AIDC

INTEL AI DEVCON 2018

AIDC
INTEL AI DEVCON 2018

MAKING SENSE OF VISUAL DATA

Yurong Chen Ph.D

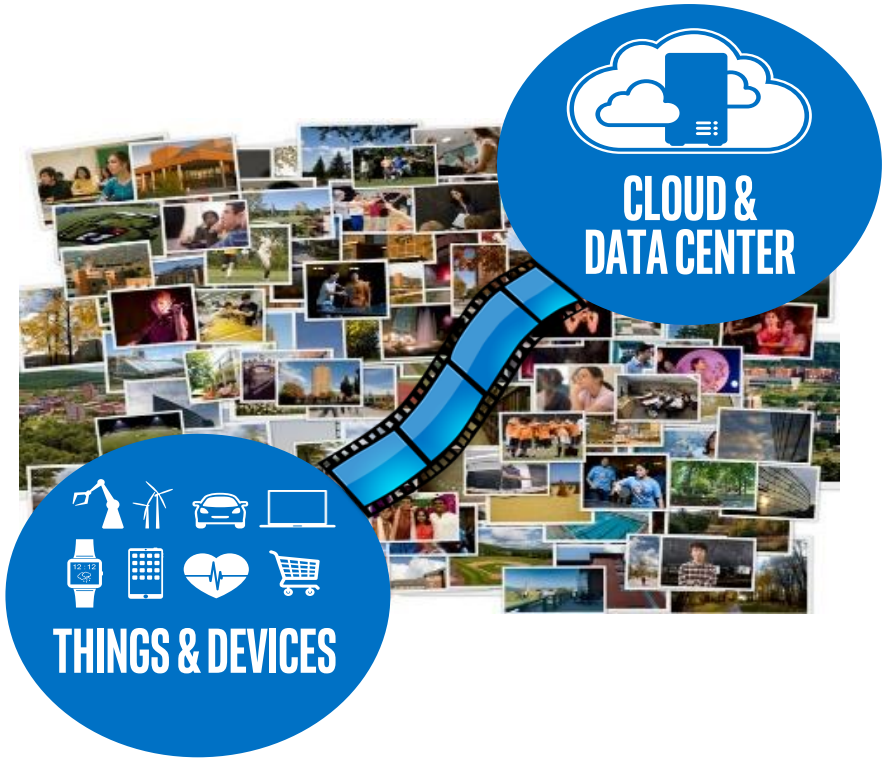
May 24, 2018

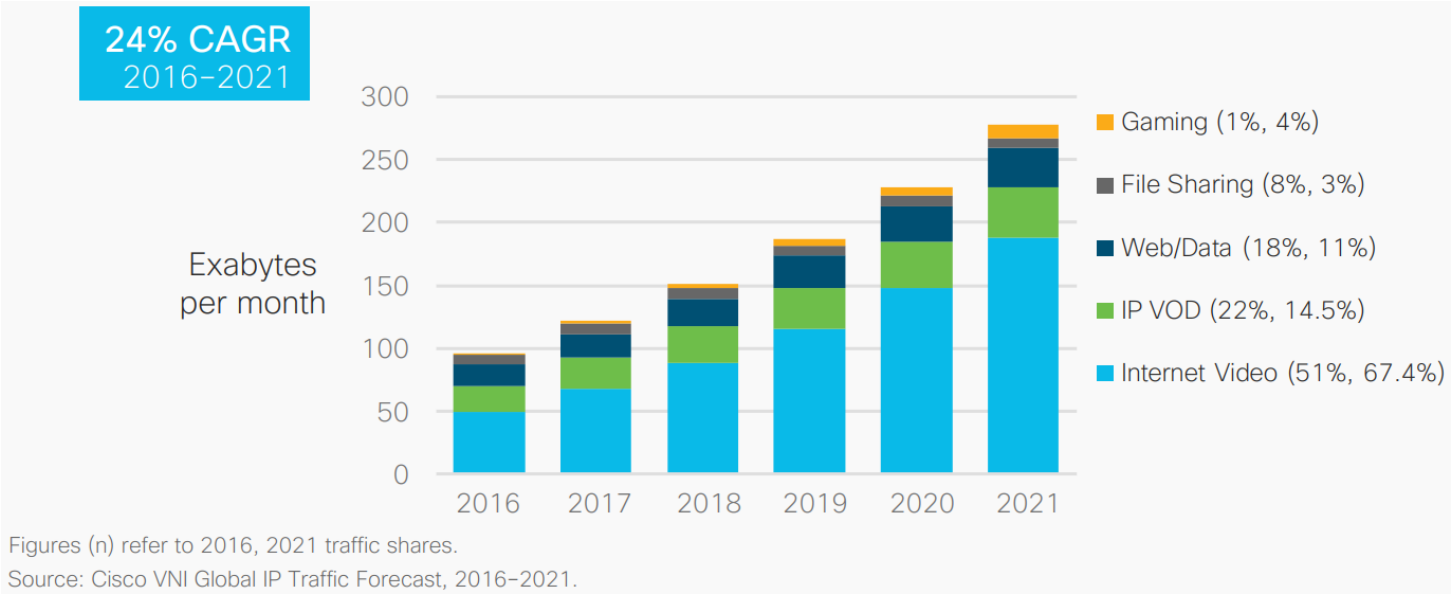
intel AI

AGENDA

- Introduction
- Face Analysis & Emotion Recognition
- Deep Learning based Visual Recognition
- Visual Parsing & Multimodal Analysis
- Summary

VISUAL DATA EXPLOSION





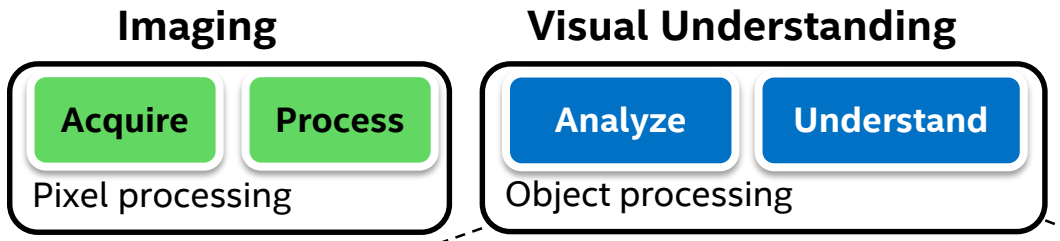
Source: Cisco white paper "The Zettabyte Era: Trends and Analysis", June 2017.

The vast majority of data in cloud and on edges/devices is visual !!

Key challenge: *How to process and understand these visual data?*

VISUAL UNDERSTANDING – WHAT IS IT?

Computer Vision (CV) is a field that includes methods for acquiring, processing, analyzing and understanding images and, in general, high-dimensional data from the real world in order to produce numerical or symbolic information Wikipedia.org



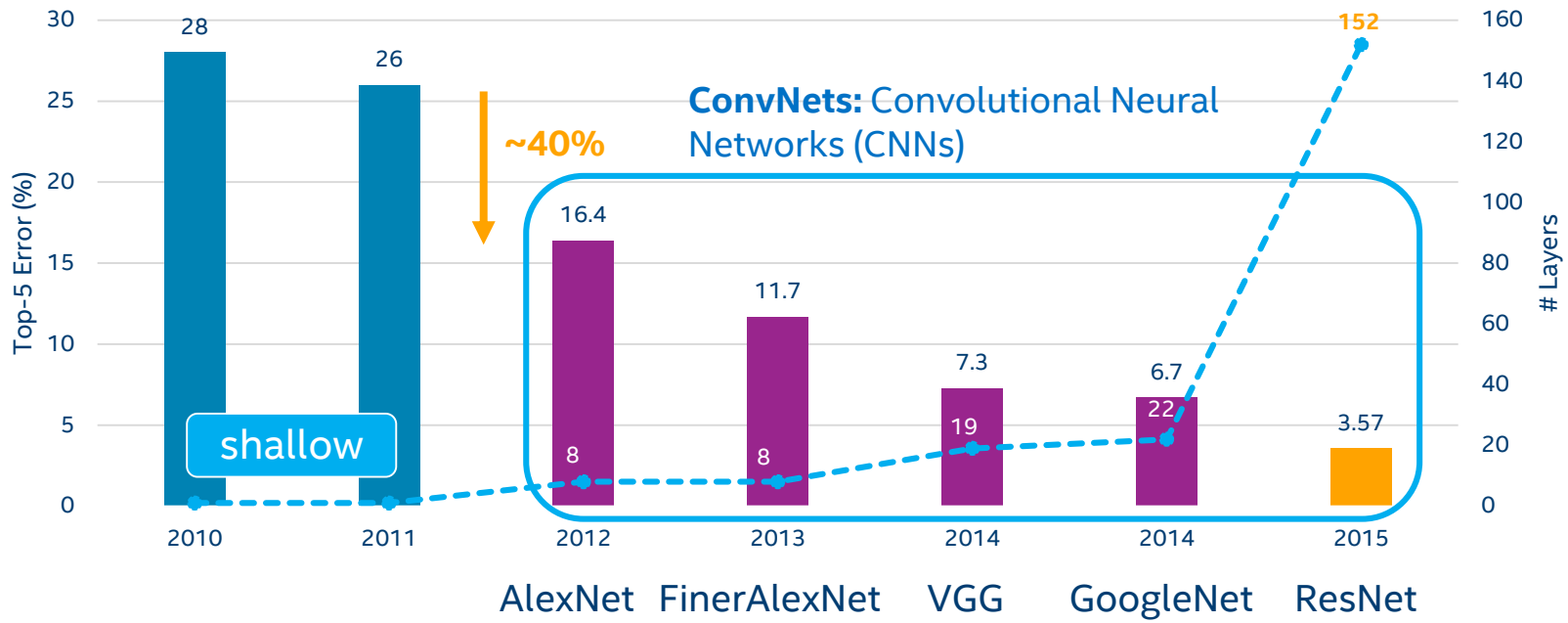
Sample Capabilities	
Face Detection & Recognition	Emotion Recognition
Text Recognition in the wild	Object Classification/Detection
Action Recognition	Scene Classification/Understanding
Video Classification/Summarization	•••

Classification: Person, Camera

Action: Taking pictures

Objective – Derive knowledge out of images/videos of the real world

DEEP LEARNING BREAKTHROUGHS FOR VISUAL RECOGNITION



ImageNet Large Scale Visual Recognition Challenge (ILSVRC): 1000-catg Object Classification

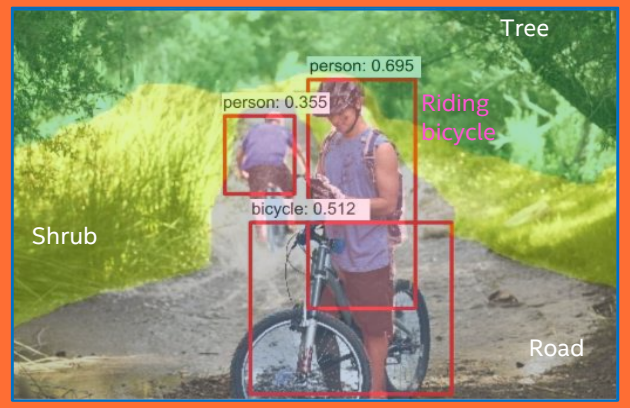
"ConvNets are now the dominant approach for almost all recognition and detection tasks and approach human performance on some tasks."

