorporate the information of A by considering a regularized term in the learning loss function as

$$\mathcal{L}_{\text{GL}} = \sum_{i,j=1}^{n} \|x_i - x_j\|_2^2 S_{ij} + \gamma \|S\|_F^2 + \beta \|S - A\|_F^2 \quad (8)$$

GLCN architecture

As shown in Figure 1, GLCN contains one graph learning layer, several graph convolution layers and one final perceptron layer. The graph learning layer aims to provide an optimal adaptive graph representation S for graph convolutional layers. That is, in the convolutional layers, it conducts the layer-wise propagation rule based on the adaptive neighbor graph S returned by graph learning layer, i.e.,

$$X^{(k+1)} = \sigma(S X^{(k)} W^{(k)}) \quad (13)$$

The original GCN network are trained by minimizing the cross-entropy loss function as,

$$\mathcal{L}_{\text{Semi-GCN}} = -\sum_{i \in L} \sum_{j=1}^{c} Y_{ij} \ln Z_{ij} \quad (3)$$

The whole network parameters are jointly trained by minimizing the following loss function as

$$\mathcal{L}_{\text{Semi-GLCN}} = \mathcal{L}_{\text{Semi-GCN}} + \lambda \mathcal{L}_{\text{GL}} \quad (15)$$

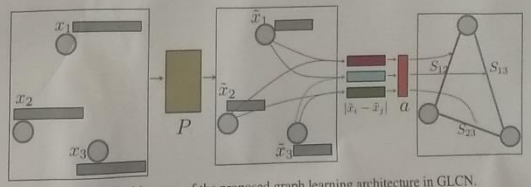


Figure 2. Architecture of the proposed graph learning architecture in GLCN.

Introduction

Result

Table 1. Comparison results of semi-supervised learning on dataset Citeseer, Cora and Pubmed.

Methond	Citeseer	Cora	Pubmed
ManiReg [2]	60.1%	59.5%	70.7%
LP [23]	45.3%	68.0%	63.0%
DeepWalk [18]	43.2%	67.2%	65.3%
GCN [11]	68.9%	82.9%	77.9%
GAT [21]	71.0%	83.2%	78.0%
GLCN	**72.0%**	**85.5%**	**78.3%**

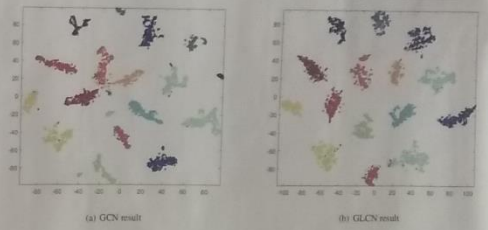


Figure 1. 2D t-SNE visualizations of the feature map output by the first convolutional layer of GCN and GLCN respectively on Scene15 dataset. Different classes are marked by different colors.

