NVIDIA DEEP LEARNING INSTITUTE

GTC 2017 ディープラーニング最新情報

エンタープライズ事業部　事業部長　井﨑 武士
NVIDIA

DEEP LEARNING関連セッション

- 合計670セッション中320セッション

種類	件数
講演	277
パネル	3
ハンズオン	24
ハングアウト	10
チュートリアル	6

SESSION 1

TRAINING OF DEEP NETWORKS WITH HALF-PRECISION FLOAT

Boris Ginsburg - Deep Learning Engineer, NVIDIA

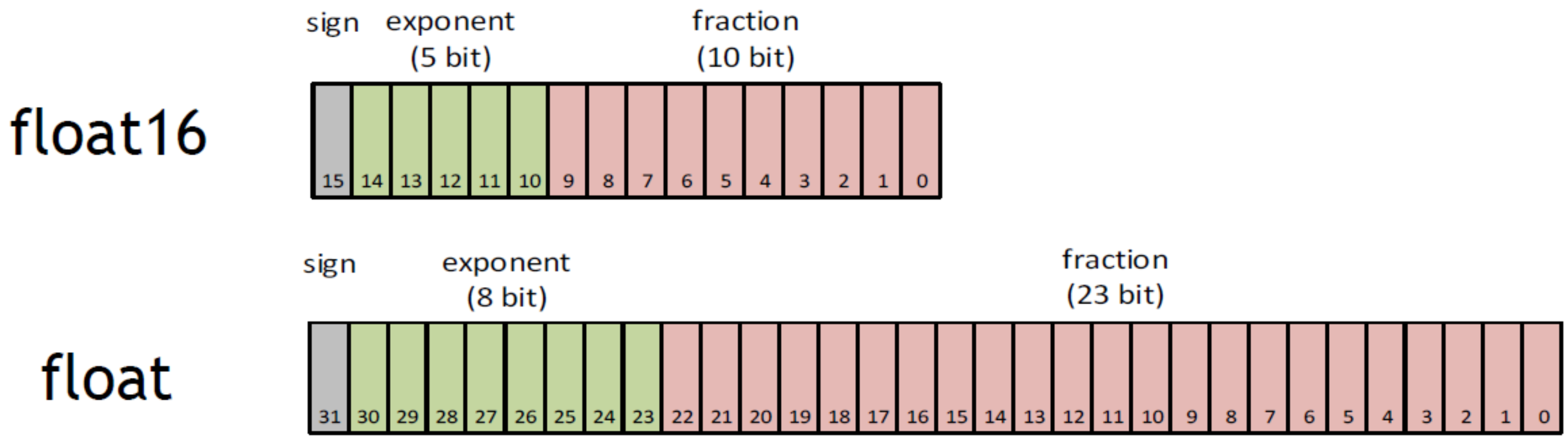
INTRODUCTION

Training with FLOAT16 has many potential benefits:

1. Smaller memory footprint:
   - ~2x if we keep weights, activations and gradients in FLOAT16 instead FLOAT
2. Faster training:
   - compute bounded layers (if HW supports FLOAT16 math - GP100)
   - memory bounded layers (ReLU, BatchNorm, …)
   - multi-GPU synchronization

Main challenge: narrow numerical range can result in underflow or overflow.

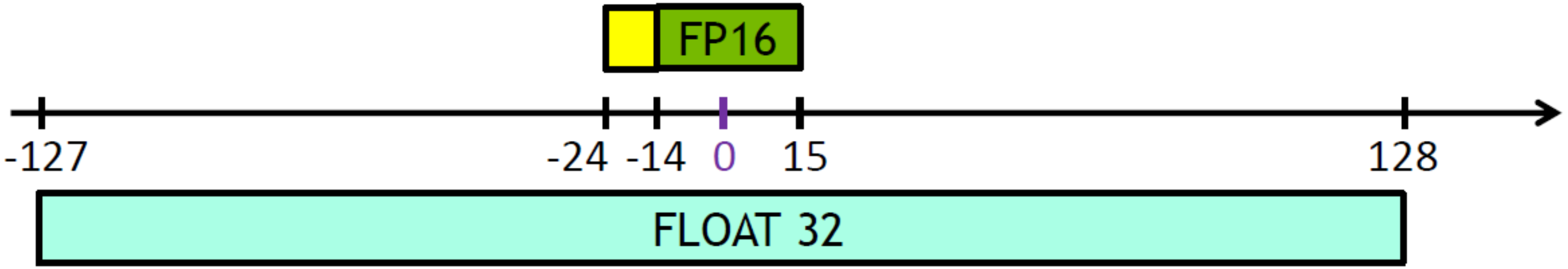
HALF-PRECISION FLOAT (FLOAT16)