LeCun, Bengio, Hinton, Deep Learning, Nature, May 2015

VISUAL UNDERSTANDING & SYNTHESIS RESEARCH

Research innovation in ***smart visual data processing*** technologies on Intel Platforms

Visual Understanding



- Object Recognition/Detection
- Action/Activity Recognition
- Semantic Segmentation
- Geometric Layout Estimation
- High-level Scene Understanding
- Visual-centric Multimodal Understanding (Emotion, Visual Content…)

Foundational Components
- CNN architectures, Visual Odometry, SLAM, Visual Indexing…

Image/Video Synthesis

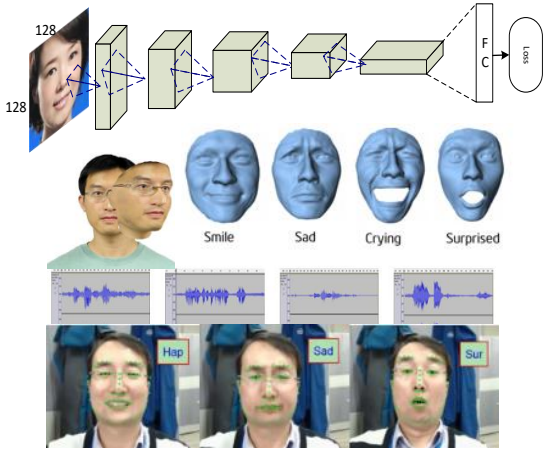
- 3D Modeling & Reconstruction
- Geometry processing
- Animation & Rendering

SW/HW Co-Design (w/ BUs…) | **SW API/Tools** | **Application/System Prototyping** | **Visual-based Decision-Making/Control**

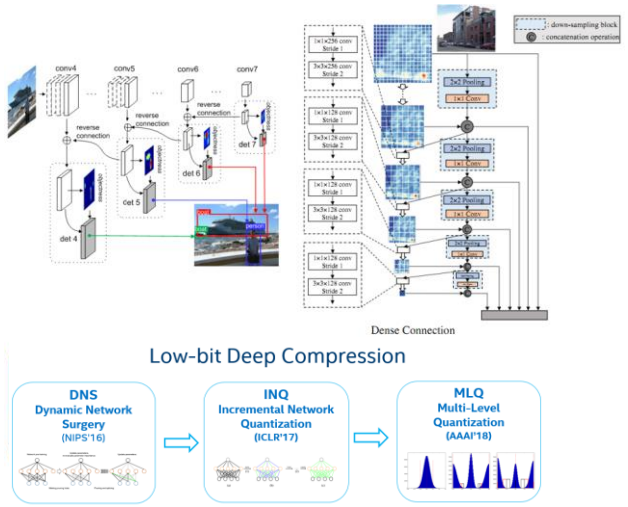
VISUAL UNDERSTANDING RESEARCH

Innovate in cutting-edge *visual cognition* & *machine learning* technologies for *smart computing* to enable *novel usages* and *user experience*

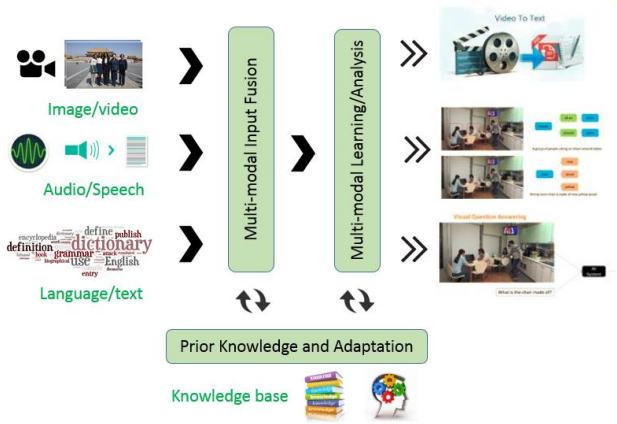
Face Analysis & Emotion Recognition



Deep Learning based Visual Recognition



Visual Parsing & Multimodal Analysis

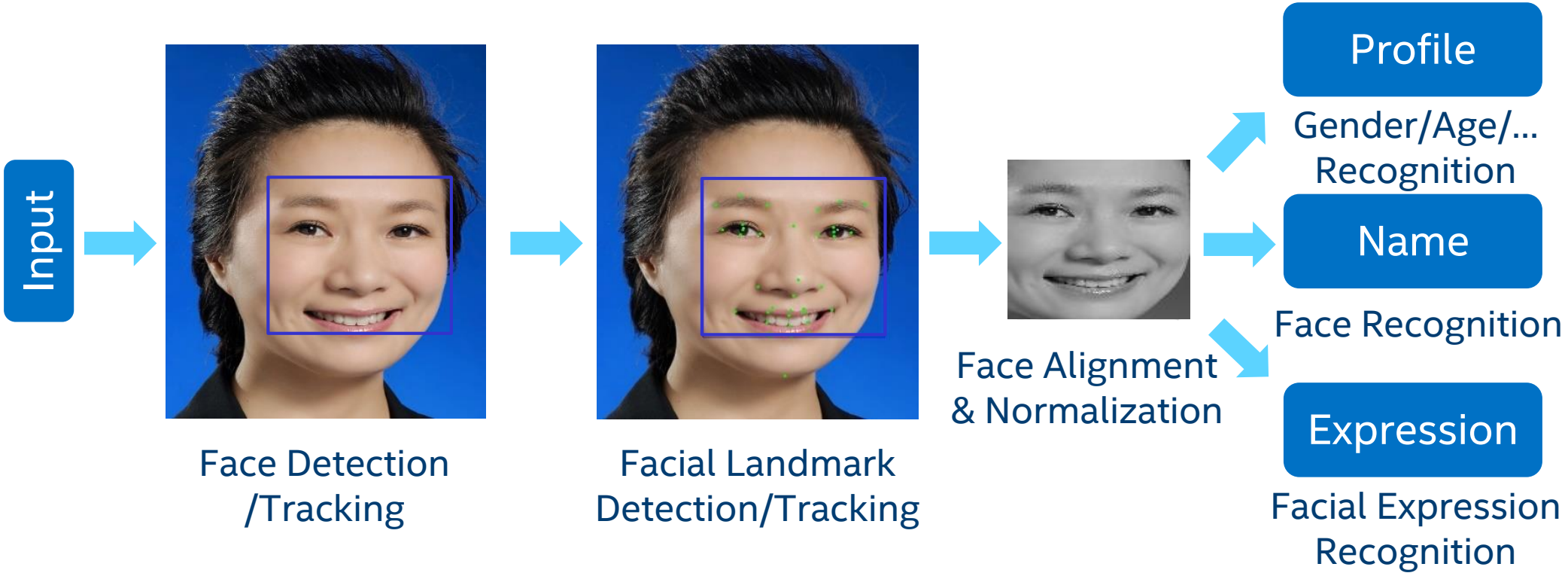


AGENDA

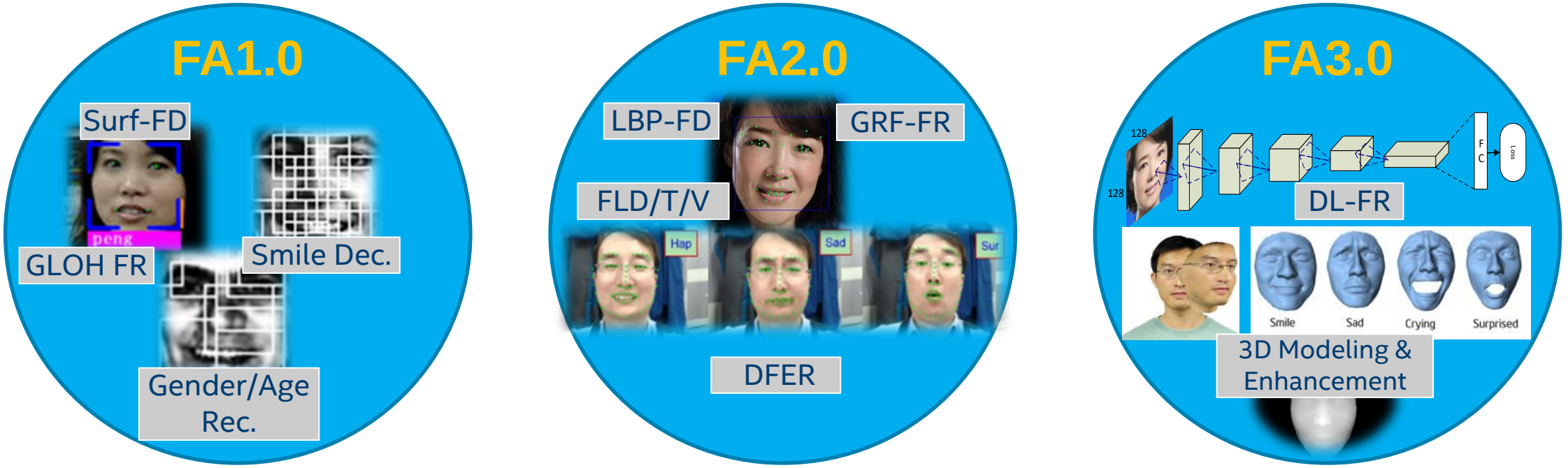
- Introduction
- Face Analysis & Emotion Recognition
- Deep Learning based Visual Recognition
- Visual Parsing & Multimodal Analysis
- Summary

FACE ANALYSIS TECHNOLOGY RESEARCH

Intel China Labs (ILC) developed a full Face Analysis pipeline with best in class algorithms (20+ IPs)



ILC FACE ANALYSIS TECHNOLOGY EVOLUTION



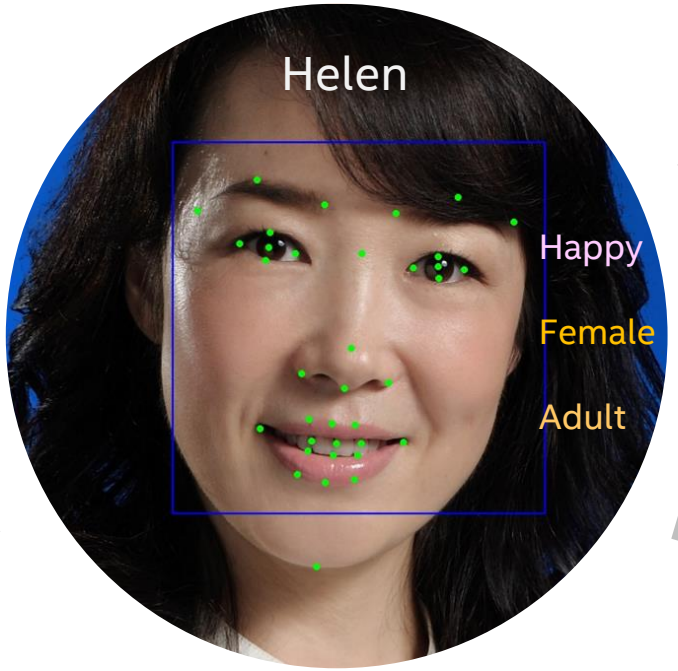
MAKE INTEL PLATFORMS GREAT WITH FACE ANALYSIS