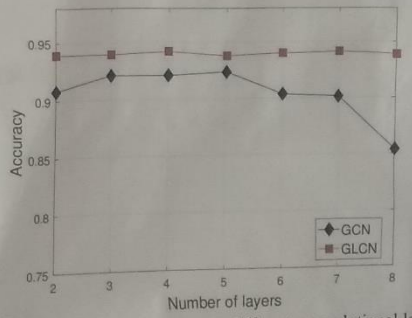
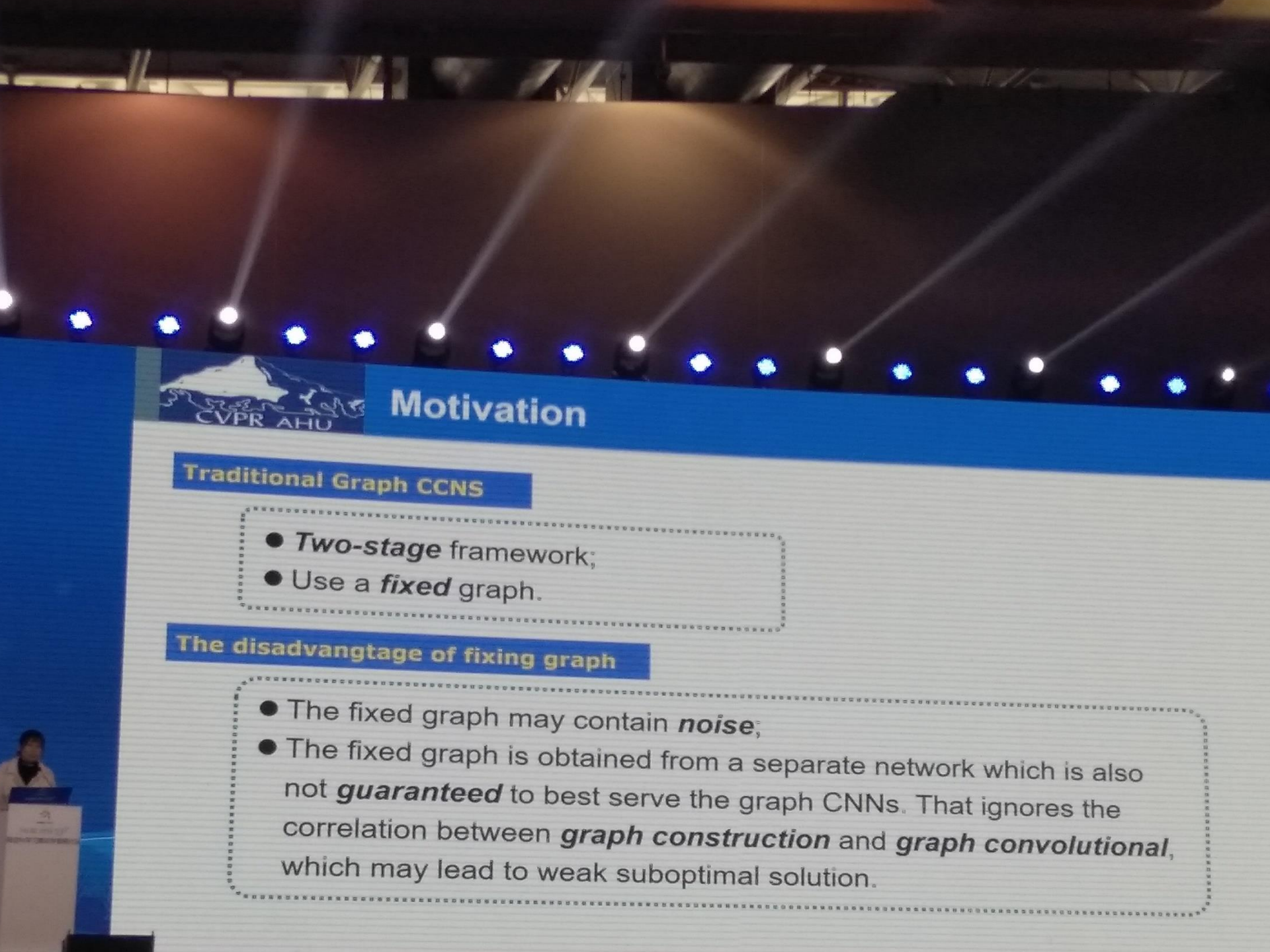


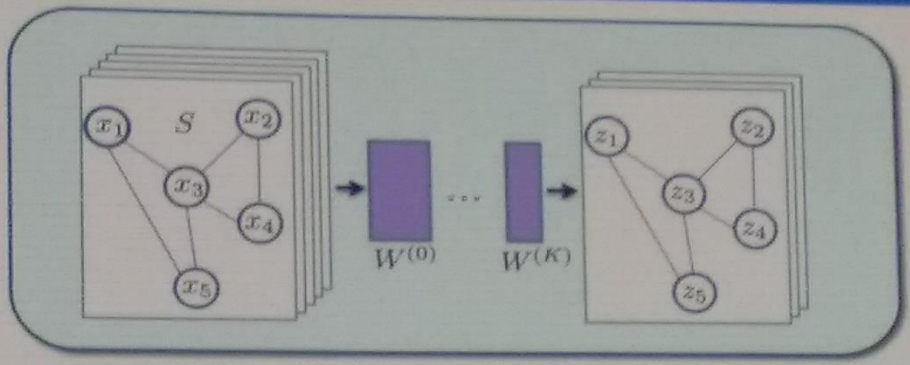
Figure 2. Results of GLCN across different convolutional layers on MNIST dataset.