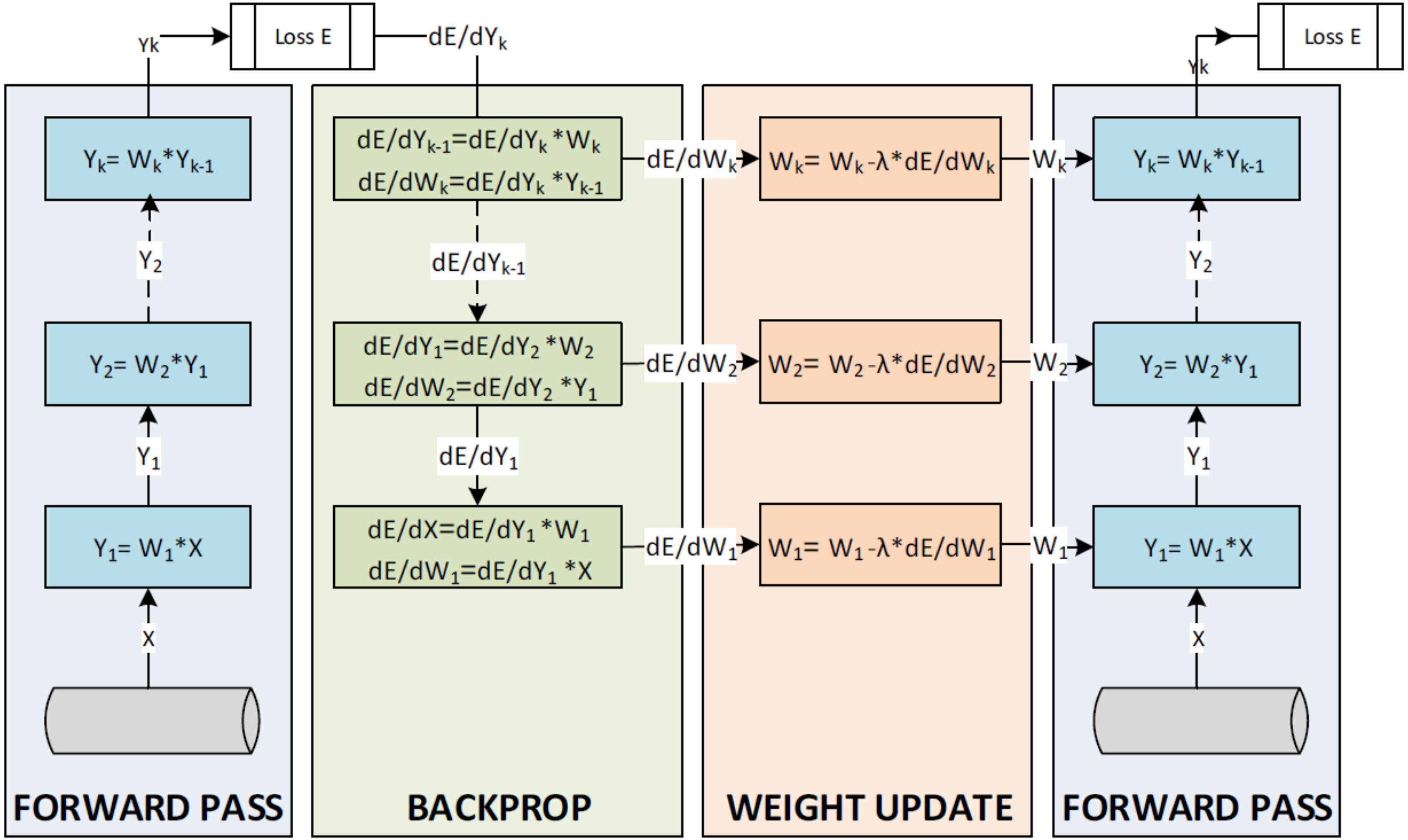
FLOAT16 has very narrow numerical range

Normal range: $[\ 6\times10^{-5}\ ,\ 65504\ ]$

Sub-normal range: $[\ 6\times10^{-8}\ ,\ 6\times10^{-5}\ ]$



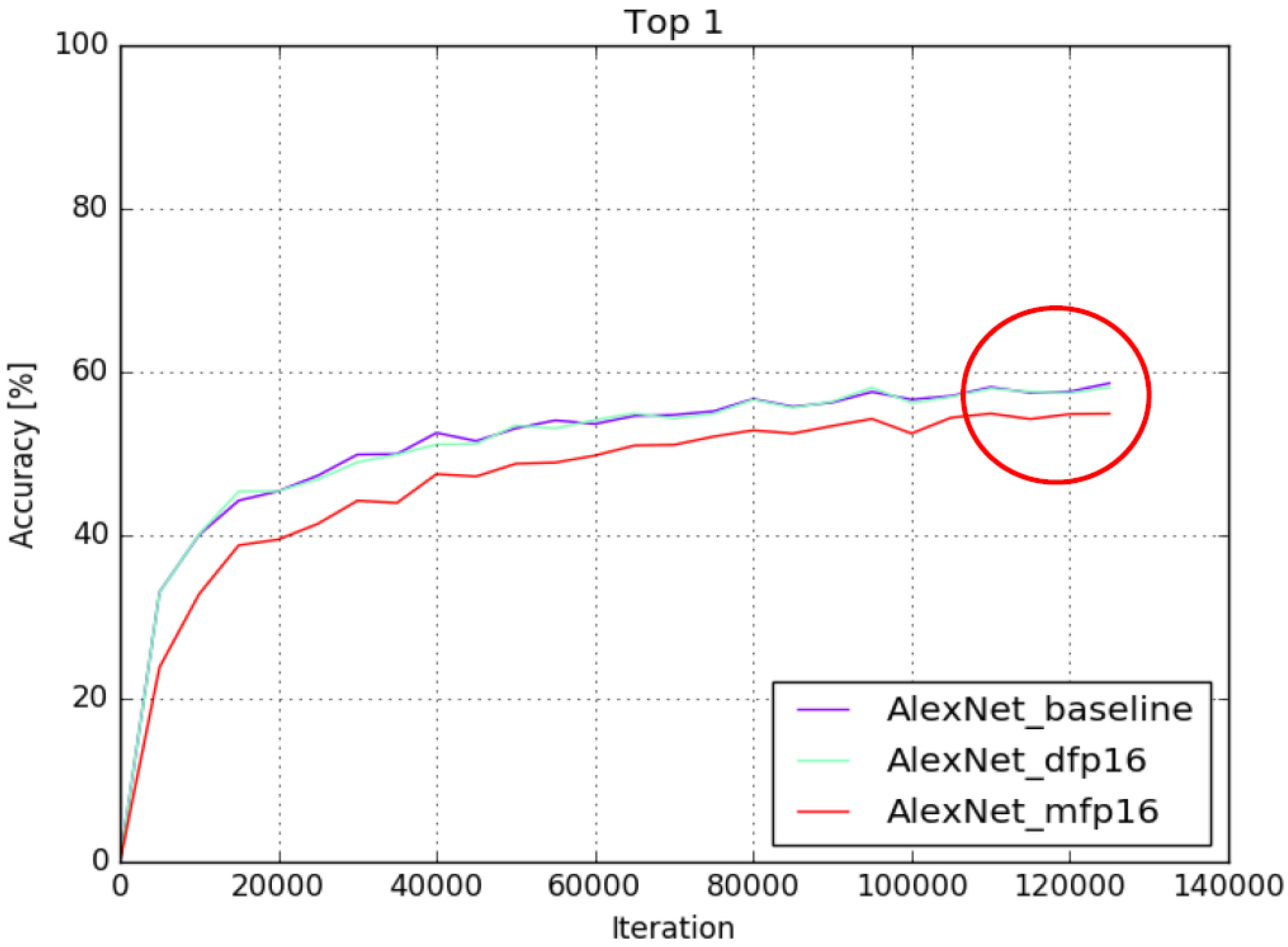
TRAINING FLOW



FLOAT16 MODES

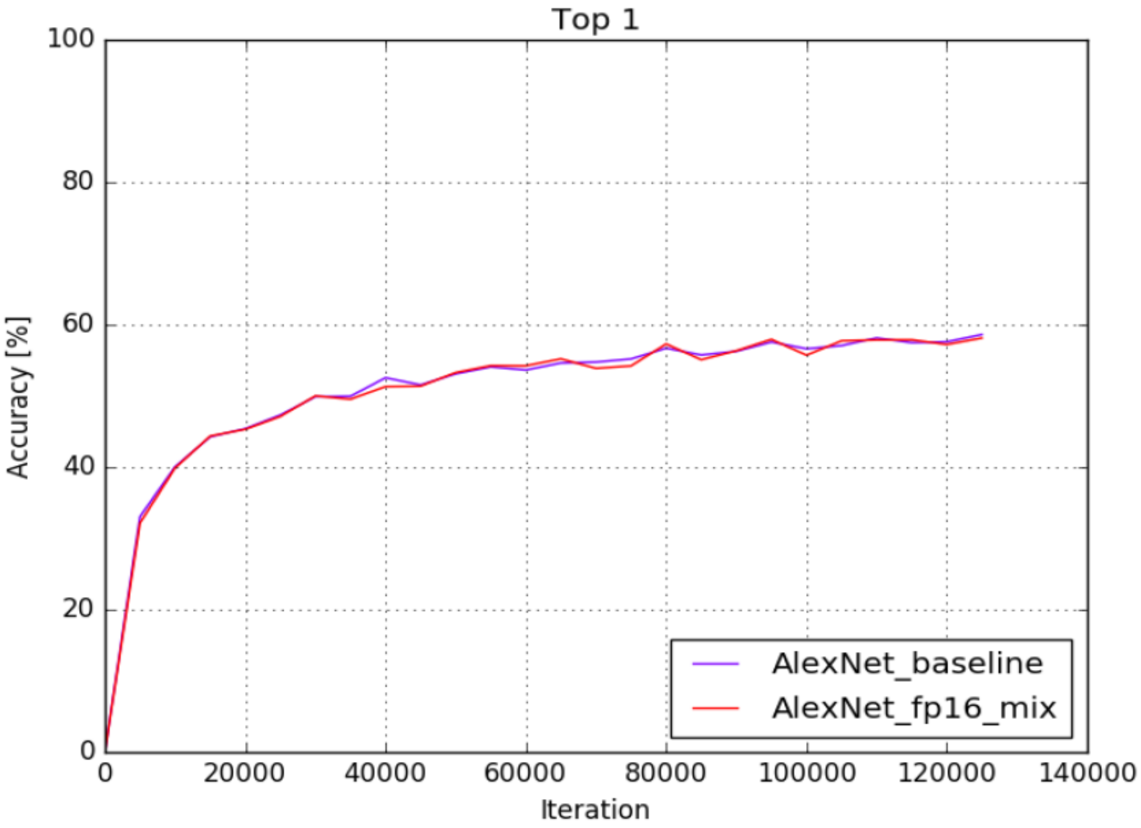
Mode	Data	Math	Update	Comment
Float	32	32	32	Baseline: all float
Dfp16	16	32	32	2 copy of weights: float16 for forward-backward and float for update
Mfp16	16	16	32	For GPUs with FP16 math
Nfp16	16	16	16	“Native” float16
Sfp16	16	32	16	For GPUs without FP16 math

ALEXNET: FLOAT16 MATH



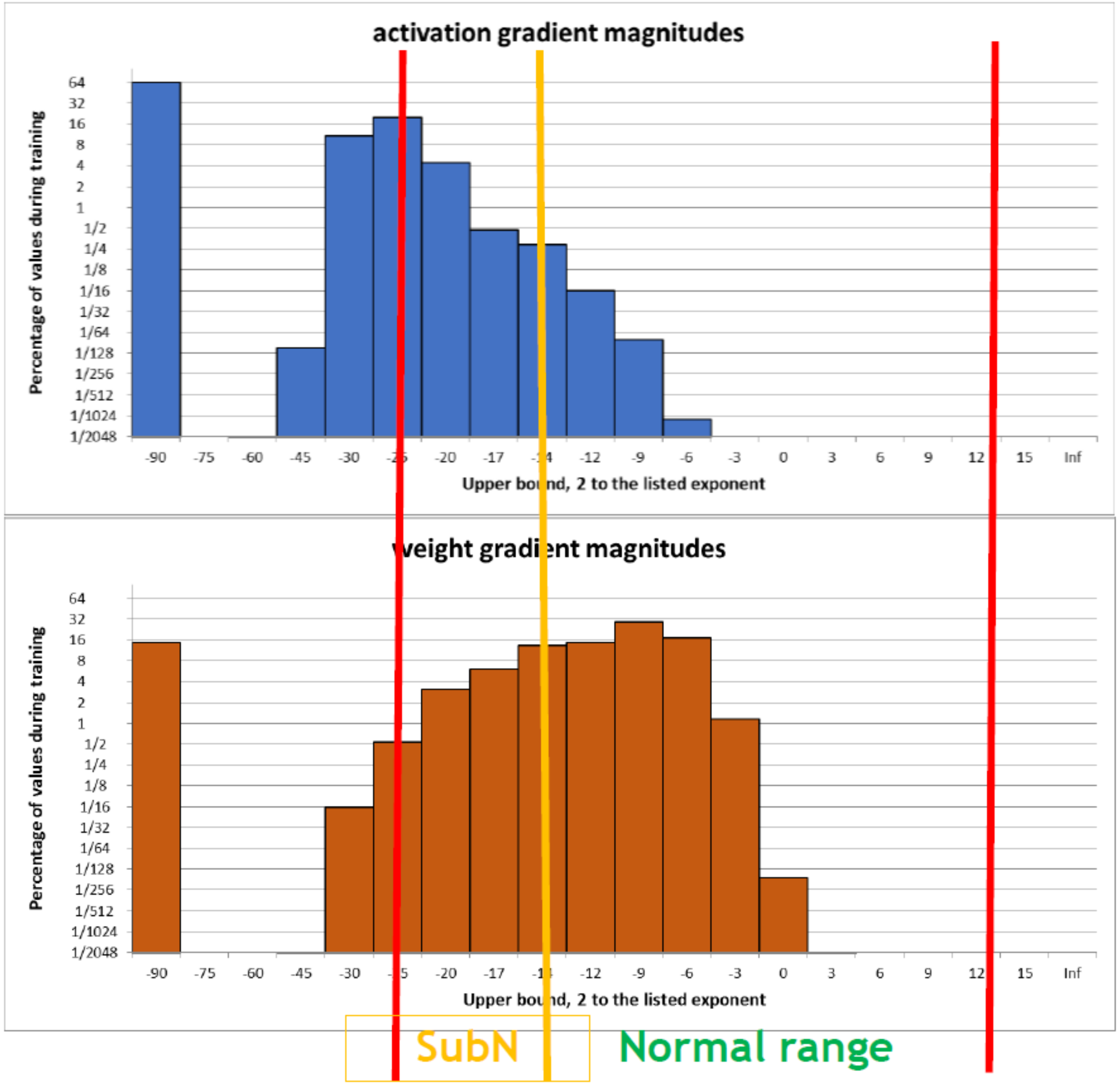
ALEXNET: MIXED MATH

Let's change backward_math from FLOAT16 to FLOAT



Accuracy is back! The problem is in the back-propagation

OBSERVATIONS ON GRADIENT VALUES



FP16 range is large (2^{40} with denorms)

Gradients use only low part of FP16 range

We can "shift" gradients to the right without overflowing

ALEXNET : FLOAT16 WITH SCALING

To shift gradients dE/dX we will scale up the loss function by constant (e.g. by 1000):

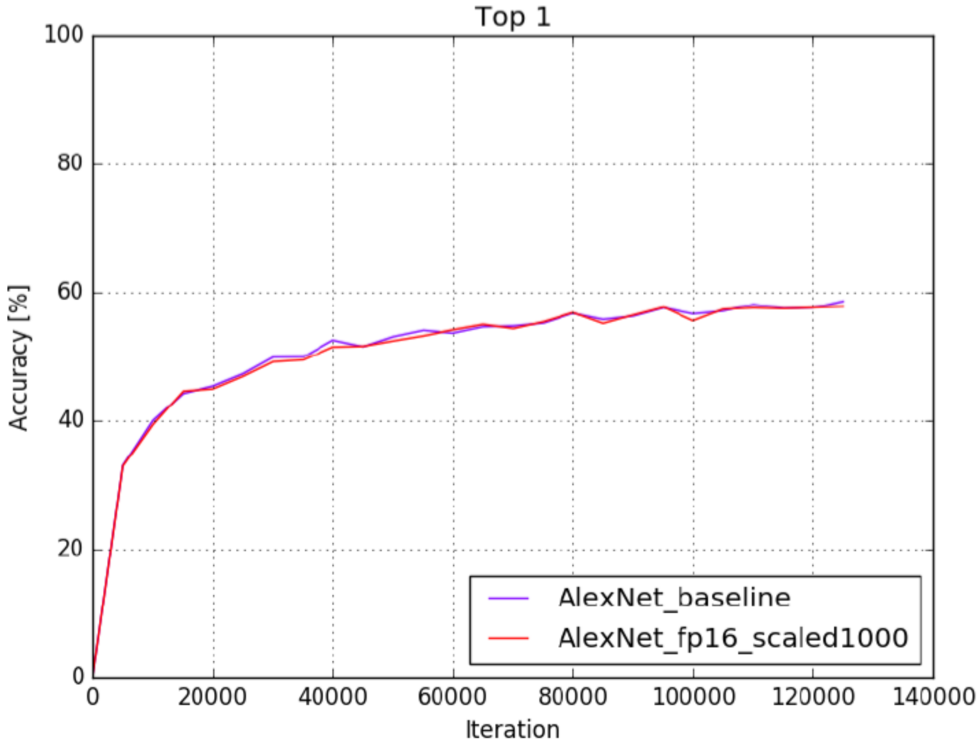
```
layer {
  type: "SoftmaxWithLoss"
  loss_weight: 1000.
}
```

and adjust learning rate and weight decay accordingly:

```
base_lr: ~~0.01~~ 0.00001            #    0.01 / 1000
weight_decay: ~~0.0005~~ 0.5         # 0.0005 * 1000
```

ALEXNET : FLOAT16 WITH SCALING

Mfp16 with scaling has the same accuracy as float!