GenX Acceleration

SDK Integration

intel REALSENSE TECHNOLOGY

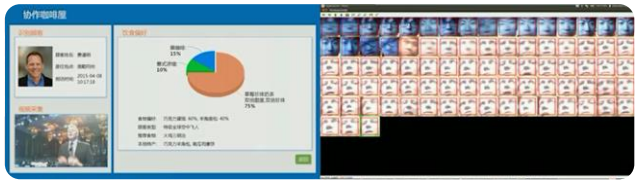
Intel® Computer Vision SDK



ILC Leading Face Analysis Technology

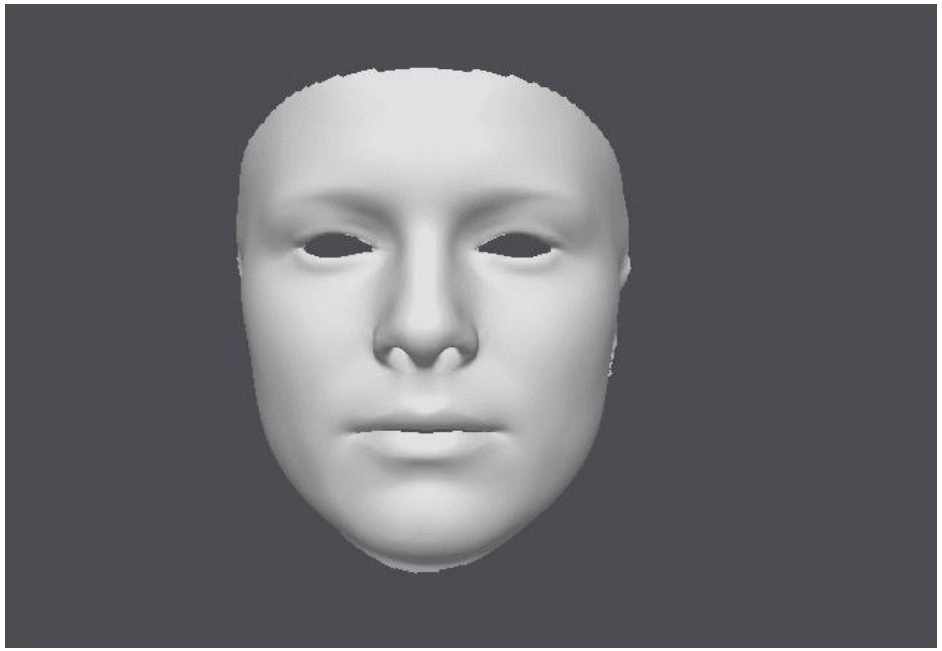
App Prototyping

Video E2E Solution



3D FACE TECHNOLOGY

Real-time 3D face *modeling*, *tracking* and *enhancement* for real-life applications in VR/AR/Gaming…



REAL-TIME 3D FACIAL EFFECTS EXPERIENCE DEMO

3D FACE TECHNOLOGY USE CASE

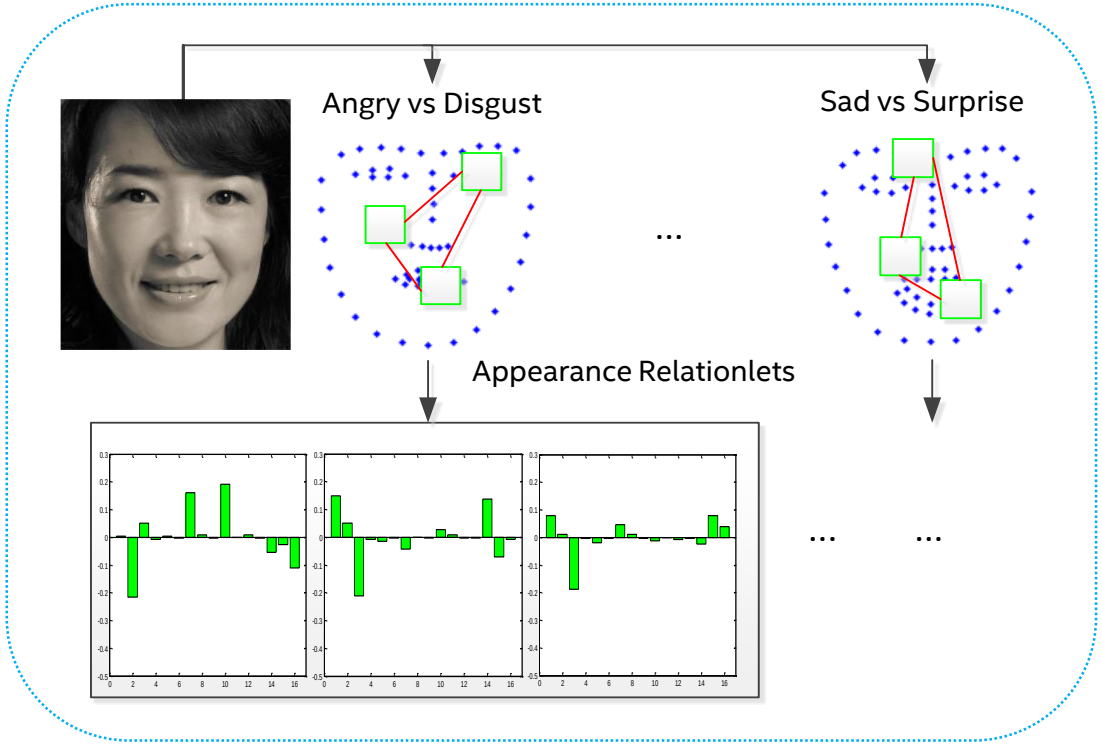
World 1st Intel AI MV
Powered by
ILC 3D Face Technology

Chris Lee
"*Rainy Day, But We Are Together*"

VISUAL EMOTION RECOGNITION

Smart world must be one with *Emotions*...

AU-Aware Features and Interactions (*AUAFI*, *ACM ICMI'15*) with multi-task learning to decode facial muscle movements & their inherent interactions



Methods	Overall recognition rate (%)
FDM (ECCV'14)	97.70
DTAGN (ICCV'15)	96.94
AUAFI	**98.70**

CK+ dataset: 327 videos (neutral-onset-peak), 7 basic facial expressions, 118 subjects with frontal pose only

Methods	Overall recognition rate (%)
STM- ExpLet (CVPR'14)	75.12
DTAGN (ICCV'15)	66.33
AUAFI	**80.27**

MMI dataset: 205 videos (neutral-onset-peak-offset-neutral), 6 basic facial expressions, 23 subjects with large pose variations

VISUAL-AUDIO SOLUTION FOR EMOTION RECOGNITION

Proposed “Importance-Aware Features” (IAF) with selective grouping for audio solution and designed a framework to fuse visual & audio solutions

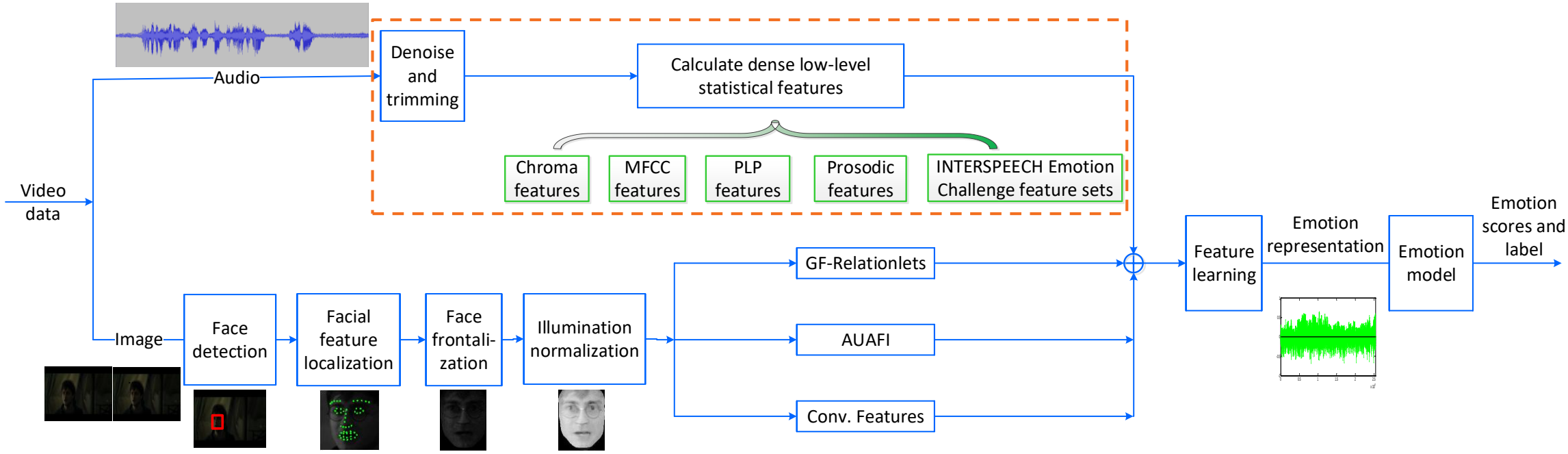


Image source: EmotiW 2015

EMOTION RECOGNITION IN THE WILD CHALLENGE

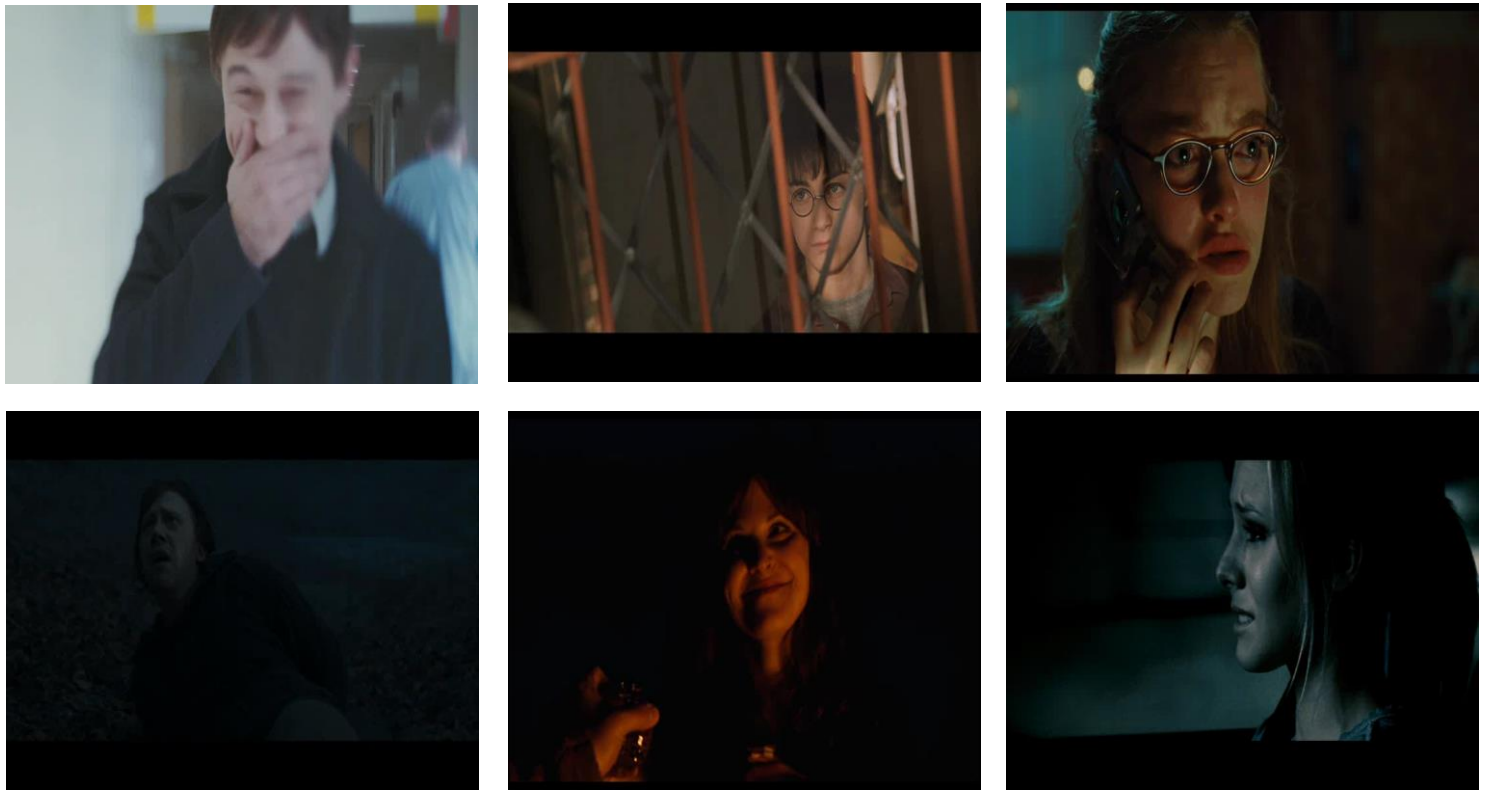
Won EmotiW 2015 (ACM ICMI 2015) in the audio-video based task, and competitors included 74 teams (CMU, UIUC, MSR, etc.) across the world