CVPR AHU
Motivation
Traditional Graph CCNS
Two-stage framework;
Use a fixed graph.
The disadvangtage of fixing graph
The fixed graph may contain noise;
The fixed graph is obtained from a separate network which is also not guaranteed to best serve the graph CNNs. That ignores the correlation between graph construction and graph convolutional, which may lead to weak suboptimal solution.

The Overall Architecture

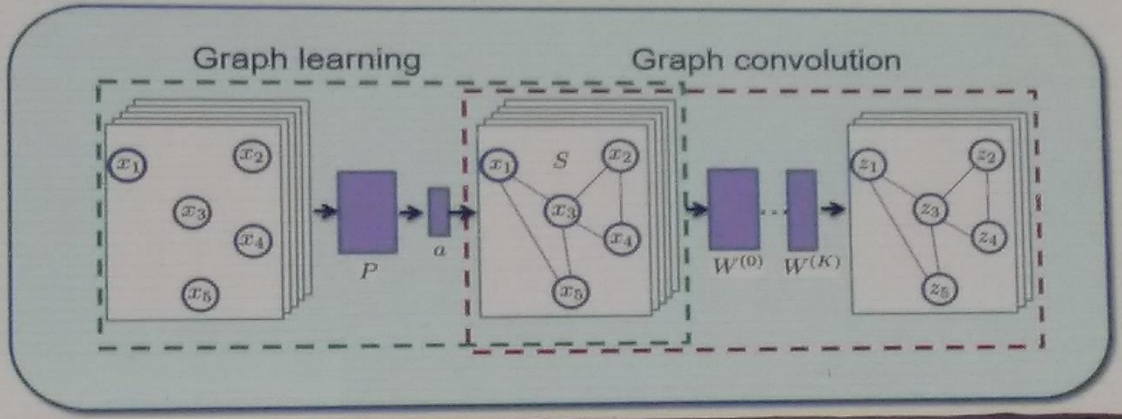
GCN

Graph convolution



GLCN

Graph learning + Graph convolution

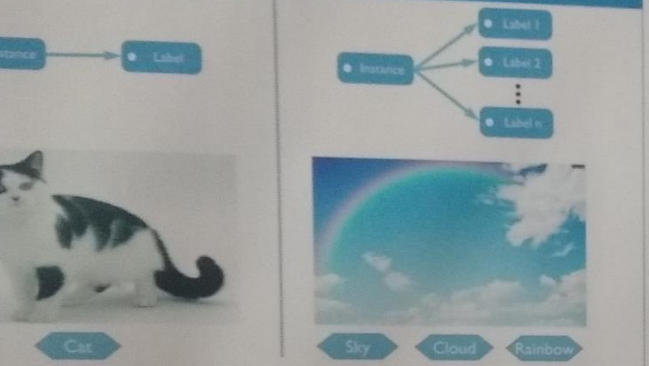


Multi-Label Image Recognition with Graph Convolutional Networks

Zhao-Min Chen[1,2], Xiu-Shen Wei[2*], Peng Wang[3], Yanwen Guo[1*]

[1]Nanjing University, [2]Megvii Research Nanjing, [3]The University of Adelaide

is Multi-label Image Recognition?



label image recognition is an important and cal task in computer vision and pattern nition, where the aim is to predict a set of ts present in an image.

tributions

e propose a novel end-to-end multi-label image cognition framework (ML-GCN), which exploits CN to map a label graph into a set of label-pendent object classifiers.