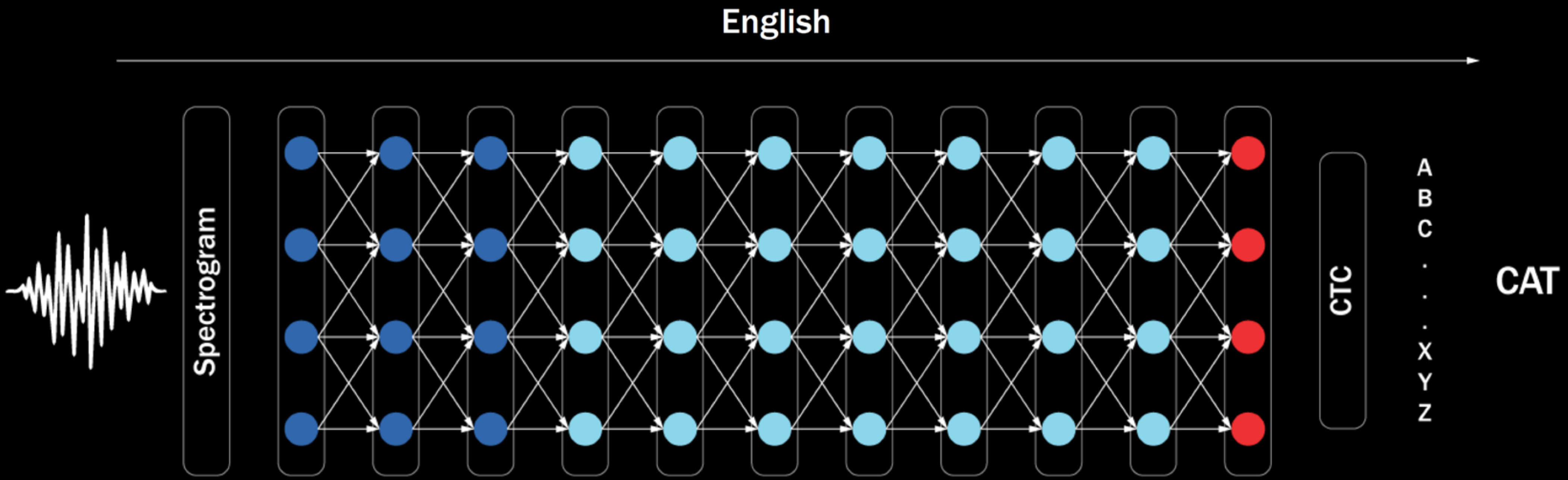


SESSION 2

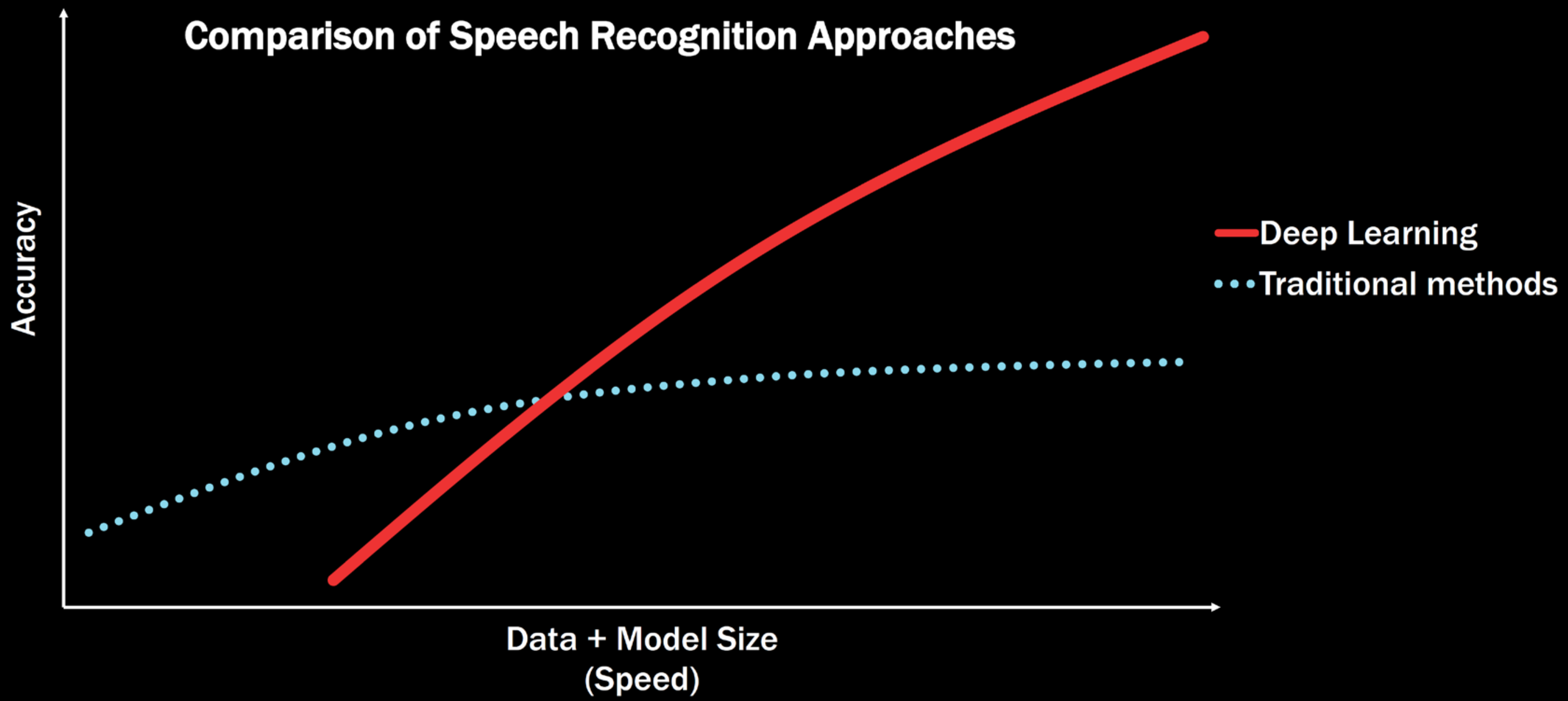
EXPLORING SPARSITY IN RECURRENT NEURAL NETWORKS