- **Task 1: EmotiW 2015 AFEW dataset**
  7 basic facial expressions completely shown in movie clips
- **Task 2: EmotiW 2015 SFEW dataset**
  7 basic facial expressions completely shown in static images

Overall Recognition Rate (%) on EmotiW2015 Test Set

Methods	AFEW	SFEW
Baseline	39.33	39.13
Winner 2014	50.37	N/A
ILC solution	**53.80**	**55.38**

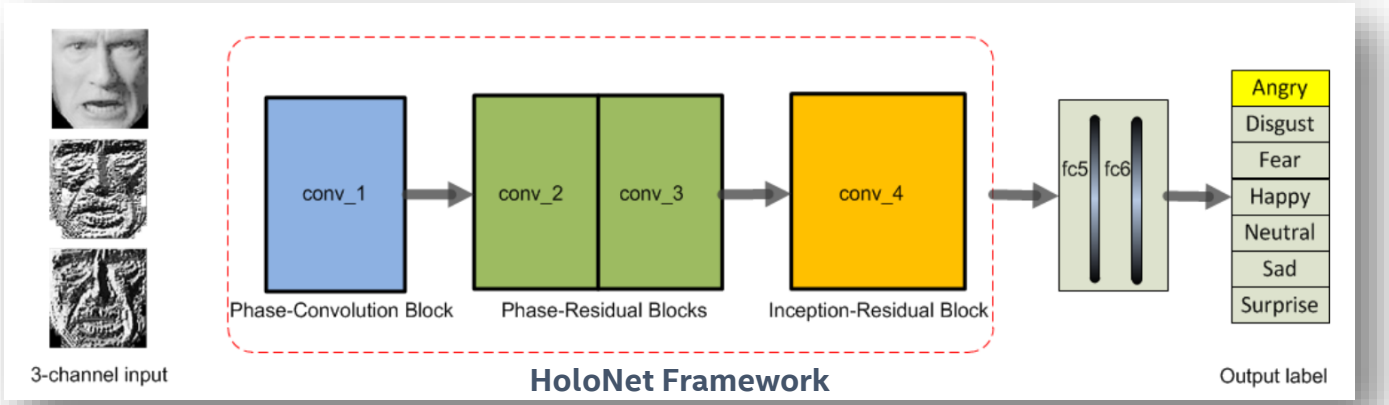
Examples of EmotiW2015 Video Clips



Video source: EmotiW 2015

HOLONET FOR ROBUST EMOTION RECOGNITION

Invented a deep yet computationally efficient CNN framework, HoloNet for robust emotion recognition (*EmotiW 2016 Most Influential paper*)



"... You showed me a really novel method, no use of extra data and its speed is hundreds of times faster than the other competitors.

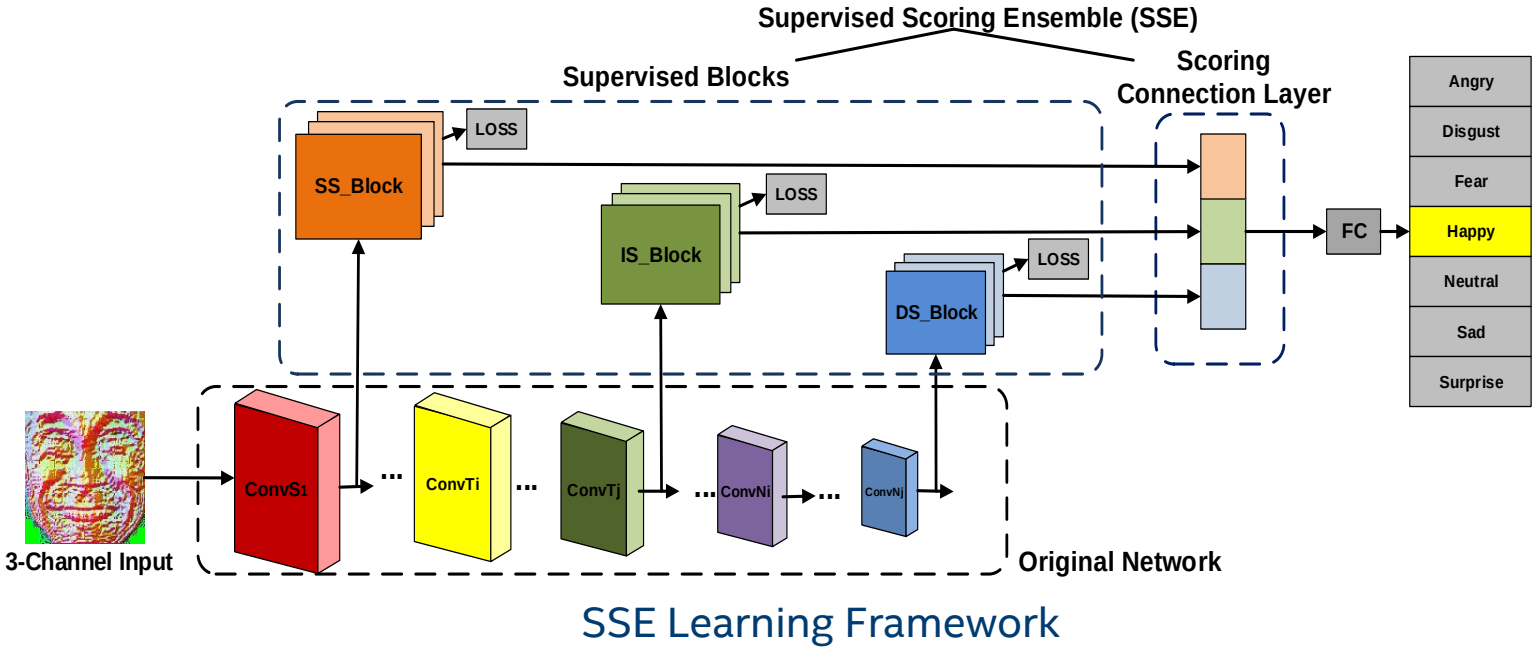
Abhinav, EmotiW 2016 Chair

Submission#	Validation (%)	Test (%)	Method
1	47.78	54.30	Fusion of HoloNet model A + 1 audio model
2	48.83	55.14	Fusion of HoloNet model B + 1 audio model
3	50.13	56.83	1st Fusion of HoloNet model A&B + 1 audio model
4	50.91	55.14	2nd Fusion of HoloNet model A&B + 1 audio model
5	**51.96**	**57.84**	Fusion of HoloNet model A&B + 1 audio model + 1 iDT model

Total recognition accuracy of our 5 submissions to AFEW 6.0, both on the validation and the test sets.

SSE: SUPERVISED SCORING ENSEMBLE

Invented SSE to enable discriminative learning within one single CNN for accurate emotion recognition (*EmotiW 2017 winner*)



SSE Learning Framework

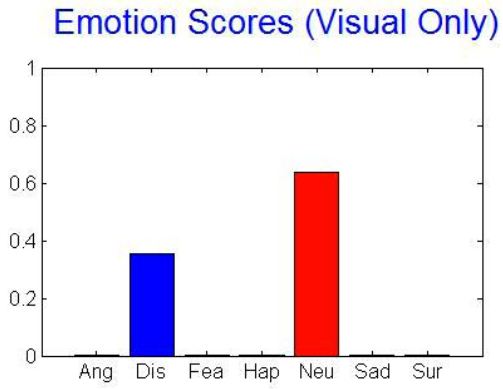
5.5% better than HoloNet (single model)

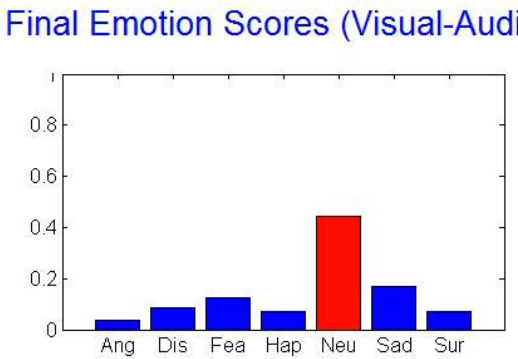
60.3% accuracy in EmotiW 2017

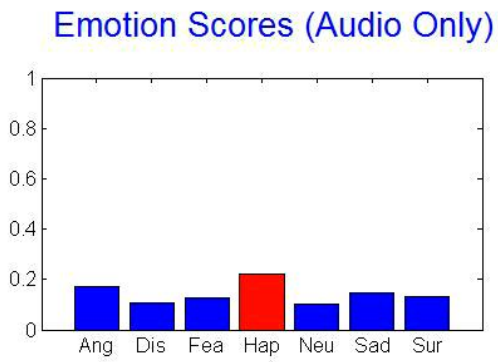
P. Hu, D. Cai, S. Wang, A. Yao, Y. Chen, "Learning Supervised Scoring Ensemble for Emotion Recognition in the Wild", ACM ICMI 2017.