/e propose a data-dependent scheme to build the aph over the object labels and have the in-depth udy on the correlation matrix construction.

/e comprehensively evaluate our method on two andard multi-label image recognition datasets, and ur proposed method consistently achieves uperior performance over state-of-the-art pproaches.

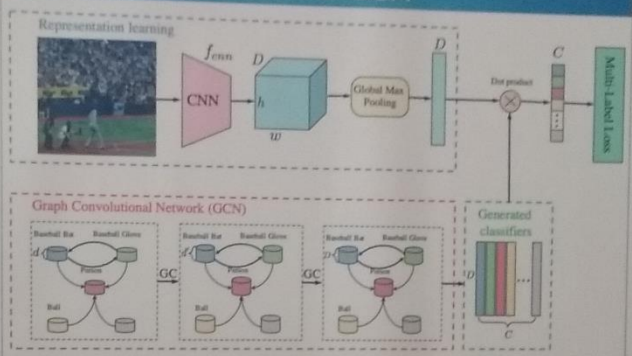
n sources

er

https://github.com/chenzhaomin123/ML_GCN

Code

The proposed method: ML-GCN



➢ Graph Convolution:

$$H^{l+1} = h(\widehat{A} H^l W^l)$$

Graph Convolution takes feature descriptions H^l and adjacency matrix A as input, then outputs node-level feature descriptions H^{l+1}.

➢ Correlation Matrix:

$$P_i = M_i / N_i$$

Obtaining the conditional probability matrix.

$$A_{ij} = \begin{cases} 0, & \text{if } P_{ij} < \tau \\ 1, & \text{if } P_{ij} \geq \tau \end{cases}$$

Filtering noisy edges and convert the conditional probability matrix into binary.

$$A'_{ij} = \begin{cases} p / \sum_{j=1, i \neq j}^{C} A_{ij}, & \text{if } i \neq j \\ 1 - p, & \text{if } i = j \end{cases}$$

Re-weighting binary matrix to alleviate over-smoothing problem.

➢ Multi-label classification loss:

$$\mathcal{L} = \sum_{c=1}^{C} y^c \log(\sigma(\hat{y}^c)) + (1 - y^c) \log(1 - \sigma(\hat{y}^c))$$

Experiments

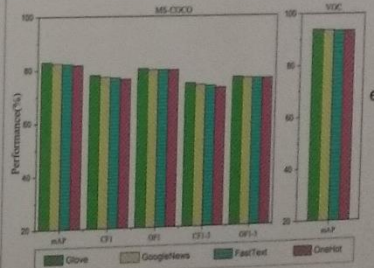
MS-COCO Dataset:

Methods	All							Top-3					
	mAP	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
CNN-RNN [28]	61.2	–	–	–	–	–	–	66.0	55.6	60.4	69.2	66.4	67.8
RNN-Attention [29]	–	–	–	–	–	–	–	79.1	58.7	67.4	84.0	63.0	72.0
Order-Free RNN [1]	–	–	–	–	–	–	–	71.6	54.8	62.1	74.2	62.2	67.7
ML-ZSL [15]	–	–	–	–	–	–	–	74.1	**64.5**	69.0	–	–	–
SRN [35]	77.1	81.6	65.4	71.2	82.7	69.9	75.8	85.2	58.8	67.4	87.4	62.5	72.9
ResNet-101 [10]	77.3	80.2	66.7	72.8	83.9	70.8	76.8	84.1	59.4	69.7	89.1	62.8	73.6
Multi-Evidence [6]	–	80.4	70.2	74.9	85.2	72.5	78.4	84.5	62.2	70.6	89.1	64.3	74.7
ML-GCN (Binary)	80.3	81.1	70.1	75.2	83.8	74.2	78.7	84.9	61.3	71.2	88.8	65.2	75.2
ML-GCN (Re-weighted)	**83.0**	**85.1**	**72.0**	**78.0**	**85.8**	**75.4**	**80.3**	**89.2**	**64.1**	**74.6**	**90.5**	**66.5**	**76.7**