Sharan Narang - Researcher, Baidu

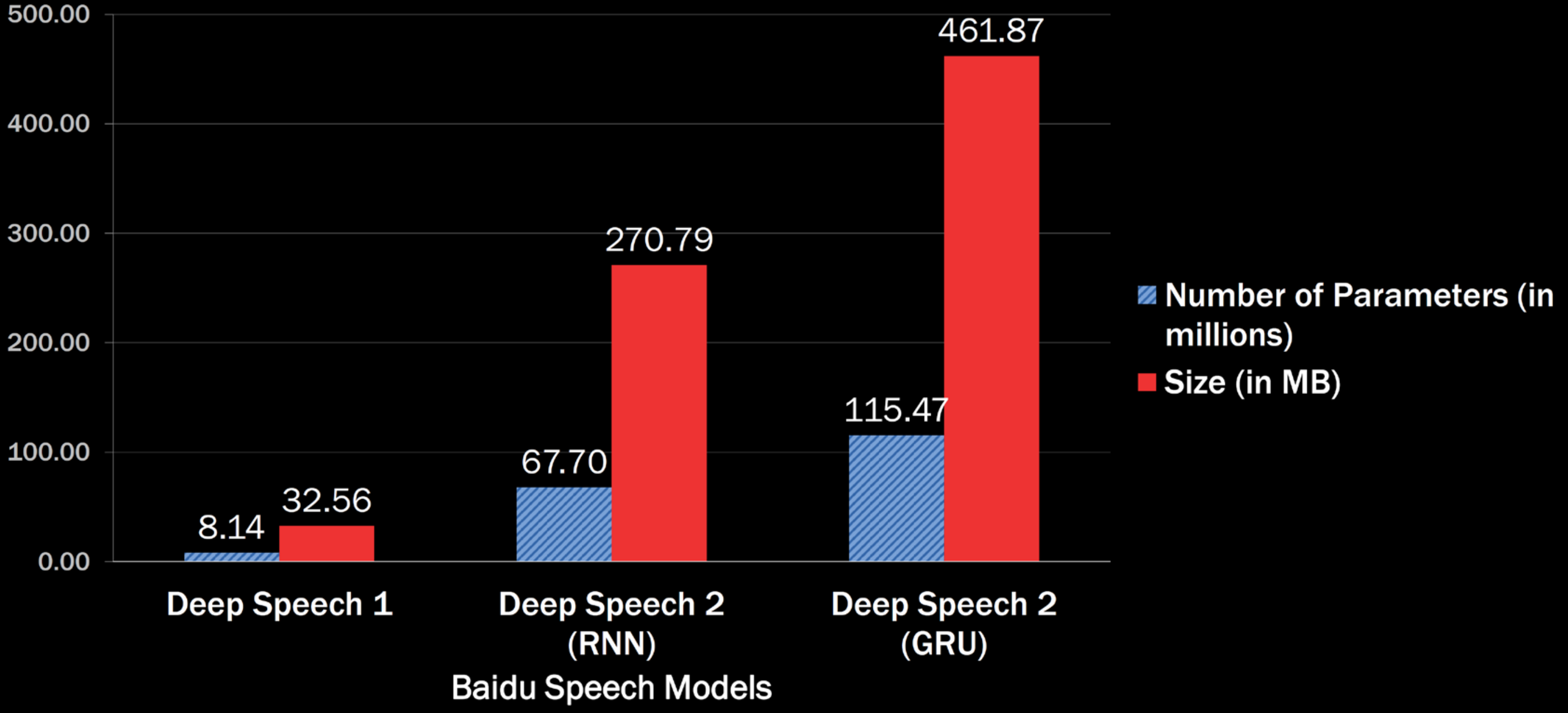
Speech Recognition with Deep Learning



Scaling with Data



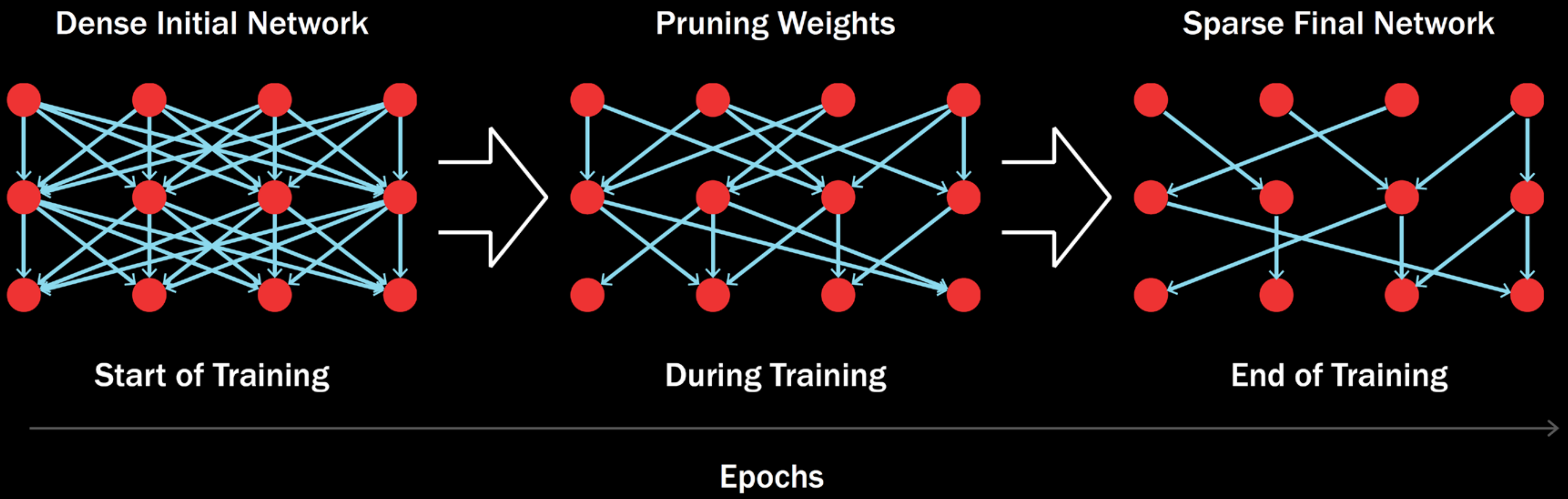
Model Sizes



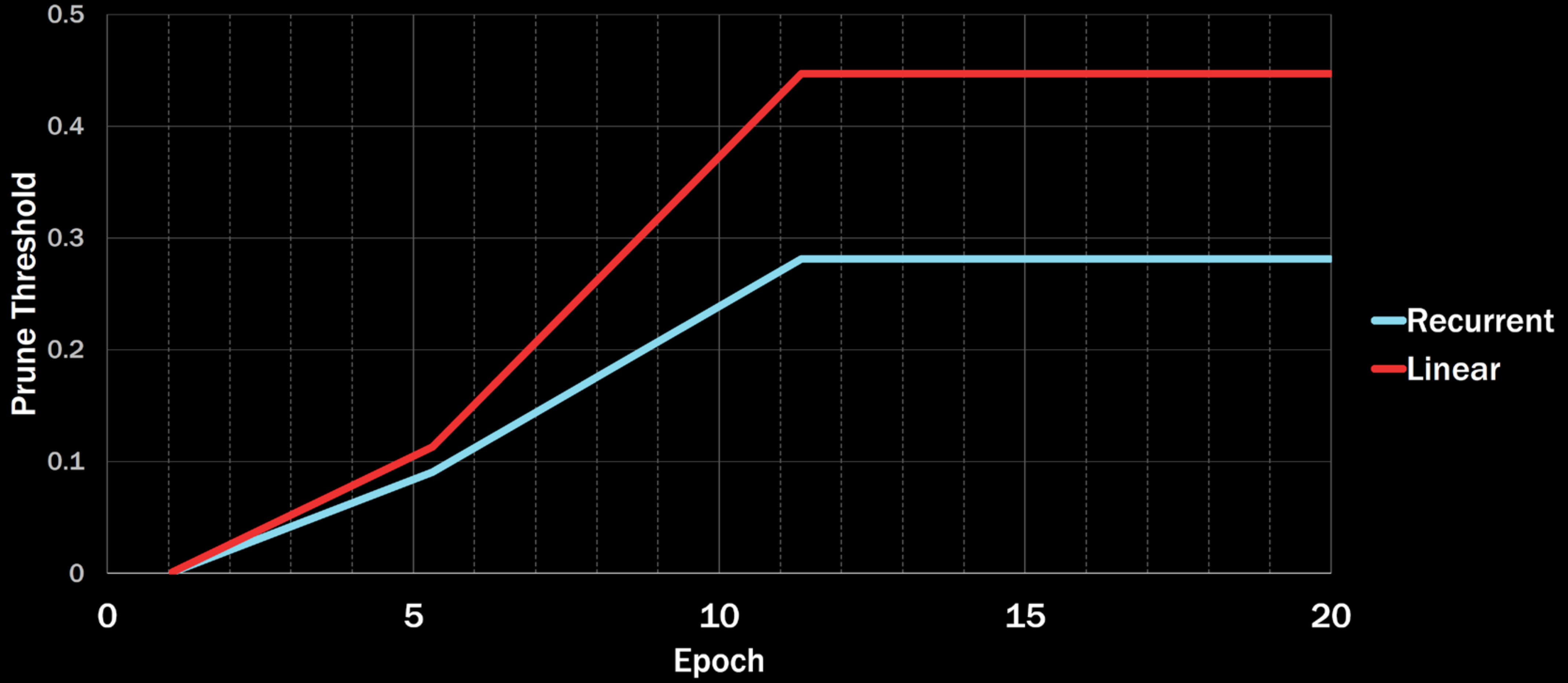
500.00
400.00
300.00
200.00
100.00
0.00
461.87
270.79
115.47
67.70
32.56
8.14
Number of Parameters (in millions)
Size (in MB)
Deep Speech 1
Deep Speech 2 (RNN)
Deep Speech 2 (GRU)
Baidu Speech Models

Sparse Neural Networks

Pruning Weights



Pruning Approach



Results

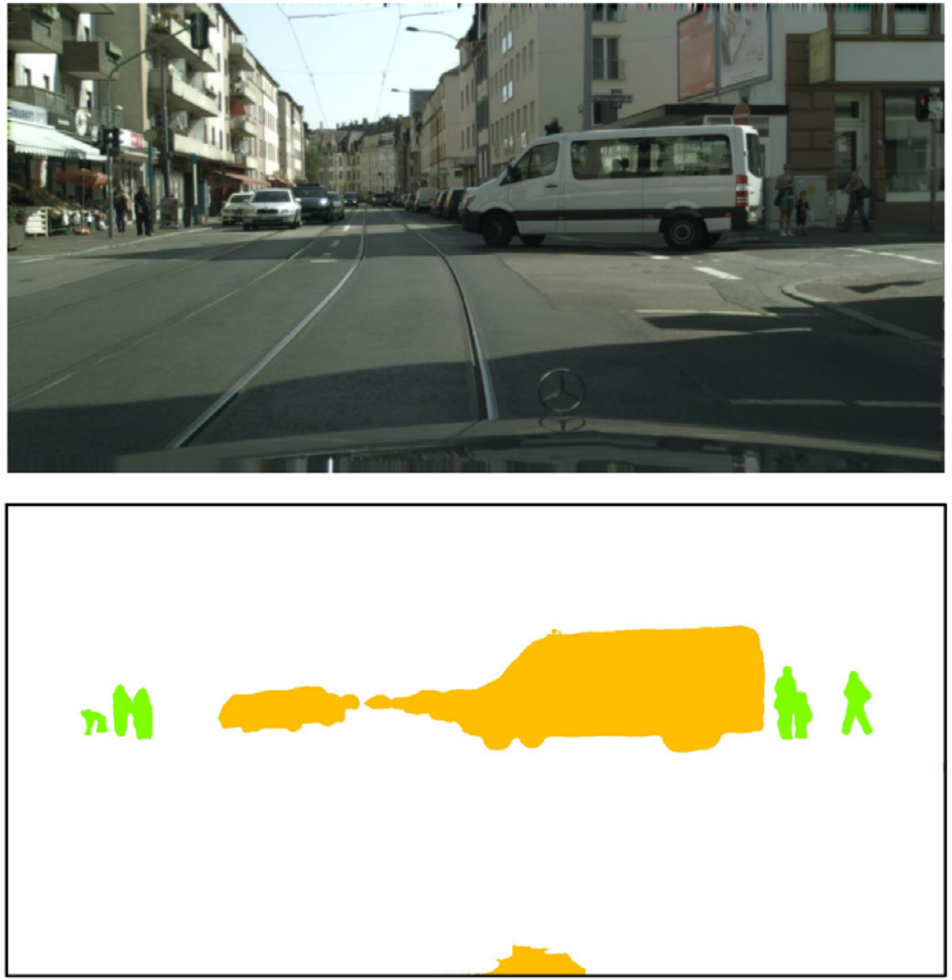
Model	Layer Size	# of Params	CER	Relative Perf
RNN Dense	1760	67 million	10.67	0.0%
RNN Sparse	1760	8.3 million	12.88	-20.71%
RNN Sparse	2560	11.1 million	10.59	0.75%
RNN Sparse	3072	16.7 million	10.25	3.95%
GRU Dense	2560	115 million	9.55	0.0%
GRU Sparse	2560	13 million	10.87	-13.82%
GRU Sparse	3568	17.8 million	9.76	-2.2%