REAL-TIME MULTIMODAL EMOTION RECOGNITION PROTOTYPE

Video source: EmotiW2015



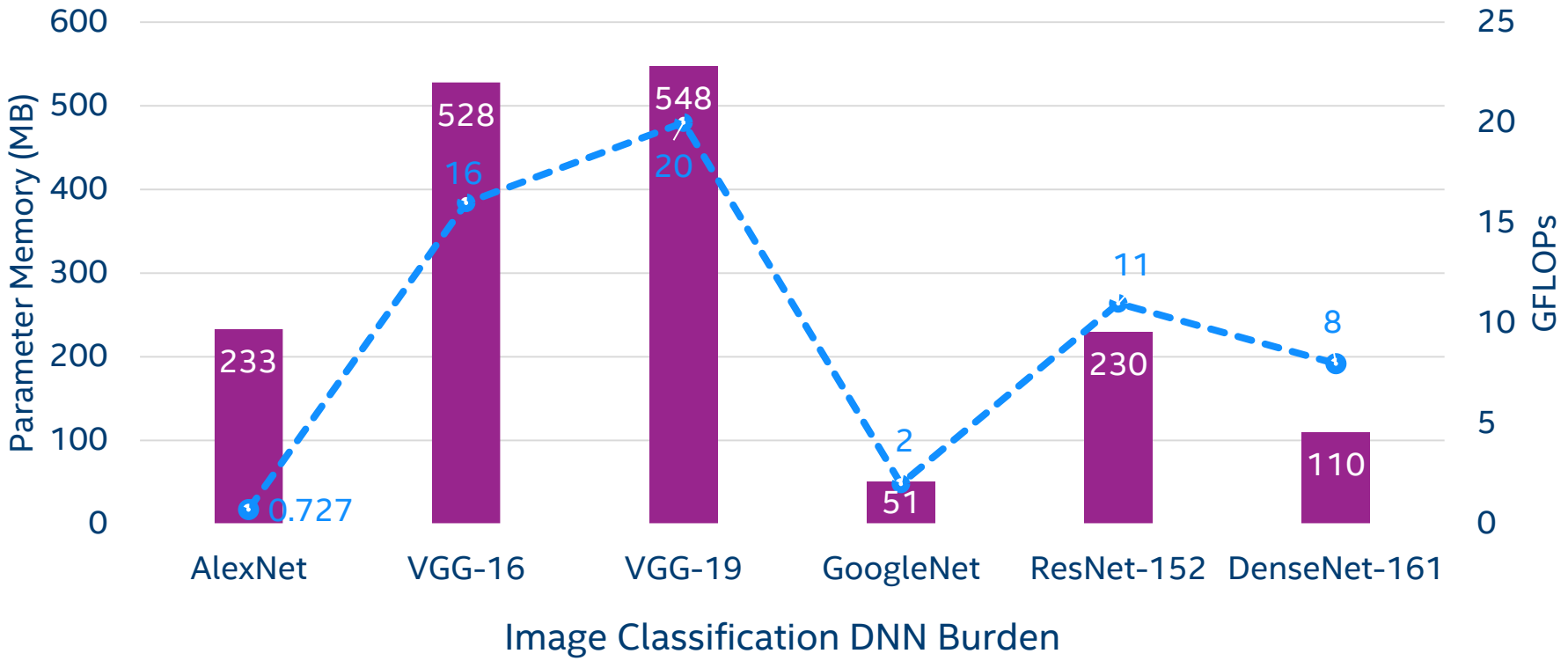




AGENDA

- Introduction
- Face Analysis & Emotion Recognition
- Deep Learning based Visual Recognition
- Visual Parsing & Multimodal Analysis
- Summary

DEEP LEARNING CHALLENGES



Deployment: Most mainstream DNNs are both *compute* and *memory intensive*, difficult to deploy to embedded/edge devices

OBJECT DETECTION ALGORITHM EVOLUTION

Region-based Methods

Region proposals to reduce se
space & region-wise subnetwor
detection

RCNN (CVPR'14), SPP (ECCV'14)
Fast/Faster RCNN (ICCV/NIPS'15)
HyperNet (CVPR'16)
R-FCN (NIPS'16

Associate the best of both methodologies
RON
(CVPR'17)

ee Methods

Single shot with FCNs
YOLO (CVPR'16)
SSD (ECCV'16)

Sliding Window + Handcrafted Features

Using different Aspect ratio/ Scale/ Stride/ Classifier

ACCURACY

SPEED

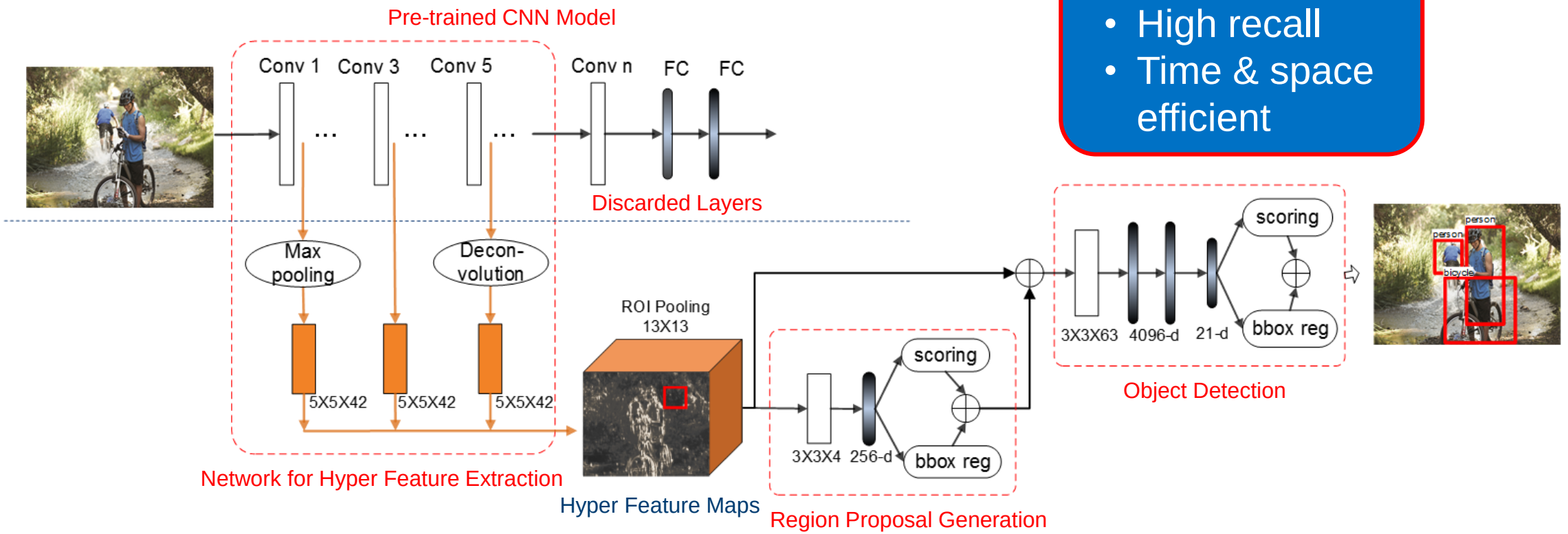
Deep Learning (CNN) based Methods

HYPERNET – AN EFFICIENT OBJECT DETECTION SOLUTION

A unified framework for region proposal selection and object detection (CVPR'16)

- Shares Hyper Feature across different tasks

Advantages:
- High recall
- Time & space efficient



Conv: Convolutional layer FC: Fully-Connected layer ROI: Region of Interest bbox reg: bounding box regression

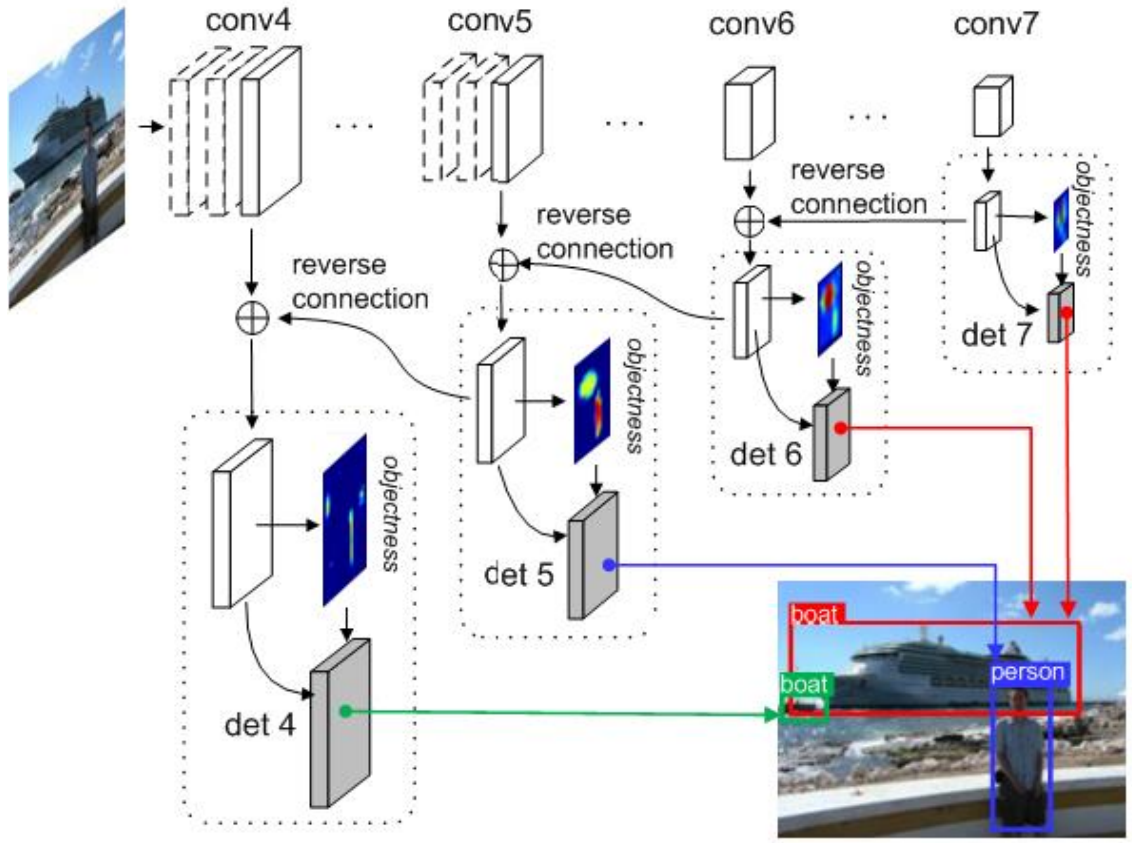
T. Kong, A. Yao, Y. Chen, F. Sun, "HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection", CVPR 2016.

RON: REVERSE CONNECTION WITH OBJECTNESS PRIOR NETWORKS

A fully convolutional framework to solve two fundamental problems (CVPR'17)

- Multi-scale object localization with Reverse Connection Pyramids
- Efficient negative space mining with Objectness Prior Networks

Achieved SOTA accuracy & speed



T. Kong, A. Yao, F. Sun, M. Lu, H. Liu, Y. Chen, "RON: Reverse Connection with Objectness Prior Networks for Object Detection", CVPR 2017.

DSOD: LEARNING DEEPLY SUPERVISED OBJECT DETECTORS FROM SCRATCH

First training from scratch OD solution (ICCV'17)

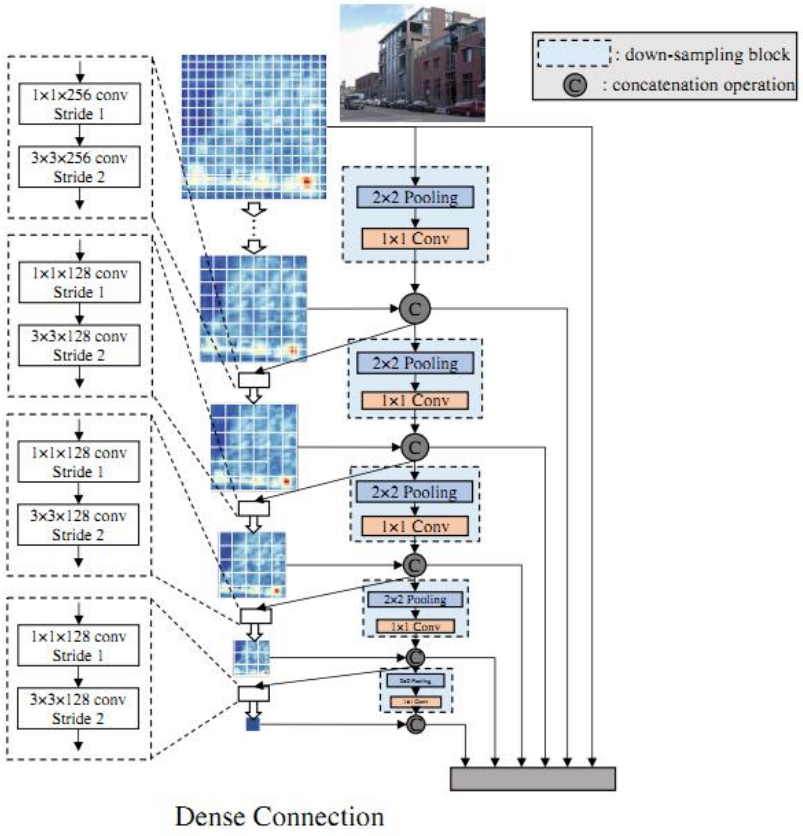
- Not require ImageNet pre-training
- Even with limited (>=10K) bbox annotations