PASCAL VOC 2007 Dataset:

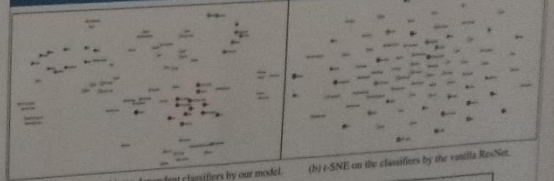
Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
CNN-RNN [28]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	91.7	59.8	93.2	75.3	99.7	78.6	84.0
RLSD [33]	96.4	92.7	93.8	94.1	71.2	92.1	94.2	95.7	74.3	90.0	74.2	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.1	87.6	88.5
VeryDeep [26]	98.9	95.0	96.8	95.4	69.7	90.4	93.5	96.0	74.2	86.6	**87.8**	96.0	96.3	93.1	97.2	70.0	92.1	80.3	98.1	87.0	89.7
ResNet-101 [10]	99.5	97.7	97.8	96.4	65.7	91.8	96.1	97.6	74.2	80.9	85.0	**98.4**	96.5	95.9	98.4	70.1	88.3	80.2	98.9	89.2	89.9
FeV+LV [32]	97.9	97.0	96.6	94.6	73.6	93.9	96.5	95.5	73.7	90.3	82.8	95.4	97.7	95.9	98.6	77.6	88.7	78.0	98.3	89.0	90.6
HCP [30]	98.6	97.1	98.0	95.6	75.3	**94.7**	95.8	97.3	73.1	90.2	80.0	97.3	96.1	94.9	96.3	78.3	94.7	76.2	97.9	91.5	90.8
RNN-Attention [29]	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	87.7	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
Atten-Reinforce [3]	98.6	97.1	97.1	95.5	75.6	92.8	96.8	97.5	78.3	92.2	87.6	96.9	96.5	93.6	98.5	81.6	93.1	81.2	98.5	89.3	92.0
VGG (Binary)	98.1	97.1	96.1	96.5	75.0	91.4	95.8	95.4	76.7	92.1	85.1	96.7	96.0	95.7	97.8	77.4	93.1	79.7	97.9	89.3	91.1
VGG (Re-weighted)	99.4	97.4	98.0	97.0	77.9	92.4	96.8	97.8	80.8	93.4	87.2	98.0	97.3	95.8	98.8	79.4	95.3	82.2	99.1	91.4	92.8
ML-GCN (Binary)	99.6	98.3	97.9	97.6	78.2	92.3	97.4	97.4	79.2	94.4	86.5	97.4	97.9	97.1	98.7	84.6	95.3	83.0	98.6	90.4	93.1
ML-GCN (Re-weighted)	99.5	98.5	98.6	98.1	80.8	94.0	97.2	98.2	82.3	95.7	88.4	98.7	98.4	98.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0

Effects of different word embedding approaches



Different word embedding approaches will hardly affect the final recognition accuracy.

Classifier Visualization



(a) t-SNE on the learned inter-dependent classifiers by our model. (b) t-SNE on the classifiers by the vanilla ResNet.

VALSE

VALSE 2019

April 11 – 14, 2019

Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition

Lei Shi, Yifan Zhang, Jian Cheng, Hanqing Lu

Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Introduction

- In existing GCN-based methods for skeleton-based action recognition task, the topology of the graph is set manually, and it is fixed over all layers and input samples. This may not be optimal for the hierarchical GCN and diverse samples in action recognition tasks. In this work, an adaptive graph convolutional network is proposed to adaptively learn the topology of the graph for different GCN layers and skeleton samples in an end-to-end manner, which can better suit the action recognition task and the hierarchical structure of the GCNs.
- The second-order information (the lengths and directions of bones) of the skeleton data, which is naturally more informative and discriminative for action recognition, is rarely investigated in existing methods. In this work, the second-order information of the skeleton data is explicitly formulated and combined with the first-order information using a two-stream framework, which brings notable improvement for the recognition performance.
- On two large-scale datasets for skeleton-based action recognition, the proposed 2s-AGCN exceeds the state-of-the-art by a significant margin. Code: https://github.com/lshiwjx/2s-AGCN.