SESSION 3

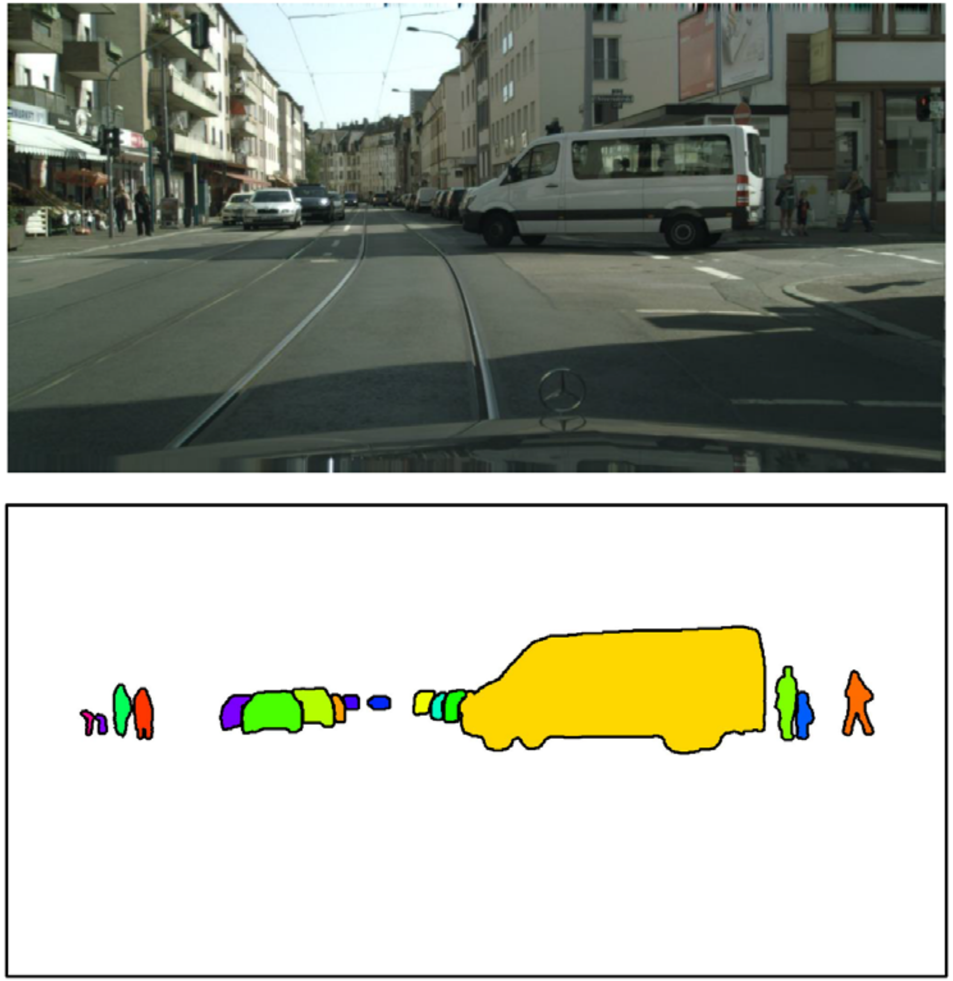
DEEP WATERSHED TRANSFORM FOR INSTANCE SEGMENTATION

Min Bai - PhD Student, University of Toronto

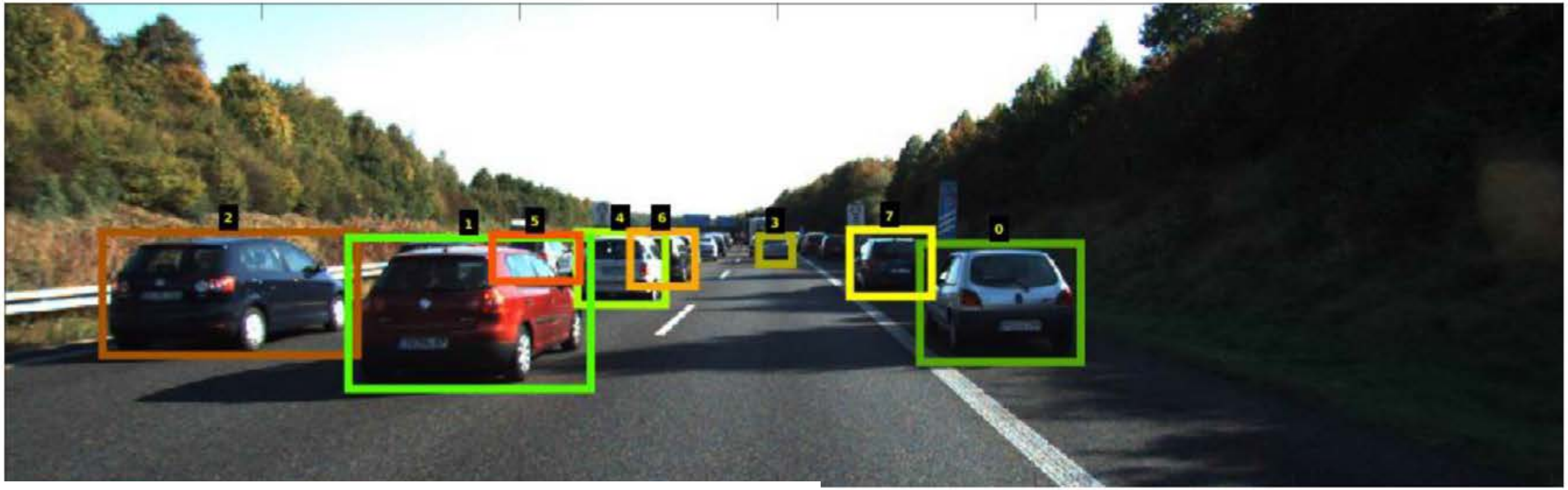
Semantic Segmentation

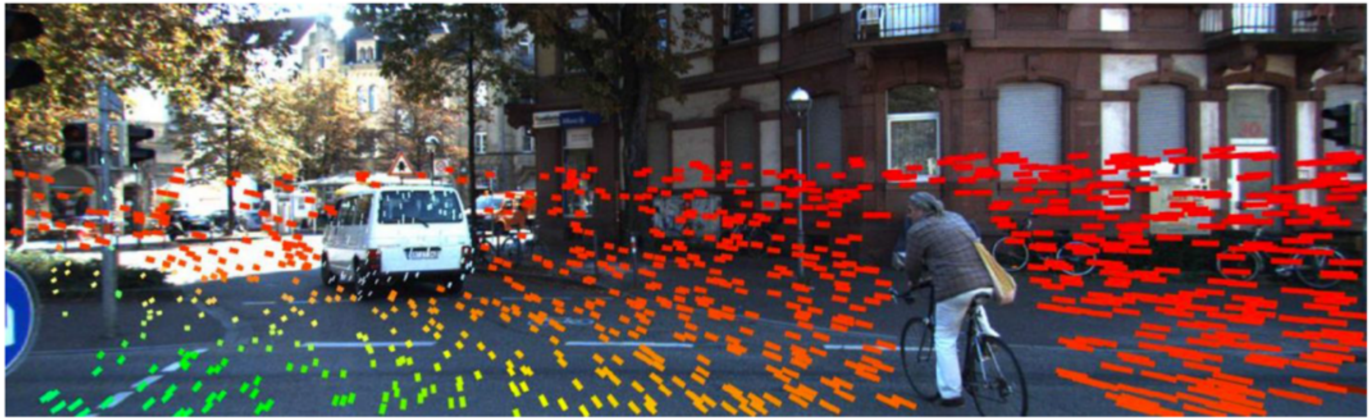


Instance Segmentation



DEEP
LEARNING
INSTITUTE


2
1
5
4
6
3
7
0



Semantic Segmentation

- Semantic segmentation is a well studied problem
  - Our instance segmentation method leverages an existing technique
  - H. Zhao et al, *Pyramid Scene Parsing Network*, https://arxiv.org/abs/1612.01105

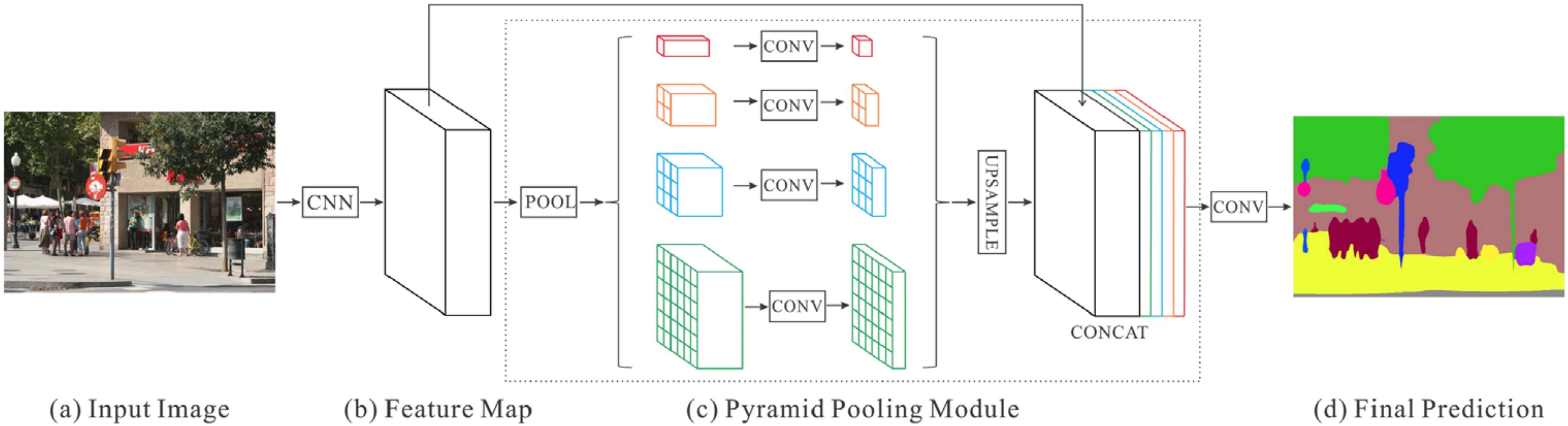
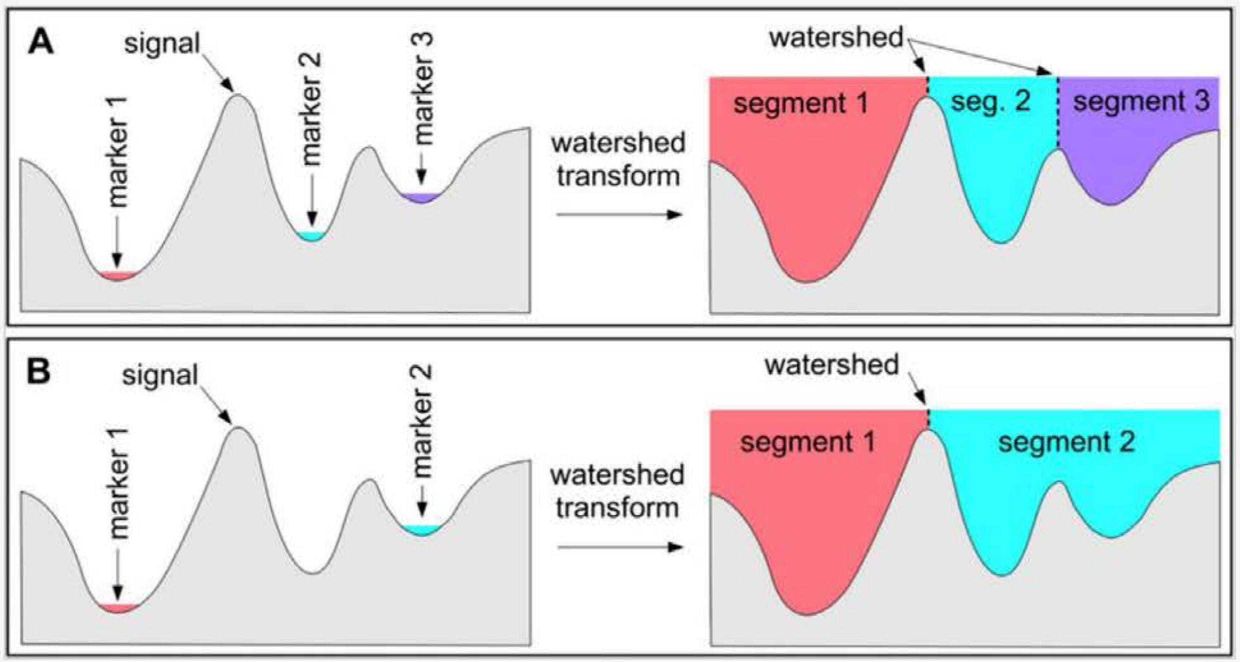
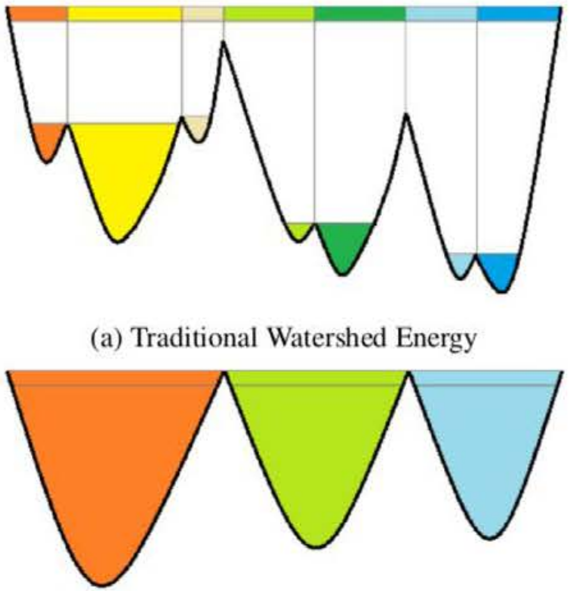