State-of-the-art Accuracy & Efficiency

- #parameter: ½ SSD, ¼ R-FCN, 1/10 Faster-RCNN
- Better accuracy than SSD/YOLOv2 on VOC/MS-COCO
- ~20fps w/o tailored optimization on Intel NUC

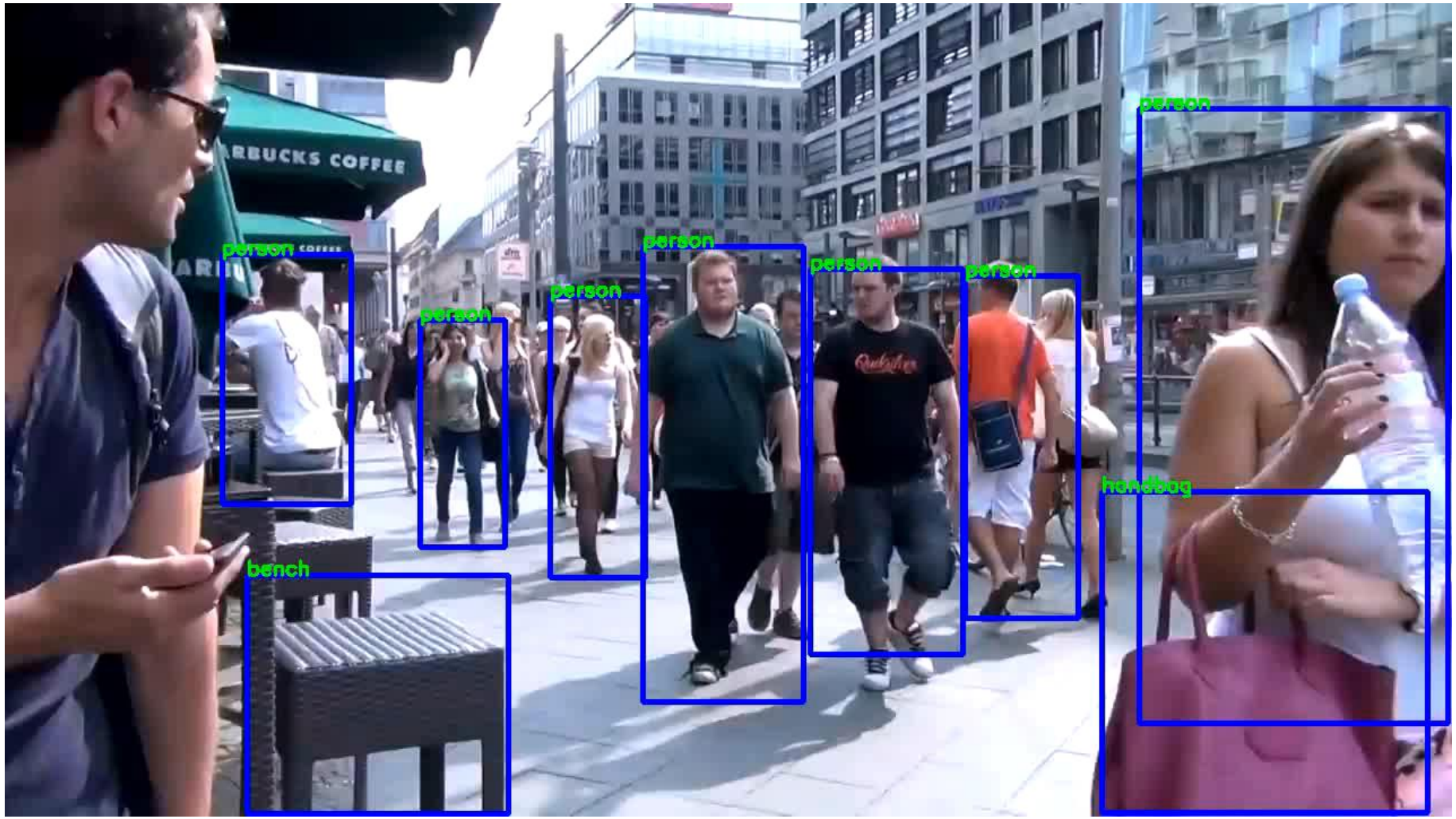
Open many possibilities

- Training with limited data for specific problems
- Other domains: depth/medical/multi-spectral images



Z. Shen, Z. Liu, J. Li, Y. Jiang, Y. Chen, X. Xue, "DSOD: Learning Deeply Supervised Object Detectors from Scratch", ICCV 2017.

MULTI-CLASS OBJECT DETECTION PROTOTYPE

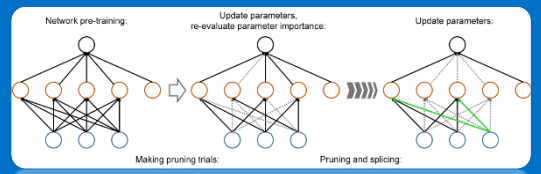


Video source: youtube

MODEL COMPRESSION: LOW-BIT DEEP COMPRESSION

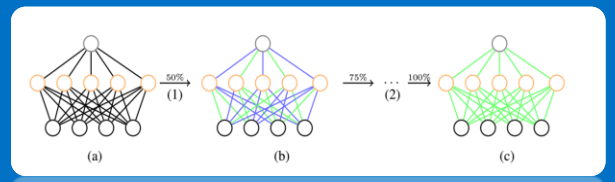
A leading and elegant solution to achieve *hundred-level lossless* compression on DNNs with *low-precision weights* and *activations*

DNS
Dynamic Network Surgery
(NIPS'16)



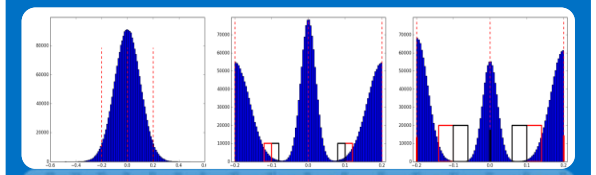
Seek optimal DNN architecture

INQ
Incremental Network Quantization
(ICLR'17)



Constrain optimal DNN with low-bit weights

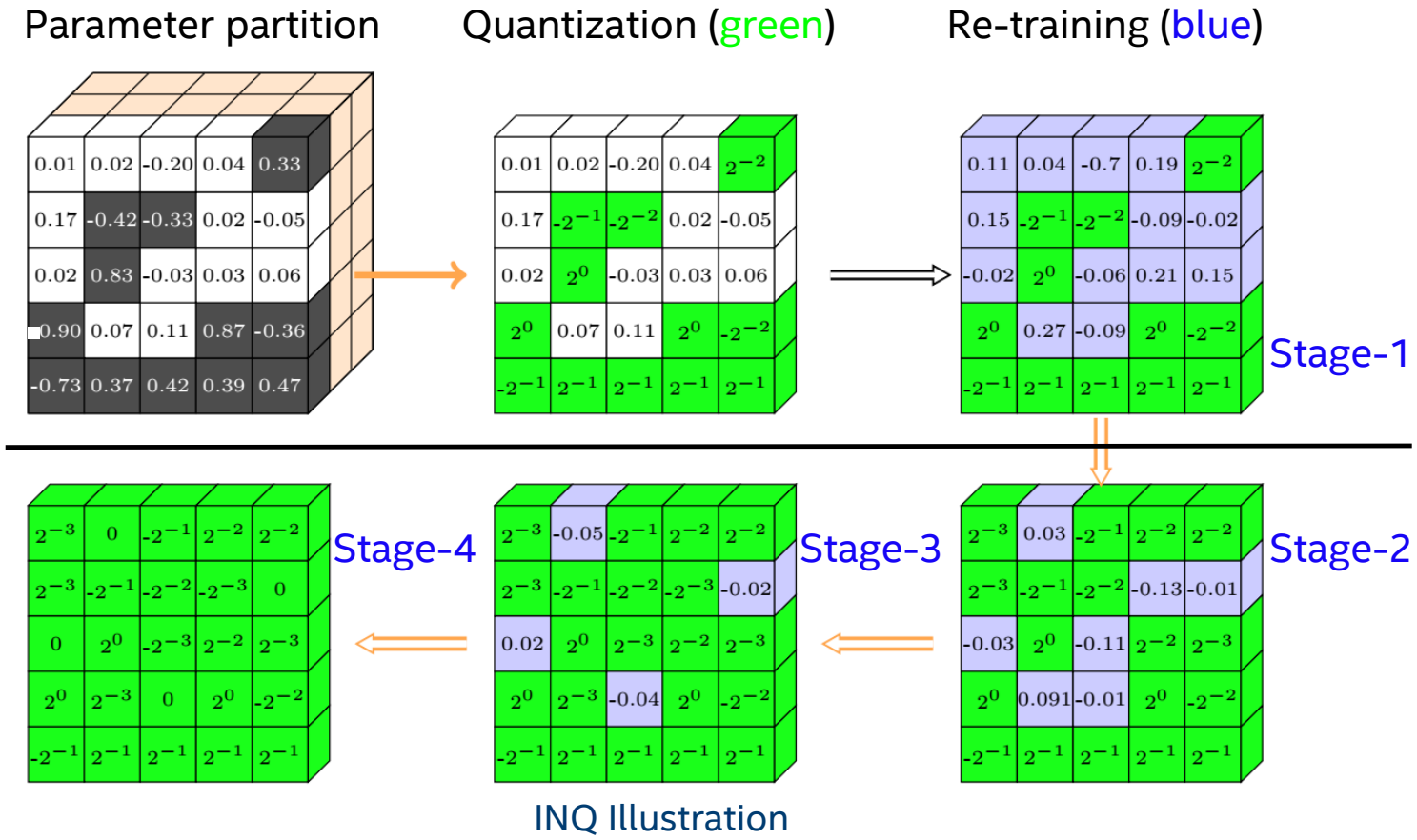
MLQ
Multi-Level Quantization
(AAAI'18)



Constrain optimal DNN with low-bit activations

INQ: INCREMENTAL NETWORK QUANTIZATION

Three novel operations: *parameter partition* (well-defined metric), *quantization* and *re-training*, making whole procedure progress in *an incremental manner*



INQ Illustration

Advantages:

- First lossless network quantization solution
- Model-free
- FP Multiplication → Binary bit shift
- Efficient training

A. Zhou, A. Yao, Y. Guo, L Xu and Y. Chen, "Incremental Network Quantization: Towards Lossless CNNs with Low-precession Weights", ICLR 2017.