Proposed Methods

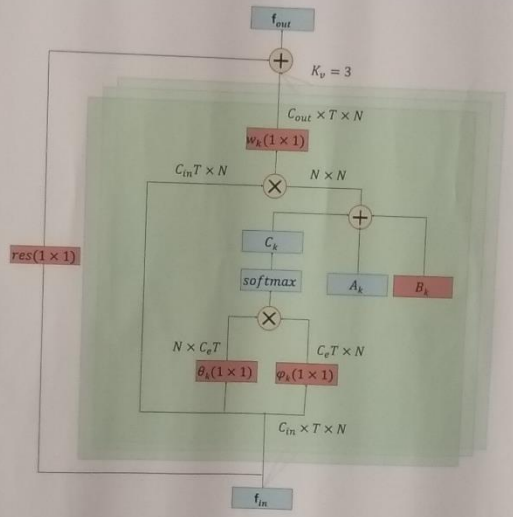
Graph Convolutional Network (GCN)

$$f_{out}(v_i) = \sum_{v_j \in \mathcal{B}_i} \frac{1}{Z_{ij}} f_{in}(v_j) \cdot w(l_i(v_j))$$

$$\mathbf{f}_{out} = \sum_k^{K_v} \mathbf{W}_k (\mathbf{f}_{in} \mathbf{A}_k) \odot \mathbf{M}_k$$

Proposed Methods

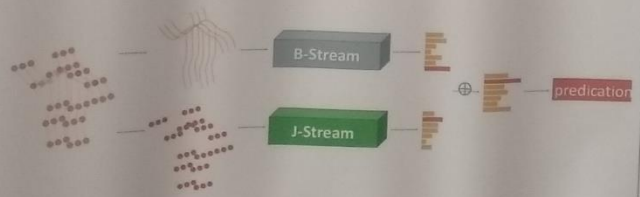
Adaptive Graph Convolutional Network (AGCN)



$$\mathbf{f}_{out} = \sum_k^{K_v} \mathbf{W}_k \mathbf{f}_{in} (\mathbf{A}_k + \mathbf{B}_k + \mathbf{C}_k)$$

$$\mathbf{C}_k = softmax(\mathbf{f_{in}}^T \mathbf{W}_{\theta k}^T \mathbf{W}_{\phi k} \mathbf{f_{in}})$$

Two-Stream Network



Experiments

Preprocessing methods

- Select two bodies in each sample according to the energy of the body.
- Subtract the coordinates of the center joint
- Rotate the body joints to a fixed viewpoints

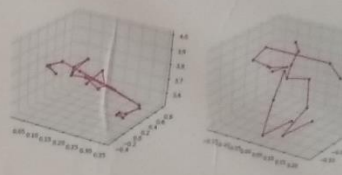


Figure 2. Example of the data preprocessing on the NTU- dataset. The left is the original skeleton, and the right is t processed skeleton.

Results on NTU-RGB-D

Methods	Accuracy (%)
ST-GCN	92.7
ST-GCN wo/M	91.1
AGCN wo/A	93.4
AGCN wo/B	93.3
AGCN wo/C	93.4
AGCN	93.7

Methods	Accuracy (%)
Js-AGCN	93.7
Bs-AGCN	93.2
2s-AGCN	95.1

Visualization

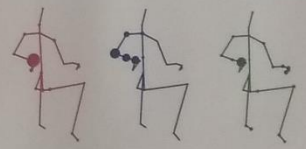


Figure 8. Visualization of the graphs for different layers.