Image credit: H. Zhao et al.

Watershed Transform

- Classical image segmentation technique





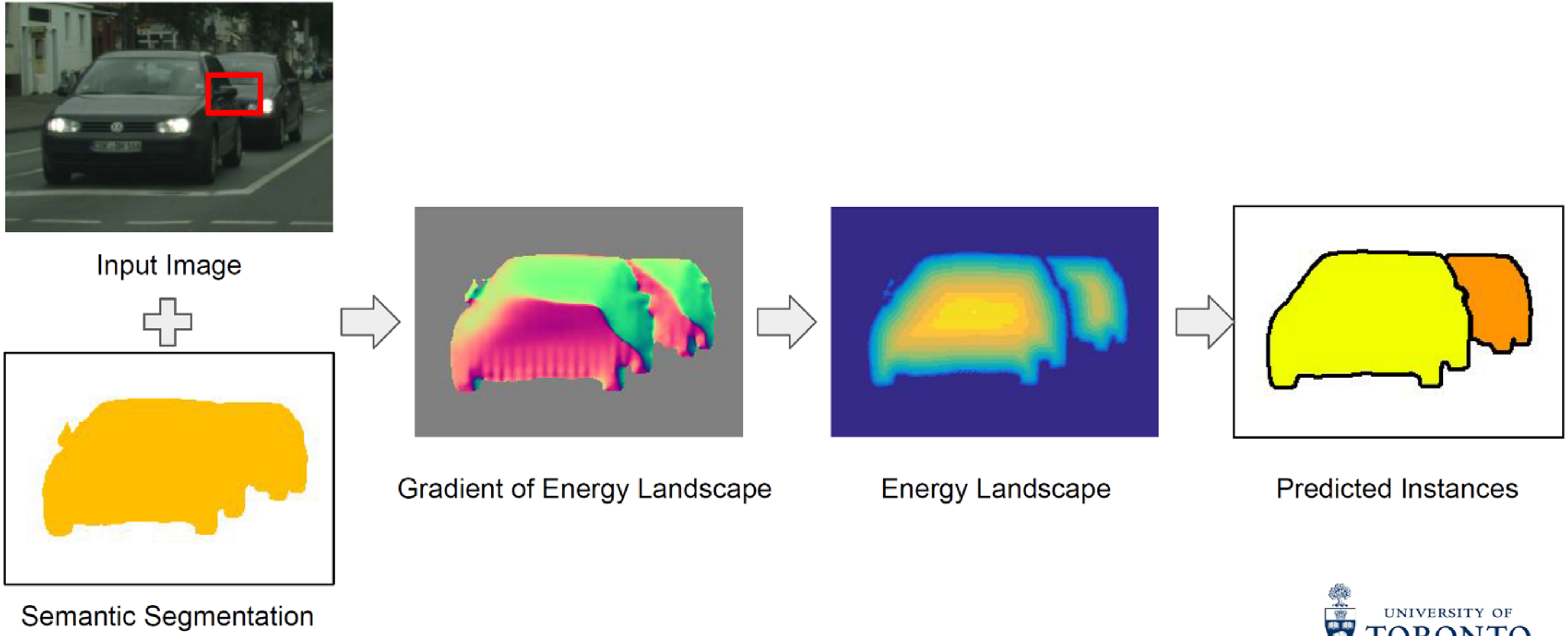
(a) Traditional Watershed Energy

(b) Our learned energy

Image (left) credit: Adrian Fisher

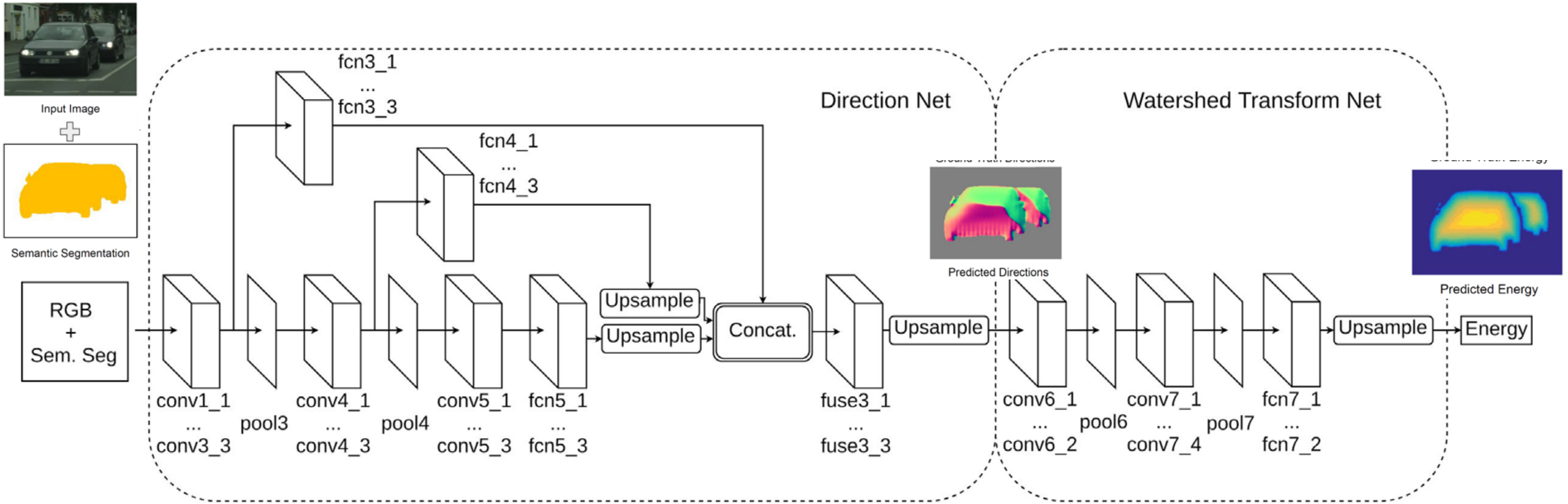
UNIVERSITY OF TORONTO

Overview of Approach



UNIVERSITY OF TORONTO

Overall Network



UNIVERSITY OF TORONTO

Cityscapes Instance Segmentation Leaderboard

	AP*	AP* @ 50%	AP* @ 50m	AP* @ 100m
van den Brand et al.	2.3%	3.7%	3.9%	4.9%
Cordts et al.	4.6%	12.9%	7.7%	10.3%
Uhrig et al.	8.9%	21.1%	15.3%	16.7%
Ours	**19.4%**	**35.3%**	**31.4%**	**36.8%**

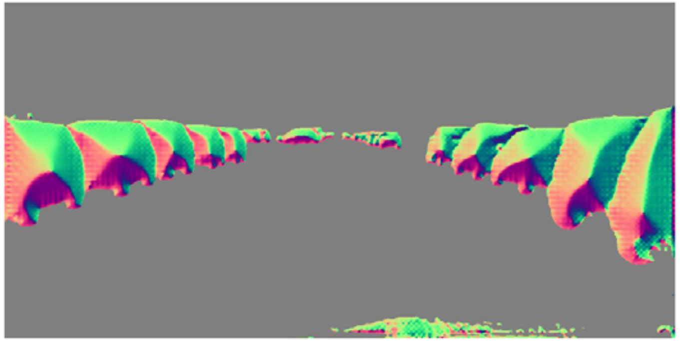
* Average Precision (AP): higher is better

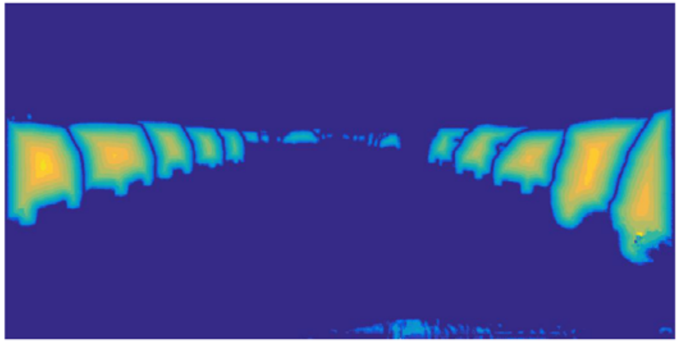
Recently, new approaches have achieved even higher performance.