INQ RESULTS

Achieved *improved accuracy with 5-bit quantization* (actually 4 bit + 1 zero) for popular DNNs, and obtained strongly comparable low-precision models against the full-precision reference ResNet-18 model using 5/4/3-bit even 2-bit

5-bit results

Network	Bit-width	Top-1 error	Top-5 error
AlexNet ref	32	42.76%	19.77%
AlexNet	5	**42.61%**	**19.54%**
VGG-16 ref	32	31.46%	11.35%
VGG-16	5	**29.18%**	**9.70%**
GoogleNet ref	32	31.11%	10.97%
GoogleNet	5	**30.98%**	**10.72%**
ResNet-18 ref	32	31.73%	11.31%
ResNet-18	5	**31.02%**	**10.90%**
ResNet-50 ref	32	26.78%	8.76%
ResNet-50	5	**25.19%**	**7.55%**

Different bit results

Model	Bit-width	Top-1 error	Top-5 error
ResNet-18 ref	32	31.73%	11.31%
INQ	5	31.02%	10.90%
INQ	4	31.11%	10.99%
INQ	3	31.92%	11.64%
INQ	2 (ternary)	33.98%	12.87%

Method	Bit-width	Top-1 error	Top-5 error
BWN	1	39.20%	17.00%
TWN	2 (ternary)	38.20%	15.80%
INQ (ours)	2 (ternary)	**33.98%**	**12.87%**

A. Zhou, A. Yao, Y. Guo, L Xu and Y. Chen, “Incremental Network Quantization: Towards Lossless CNNs with Low-precession Weights”, ICLR 2017.

LOW-BIT DEEP COMPRESSION RESULTS

Outperforms the state-of-the-art Deep Compression solution* with *at least 1X absolute margin* on AlexNet, achieving *>100X compression* with 2 bits

Method	Bit-width (Conv/FC)	Bit-width (Act)	Compression ratio	Decrease in top-1 / top-5 error rate
P+Q *	8/5	32	27x	0.00% / 0.03%
P+Q+H *	8/5	32	35x	0.00% / 0.03%
Our method	**4/4**	**4**	**71x**	0.08% / 0.03%
P+Q+H *	4/2	32	-	-1.99% / -2.60%
Our method	**3/3**	**4**	**89x**	-0.52% / -0.20%
Our method	**2/2**	**4**	**142x**	-1.47% / -0.96%

Comparison of our low-bit deep compression and deep compression method (P+Q+H, LCLR'16 and ISCA'16) on AlexNet. Conv: Convolutional layer, FC: Fully connected layer, Act: Activation, P: Pruning, Q: Quantization, H: Huffman coding.

* S. Han, J. Pool, J. Tran, and W. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. Best paper in ICLR 2016.

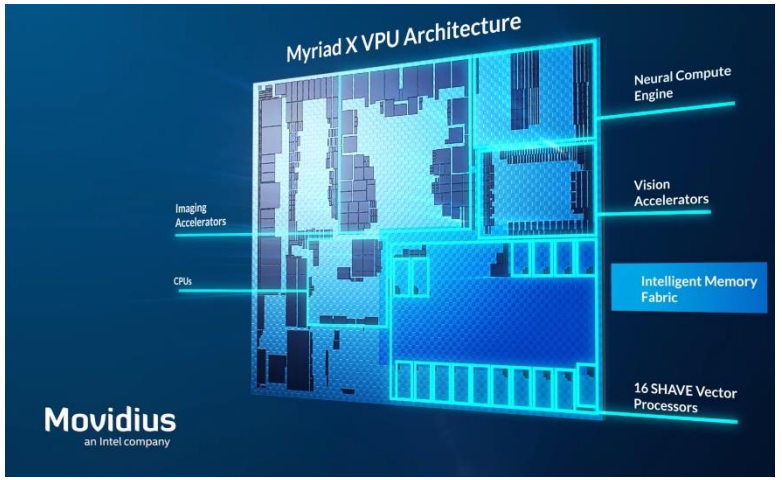
HW ACCELERATION FOR DEEP LEARNING INFERENCE

Low-bit Deep Compression lays a solid foundation for *HW acceleration of deep learning inference* in *fog/edge computing*

INTEL FPGAS

intel ARRIA 10 inside

intel + Movidius

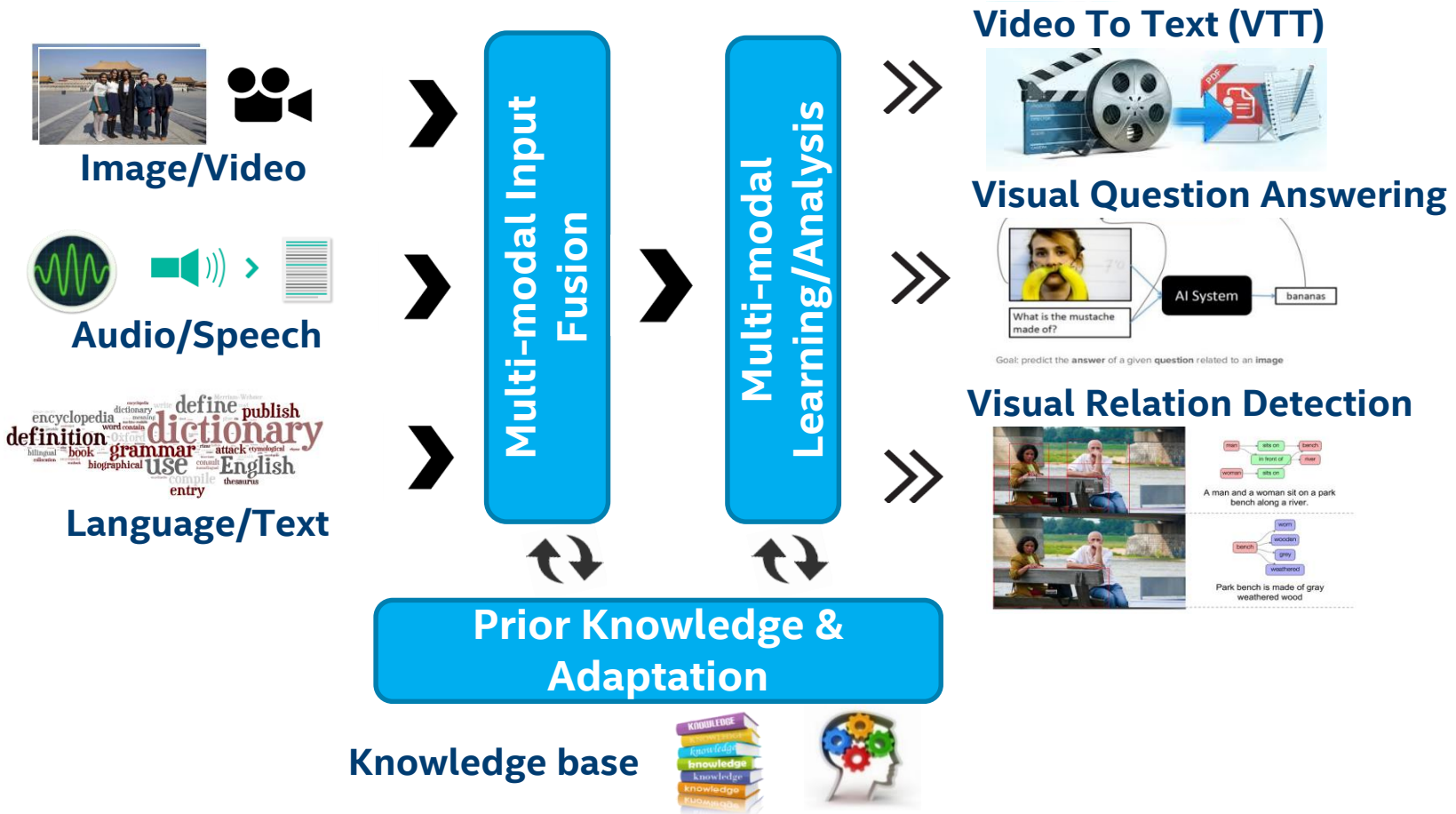


AGENDA

- Introduction
- Face Analysis & Emotion Recognition
- Deep Learning based Visual Recognition
- Visual Parsing & Multimodal Analysis
- Summary

VISUAL PARSING & MULTIMODAL ANALYSIS

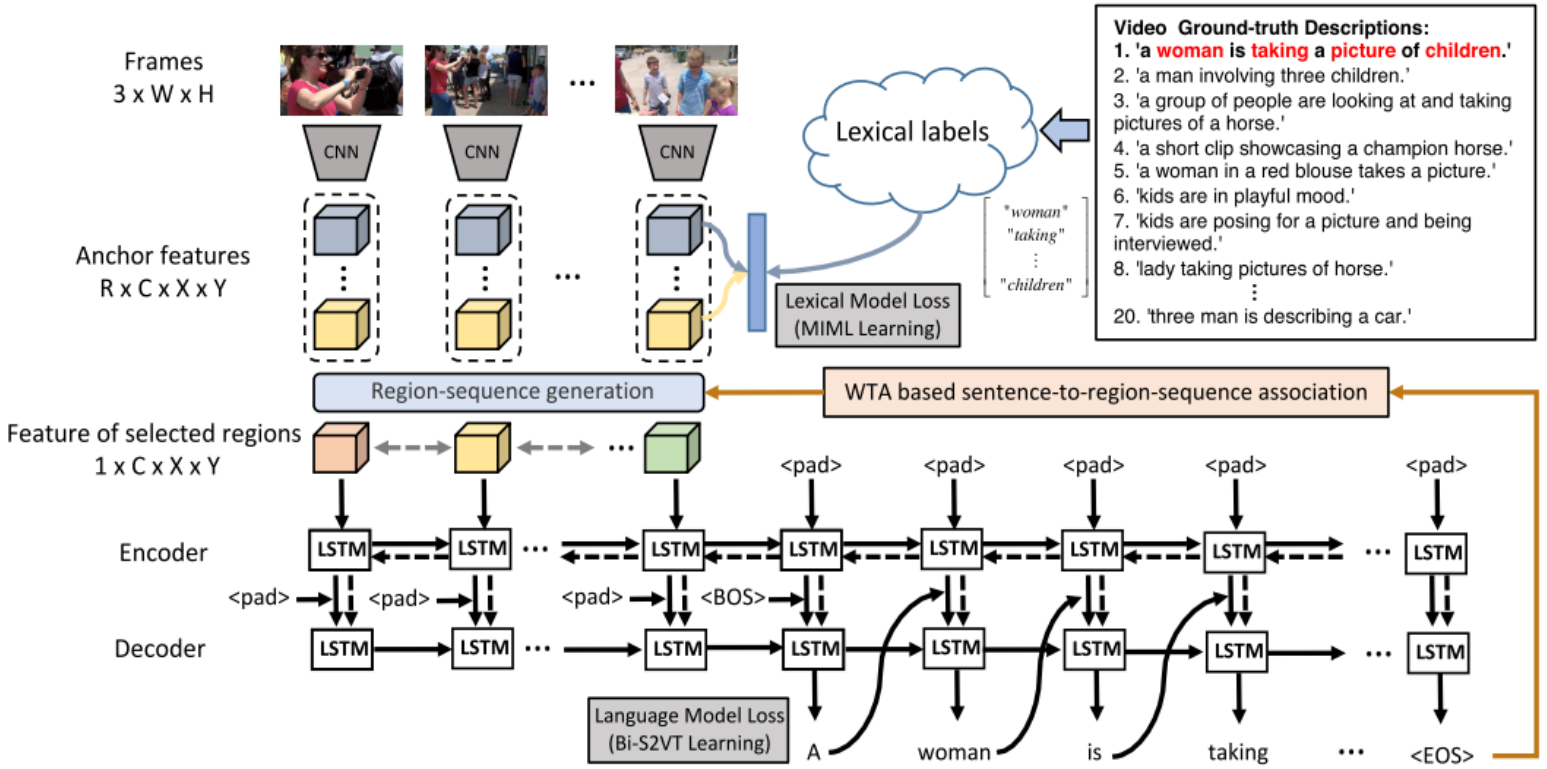
Advanced multimodal fusion & learning research to *bridge the gap* between visual recognition and visual understanding



DENSE VIDEO CAPTIONING

Invented a novel solution to produce *informative & diverse dense captions* and *outperform SOTA* single video captioning methods

Video source: MSR-VTT dataset *



Z. Shen, J. i, Z. Su, M. Li, Y. Chen, Y. Jiang, X. Xue, "Weakly Supervised Dense Video Captioning ", CVPR 2017.

* J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. CVPR 2016.