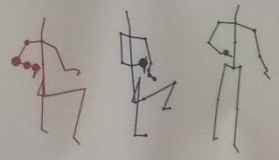


Figure 9. Visualization of the graphs for different samples.

VALSE

VALSE 2019

April 11 – 14, 2019

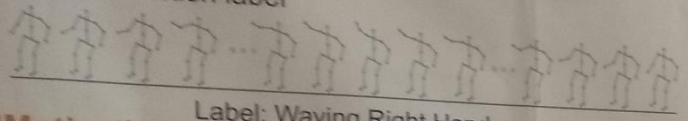
CVPR LONG BEACH CALIFORNIA June 16-20, 2019

An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition

Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, Tieniu Tan
Institute of Automation, Chinese Academy of Sciences

Overview

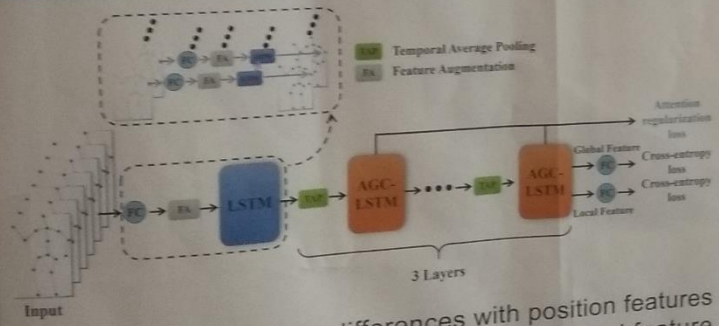
Task: Skeleton-based action recognition
- **Input:** Skeleton sequence
- **Output:** Action label



Label: Waving Right Hand

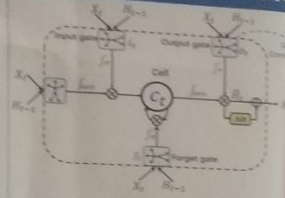
Motivation:
- The skeleton data contains abundant spatial structure information.
- Temporal continuity exists not only in the same joints (e.g., hand, wrist and elbow), but also in the body structure.
- There is a co-occurrence relationship between spatial and temporal domains.

Network Architecture



1. FA computes feature differences with position features and concatenates both position features and feature differences.
2. LSTM is used to dispel scale variance between feature differences and position features.
3. Three AGC-LSTM layers can model discriminative spatial-temporal features.
4. Temporal average pooling is the implementation of average pooling in the temporal domain.

AGC-LSTM



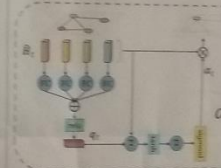
$$\mathbf{i}_t = \sigma\left(\mathbf{W}_{xi} *_\mathcal{G} \mathbf{X}_t + \mathbf{W}_{hi} *_\mathcal{G} \mathbf{H}_{t-1} + \mathbf{b}_i\right)$$

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_{xf} *_\mathcal{G} \mathbf{X}_t + \mathbf{W}_{hf} *_\mathcal{G} \mathbf{H}_{t-1} + \mathbf{b}_f\right)$$

$$\mathbf{o}_t = \sigma\left(\mathbf{W}_{xo} *_\mathcal{G} \mathbf{X}_t + \mathbf{W}_{ho} *_\mathcal{G} \mathbf{H}_{t-1} + \mathbf{b}_o\right)$$

$$\mathbf{u}_t = tanh\left(\mathbf{W}_{xc} *_\mathcal{G} \mathbf{X}_t + \mathbf{W}_{hc} *_\mathcal{G} \mathbf{H}_{t-1} + \mathbf{b}_c\right) \quad \mathbf{C}_t = \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \mathbf{u}_t$$

$$\hat{\mathbf{H}}_t = \mathbf{o}_t \odot tanh(\mathbf{C}_t) \qquad \mathbf{H}_t = f_{att}(\hat{\mathbf{H}}_t) + \hat{\mathbf{H}}_t$$