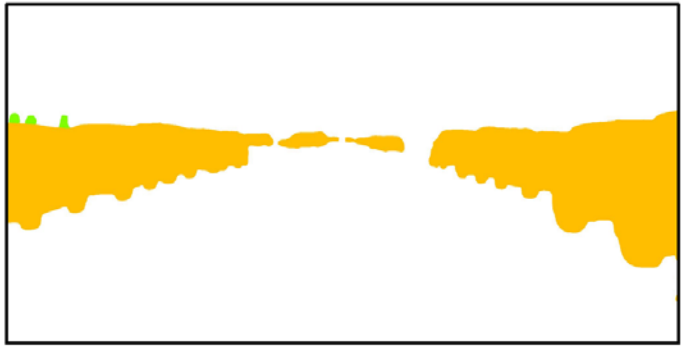
UNIVERSITY OF TORONTO

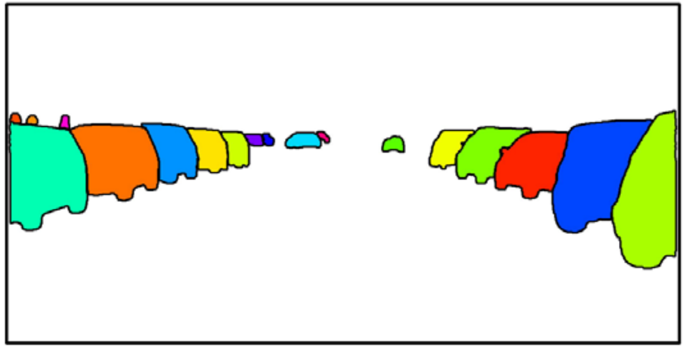
Sample Output

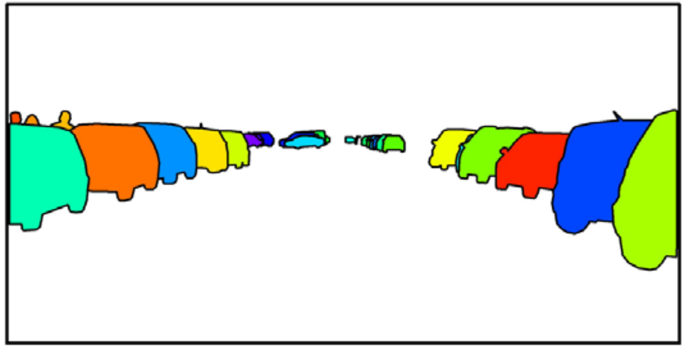

Input RGB


Direction Prediction


Energy Prediction

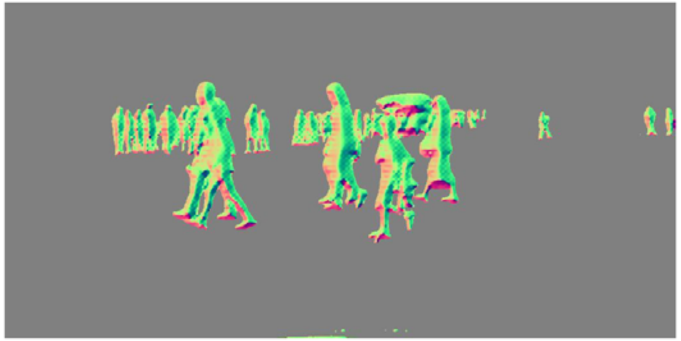

Semantic Segmentation


Predicted Instances

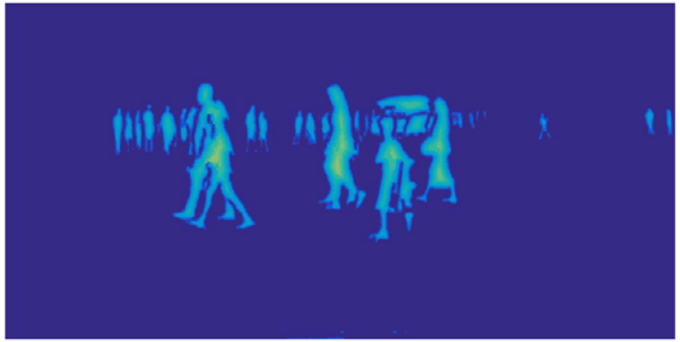

Ground Truth Instances

Sample Output

Input RGB

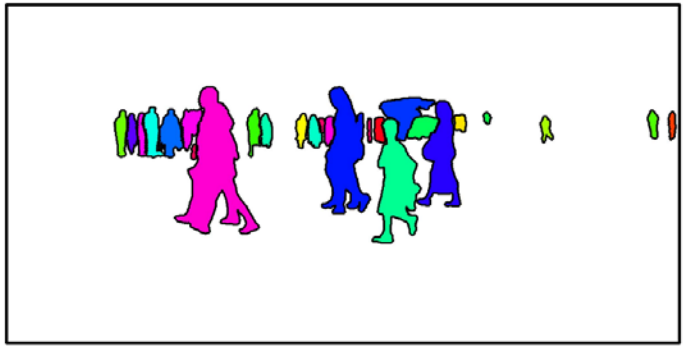


Direction Prediction

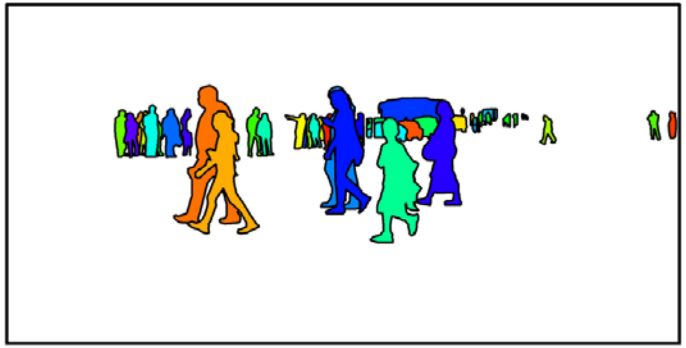


Energy Prediction

Semantic Segmentation



Predicted Instances



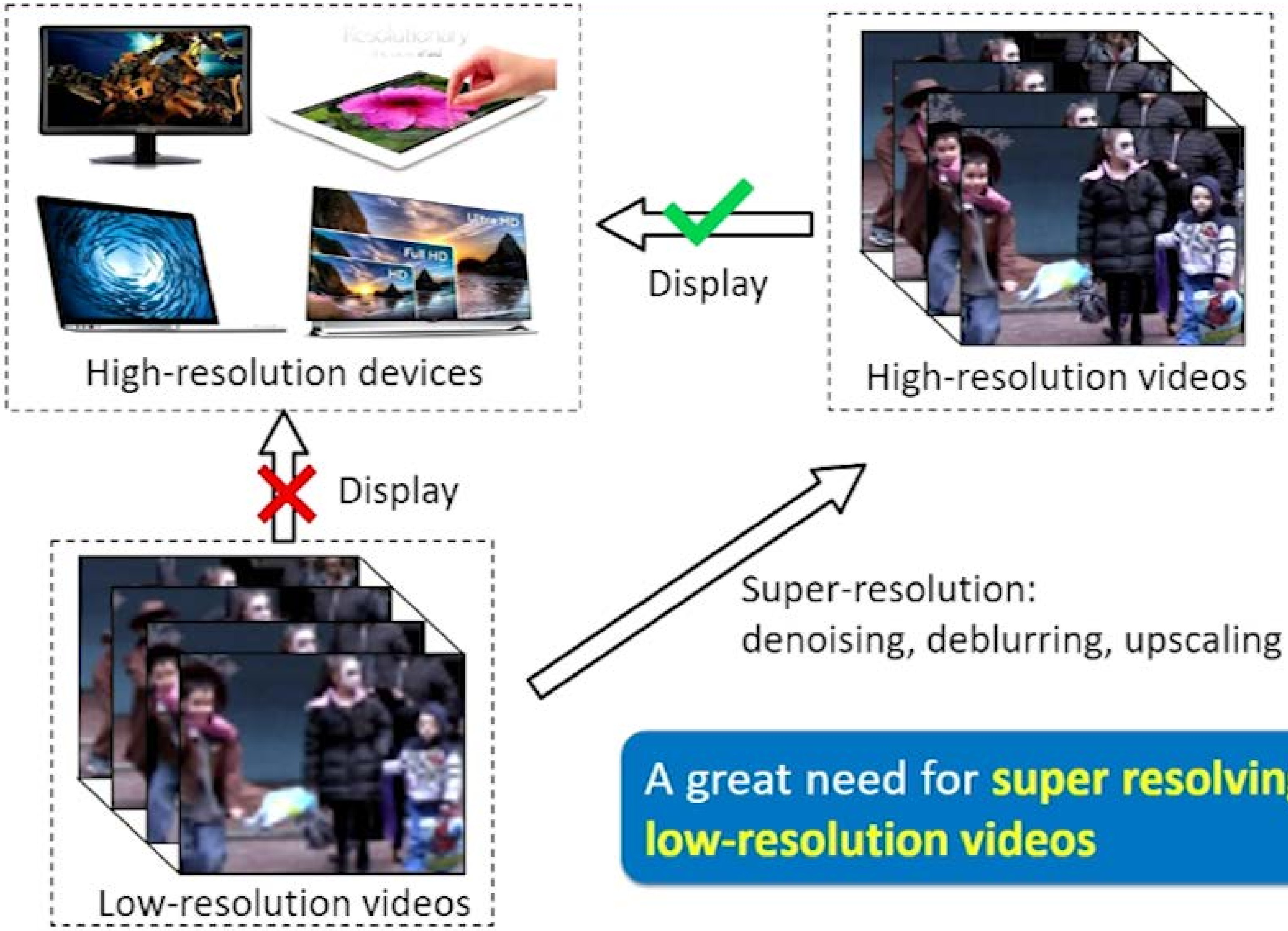
Ground Truth Instances

SESSION 4

BIDIRECTIONAL RECURRENT CONVOLUTIONAL NETWORKS AND THEIR APPLICATIONS TO VIDEO SUPER-RESOLUTION

Qi Zhang - Assistant Professor, Chinese Academy of Sciences, Institute of Automation

Video Super-Resolution



A great need for **super resolving low-resolution videos**

1. Single-Image super-resolution [1-6]

One-to-One scheme, super resolve each video frame independently

Ignore the intrinsic temporal dependency relation of video frames

Low computational complexity, fast

2. Multi-Frame super-resolution [7-11]

Many-to-One scheme, use multiple adjacent frames to super resolve a frame

Model the temporal dependency relation by motion estimation

High computational complexity, slow

Motivation

RNN: Recurrent Neural Networks
SR: Super-Resolution

- RNN can model **long-term contextual information** of temporal sequences well
- Convolutional operation can **scale to full videos** of any spatial size and temporal step

➢ Propose bidirectional recurrent convolutional networks, different from vanilla RNN:

1. Commonly-used full connections are replaced with weight -sharing convolutions

2. Conditional convolutions are added for learning visual-temporal dependency relation

Experiments

- Train the model on 25 YUV format video sequences
  - volume-based training
  - number of volumes: roughly 41,000
  - volume size: $32 \times 32 \times 10$
- Test on a variety of real world videos
  - severe motion blur
  - motion aliasing
  - complex motions

Training videos

Testing videos

PSNR Comparison

PSNR: peak signal-to-noise ratio

Table1: The results of PSNR (dB) and test time (sec) on the test video sequences.