DENSE VIDEO CAPTIONING

Invented a novel solution to produce *informative & diverse dense captions* and *outperform SOTA* single video captioning methods

Video source: MSR-VTT dataset *

Region Sequences & DenseVidCap

A **woman** in **red blouse** is **taking pictures** of **children**

A group of people are **taking pictures** of a **horse**

Kids are being **interviewed**

Z. Shen, J. i, Z. Su, M. Li, Y. Chen, Y. Jiang, X. Xue, "Weakly Supervised Dense Video Captioning ", CVPR 2017.

* J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. CVPR 2016.

DENSE VIDEO CAPTIONING

Invented a novel solution to produce *informative & diverse dense captions* and *outperform SOTA* single video captioning methods

Region Sequences & DenseVidCap

Z. Shen, J. i, Z. Su, M. Li, Y. Chen, Y. Jiang, X. Xue, "Weakly Supervised Dense Video Captioning ", CVPR 2017.

Video source: J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. CVPR 2016.

DENSE VIDEO CAPTIONING

Invented a novel solution to produce *informative & diverse dense captions* and *outperform SOTA* single video captioning methods

Results on MSR-VTT Challenge

Team	Memo	METEOR	BLEU-4	ROUGE-L	CIDEr
Ruc-UVA	RUC + UVA + ZJU	26.9	38.7	58.7	45.9
VideoLab	UCB + Austin +...	27.7	39.1	60.6	44.1
Aalto	Aalto Univ.	26.9	39.8	59.8	45.7
V2t-navigator	RUC + CMU	28.2	40.8	60.9	44.8
Ours	ILC	**28.3**	**41.4**	**61.1**	**48.9**

Z. Shen, J. i, Z. Su, M. Li, Y. Chen, Y. Jiang, X. Xue, “Weakly Supervised Dense Video Captioning ”, CVPR 2017.

VIDEO CAPTIONING DEMO

VISUAL QUESTION ANSWERING

Question objective categories

- Apparent objective (through recognition results)
- Indiscernible objective (requires knowledge confirmation due to unclear target)
- Invisible objective (requires reasoning from external knowledge)

Q: What is the color of the batter's shirt?
A: Red

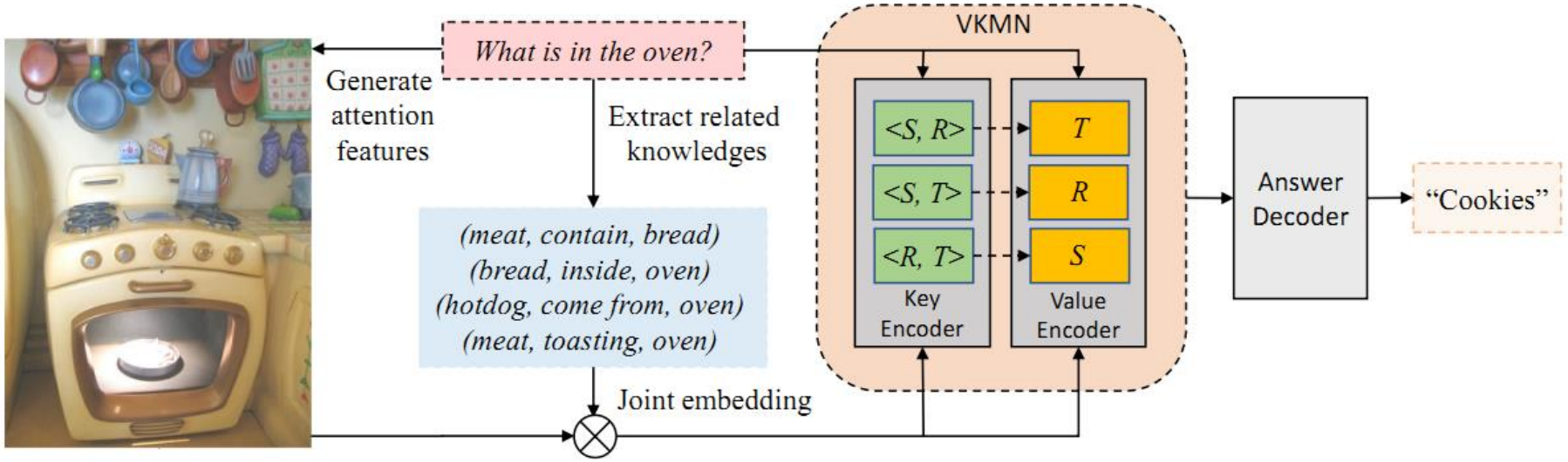
Q: What is in the oven?
A: Cookies

Q: What kind of animal would love to eat these fruits?
A: Monkey

VKMN: VISUAL-KNOWLEDGE MEMORY NETWORKS

An end-to-end learning framework seamlessly incorporates *structured human knowledge* and *deep visual features* into *memory networks*

- Input module: image/question
- Knowledge spotting module: retrieval case related knowledge
- Joint embedding module: joint visual and knowledge embedding
- Memory module: receiving query, reading memory and predicting answers



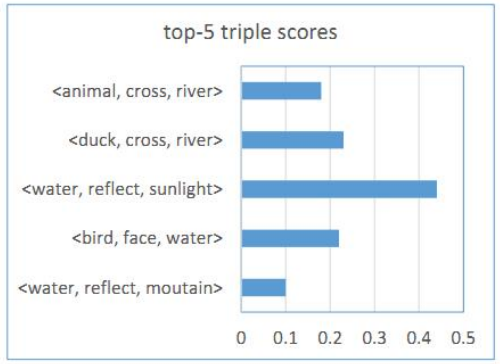
Z. Su, C. Zhu, Y. Dong, D. Cai, Y. Chen, J. Li, "Learning Visual Knowledge Memory Networks for Visual Question Answering", to appear in CVPR 2018.

VKMN RESULT ILLUSTRATIONS

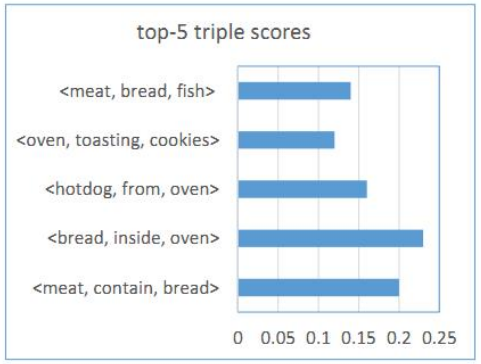
Achieved SOTA results on VQA 1.0/2.0 benchmarks and better results on questions required knowledge reasoning/confirmation

Q: Who is wearing a red hat?
MLB: Dog
Ours: Man

Q: Which animal is reflected in the water?
MLB: Dog
Ours: Duck



Q: What is in the oven?
MLB: Fruit
Ours: Cookies



Answers by our attention module MLB (ICLR'17) and VKMN (to appear in CVPR'18) with top-5 triple score

AGENDA

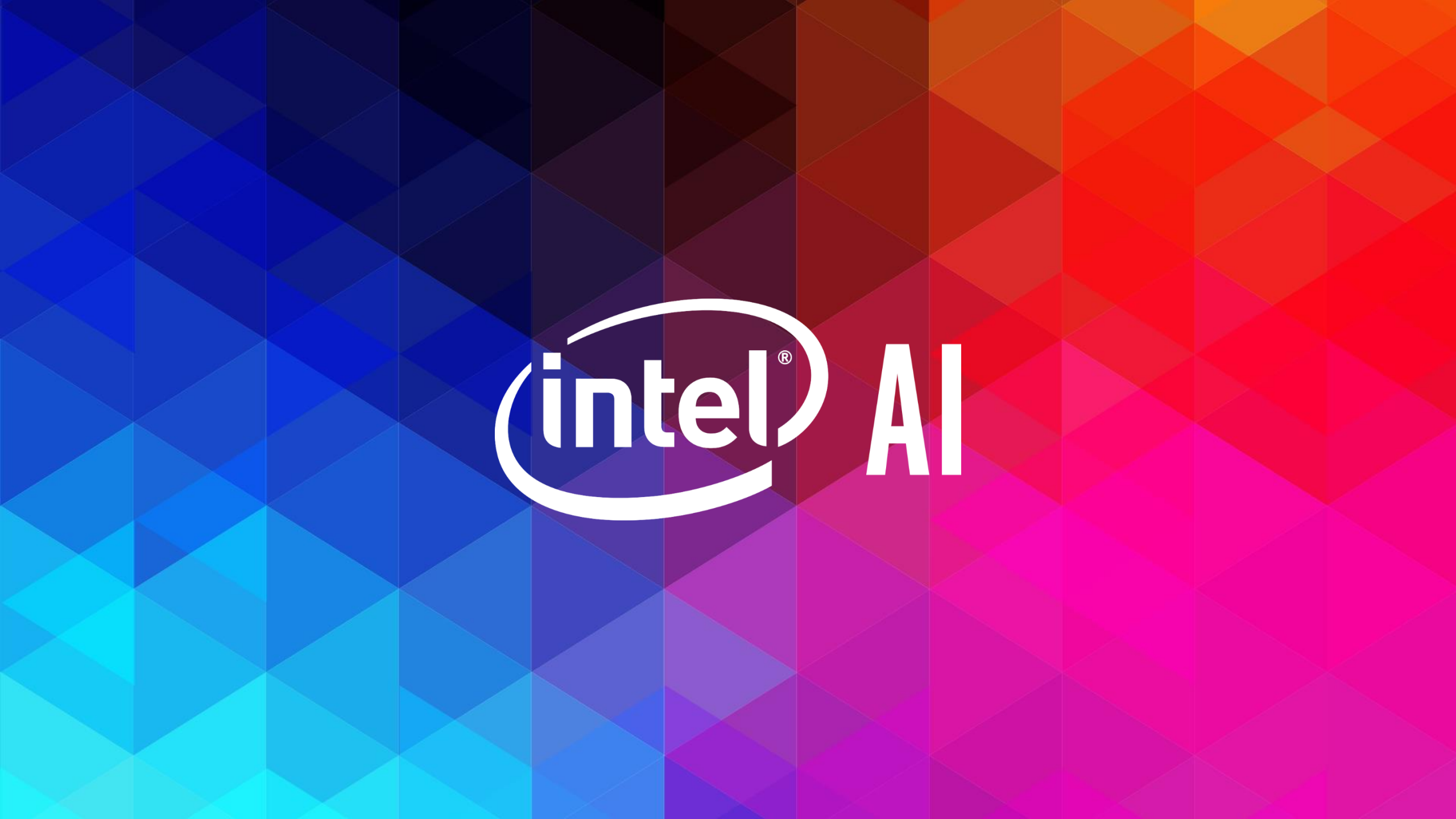
- Introduction
- Face Analysis & Emotion Recognition
- Deep Learning based Visual Recognition
- Visual Parsing & Multimodal Analysis
- Summary

SUMMARY

Visual Understanding research innovation to address visual data explosion challenges

Cutting-edge Deep Learning based VU research to impact Intel architectures/platforms/solutions and help differentiate Intel products

Call for more collaboration between universities and industry to accelerate research innovation for making sense of visual data


intel®
AI