$$\mathbf{q}_t = Relu\left(\sum_{i=1}^{N} \mathbf{W}\hat{\mathbf{H}}_{ti}\right)$$

$$\alpha_t = Sigmoid\left(\mathbf{U}_s tanh\left(\mathbf{W}_h\hat{\mathbf{H}}_t + \mathbf{W}_q\mathbf{q}_t + \mathbf{b}_s\right) + \mathbf{b}_u\right)$$

$$\mathbf{H}_{ti} = (1 + \alpha_{ti}) \cdot \hat{\mathbf{H}}_{ti}$$

The structures of AGC-LSTM unit. Compared with LSTM, the inner operator of AGC-LSTM is graph convolutional calculation. To highlight more discriminative information, the attention mechanism is employed to enhance the features of key nodes.

Learning

$$\mathcal{L} = -\sum_{t=1}^{T_3}\sum_{i=1}^{C} y_i log\hat{y}_{ti}^{g} - \sum_{t=1}^{T_3}\sum_{i=1}^{C} y_i log\hat{y}_{ti}^{l}$$

$$+\lambda\sum_{j=1}^{3}\sum_{n=1}^{N}\left(1 - \frac{\sum_{t=1}^{T_j}\alpha_{tnj}}{T_j}\right)^2 + \beta\sum_{j=1}^{3}\frac{1}{T_j}\sum_{t=1}^{T_j}\left(\sum_{n=1}^{N}\alpha_{tnj}\right)^2$$

The third term aims to pay equal attention to different joints. The last term is to limit the number of interested nodes.

Conclusion

- AGC-LSTM can not only capture discriminative features in spatial configuration and temporal dynamics but also explore the co-occurrence relationship between spatial and temporal domains.
- The temporal hierarchical architect boosts the capability to learn the high-level semantic representation and significantly reduces the computation cost.
- The attention mechanism can enhance information of key joints in each AGC-LSTM layer.

Results

- The comparison results on the NTU dataset

Methods	Cross-View	Cross-Subject
VA-LSTM (2017)	87.6	79.4
ST-GCN (2018)	88.3	81.5
SR-TSL (2018)	92.4	84.8
HCN (2018)	91.1	86.5
PB-GCN (2018)	93.2	87.5
AGC-LSTM (Joint)	**93.5**	**87.5**
AGC-LSTM (Part)	**93.8**	**87.5**
AGC-LSTM (Joint&part)	**95.0**	**89.2**

- Ablation study on the NTU dataset

	Methods	Cross-View	Cross-Subject
Joint	LSTM	89.4	80.3
	GC-LSTM	92.4	85.6
	LSTM+TH	90.4	81.4
	GC-LSTM+TH	92.9	86.3
	AGC-LSTM	93.5	87.5
Part	AGC-LSTM	93.8	87.5
	AGC-LSTM (Joint&part)	**95.0**	**89.2**

- Comparison results on the Northwestern-UCLA dataset

Methods	Accuracy
Actionlet ensemble(2014)	76.0
HBRNN-L(2015)	78.5
Visualization CNN(2017)	86.1
Ensemble TS-LSTM(2017)	89.2
AGC-LSTM (Joint)	**92.2**
AGC-LSTM (Part)	**90.1**
AGC-LSTM (Joint&part)	**93.3**

- Ablation study on the Northwestern-UCLA dataset

	Methods	Accuracy
Joint	LSTM	70.0
	GC-LSTM	87.5
	LSTM+TH	78.5
	GC-LSTM+TH	89.4
	AGC-LSTM	92.2
Part	AGC-LSTM	90.1
	AGC-LSTM (Joint&part)	93.3

- Visualizations of the attention weights of three AGC-LSTM layers on one actor of the action "handshaking"

Vertical axis denotes the joints. Horizontal axis denotes the frames. (a), (b), (c) are the attention results of the first, second and third AGC-LSTM layer, respectively.

Thank you !

Q&A