Video	Bicubic		SC [25]		K-SVD [26]		NE+NNLS [4]		ANR [23]	
	PSNR	Time	PSNR	Time	PSNR	Time	PSNR	Time	PSNR	Time
Dancing	26.83	-	26.80	45.47	27.69	2.35	27.63	19.89	27.67	0.85
Flag	26.35	-	26.28	12.89	27.61	0.58	27.41	4.54	27.52	0.20
Fan	31.94	-	32.50	12.92	33.55	1.06	33.45	8.27	33.49	0.38
Treadmill	21.15	-	21.27	15.47	22.22	0.35	22.08	2.60	22.24	0.12
Turbine	25.09	-	25.77	16.49	27.00	0.51	26.88	3.67	27.04	0.18
Average	26.27	-	26.52	20.64	27.61	0.97	27.49	7.79	27.59	0.35

Video	NE+LLE [5]		SR-CNN [6]		3DSKR [21]		Enhancer [1]		BRCN	
	PSNR	Time	PSNR	Time	PSNR	Time	PSNR	Time	PSNR	Time
Dancing	27.64	4.20	27.81	1.41	27.81	1211	27.06	-	**28.09**	3.44
Flag	27.48	0.96	28.04	0.36	26.89	255	26.58	-	**28.55**	0.78
Fan	33.46	1.76	33.61	0.60	31.91	323	32.14	-	**33.73**	1.46
Treadmill	22.22	0.57	22.42	0.15	22.32	127	21.20	-	**22.63**	0.46
Turbine	26.98	0.80	27.50	0.23	24.27	173	25.60	-	**27.71**	0.70
Average	27.52	1.66	27.87	0.55	26.64	418	26.52	-	**28.15**	1.36

Surpass state-of-the-art methods in PSNR, due to the effective temporal dependency modelling

[1] V
[4] B
[5] C
[6] D
[20] Takeda et al., Super-resolution without explicit subpixel motion estimation. IEEE TIP, 2009.
[22] Timofte et al., Anchored neighborhood regression for fast example-based super resolution. ICCV, 2013.
[24] Yang et al., Image super-resolution via sparse representation. IEEE TIP, 2010.
[25] Zeyde et al., On single image scale-up using sparse-representations. Curves and Surfaces, 2012.

Model Architecture

- **Investigate the impact of our model architecture on the performance**
- Take a simplified network containing only feedfoward (v) convolution as a benchmark
- Study its variants by successively adding the bidirectional (b), recurrent (r)and conditional (t) schemes

Table1: The results of PSNR (dB) by variants of BRCN on the testing video sequences.

Video	BRCN $\{v\}$	BRCN $\{v,r\}$	BRCN $\{v,t\}$	BRCN $\{v,r,t\}$	BRCN $\{v,r,t,b\}$
Dancing	27.81	27.98	27.99	28.09	28.09
Flag	28.04	28.32	28.39	28.47	28.55
Fan	33.61	33.63	33.65	33.65	33.73
Treadmill	22.42	22.59	22.56	22.59	22.63
Turbine	27.50	27.47	27.50	27.62	27.71
Average	27.87	27.99	28.02	28.09	28.15

Example

Upscaling factor:4

$$87 \times 157 \rightarrow 348 \times 628$$

Comparison:

Bicubic (top)

Ours (bottom)

SESSION 5

REAL-TIME LIVE VIDEO HIGHLIGHT IDENTIFICATION AT SCALE: LESSONS LEARNED FROM YAHOO ESPORTS

Yale Song - Senior Research Scientist, Yahoo Research

Bin Ni - Distinguished Software Architect, Yahoo

Overview

Esports Events

Live video → **Video Highlight Detection** → Highlights

Yahoo Esports

"This is exciting!"

Typical Scenes in Esports Video

Game


Highlight

Replay


Character Draft


Commentator

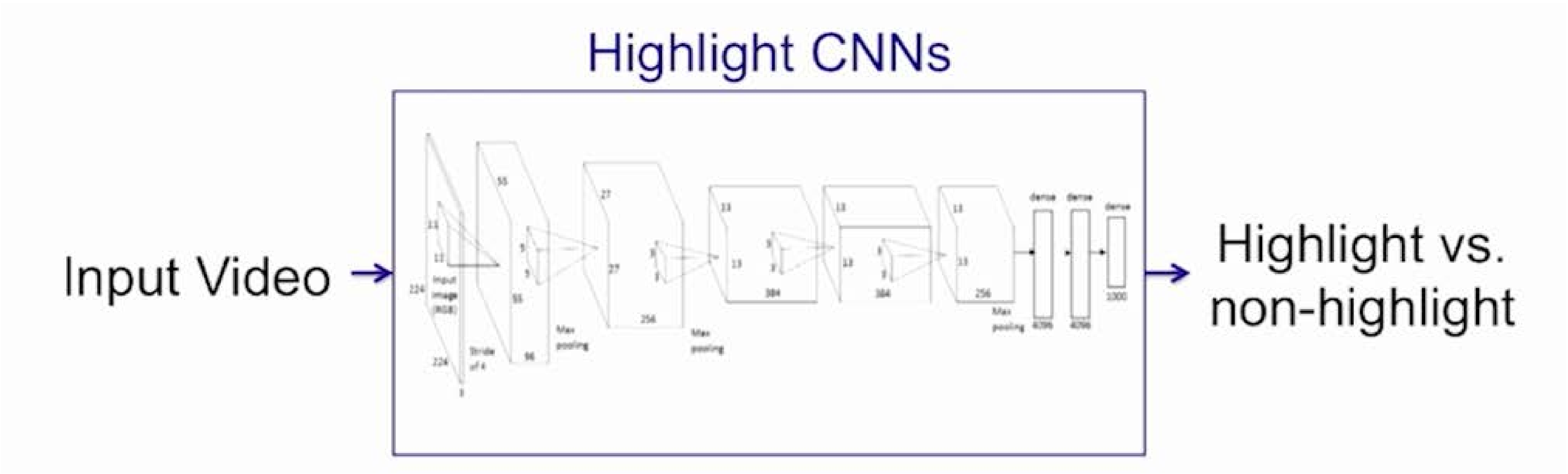

Interview


Player

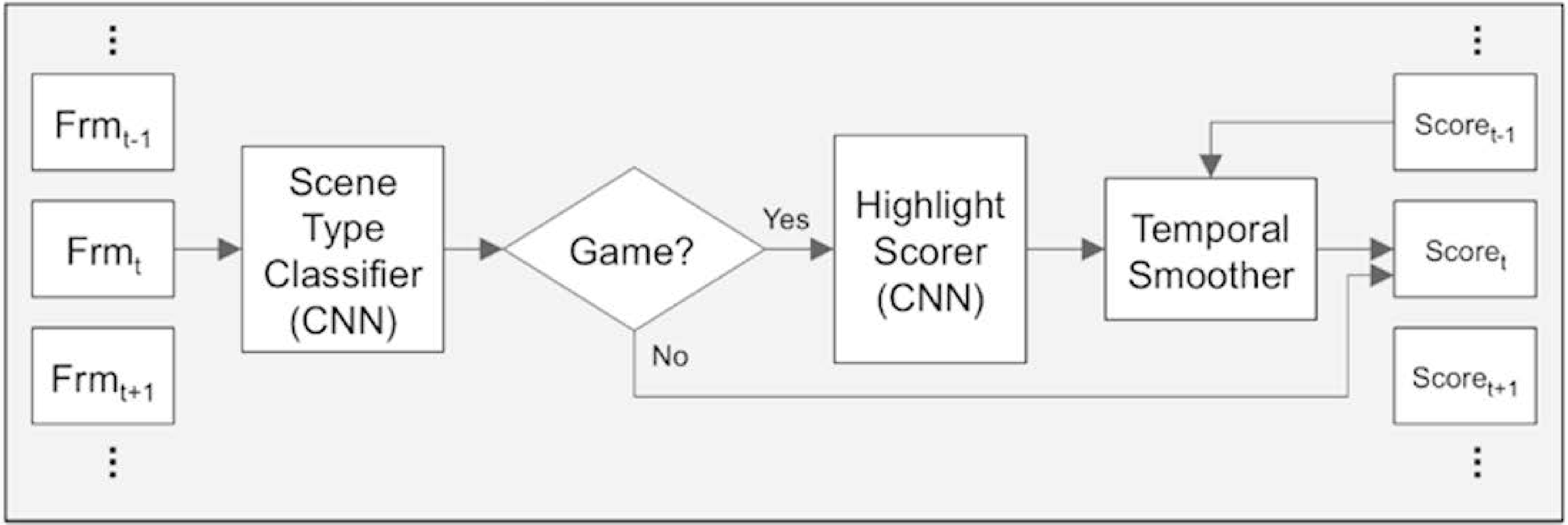

Audience

Game credit: ***Heroes of the Storm*** by Blizzard Entertainment

HIGHLIGHT CNN



Cascaded prediction



Scene type categorization

Multi-class classification (game, replay, studio, audience, …)

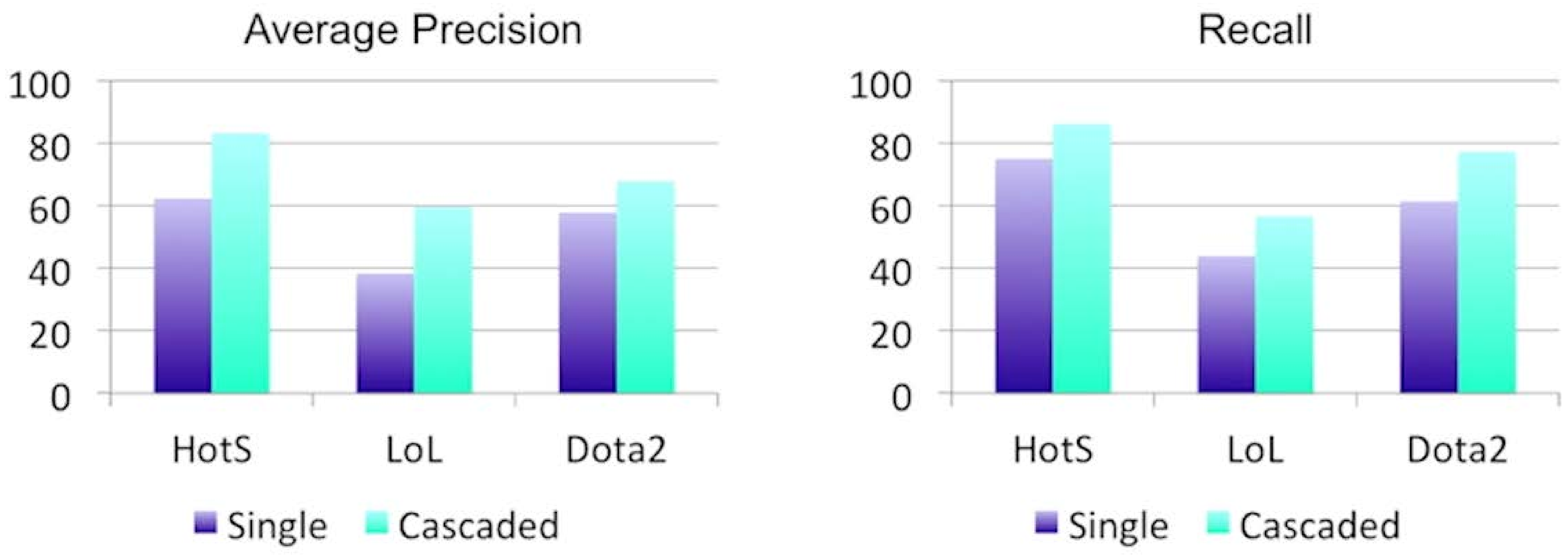
Highlight detection

Binary classification (highlight vs. non-highlight)

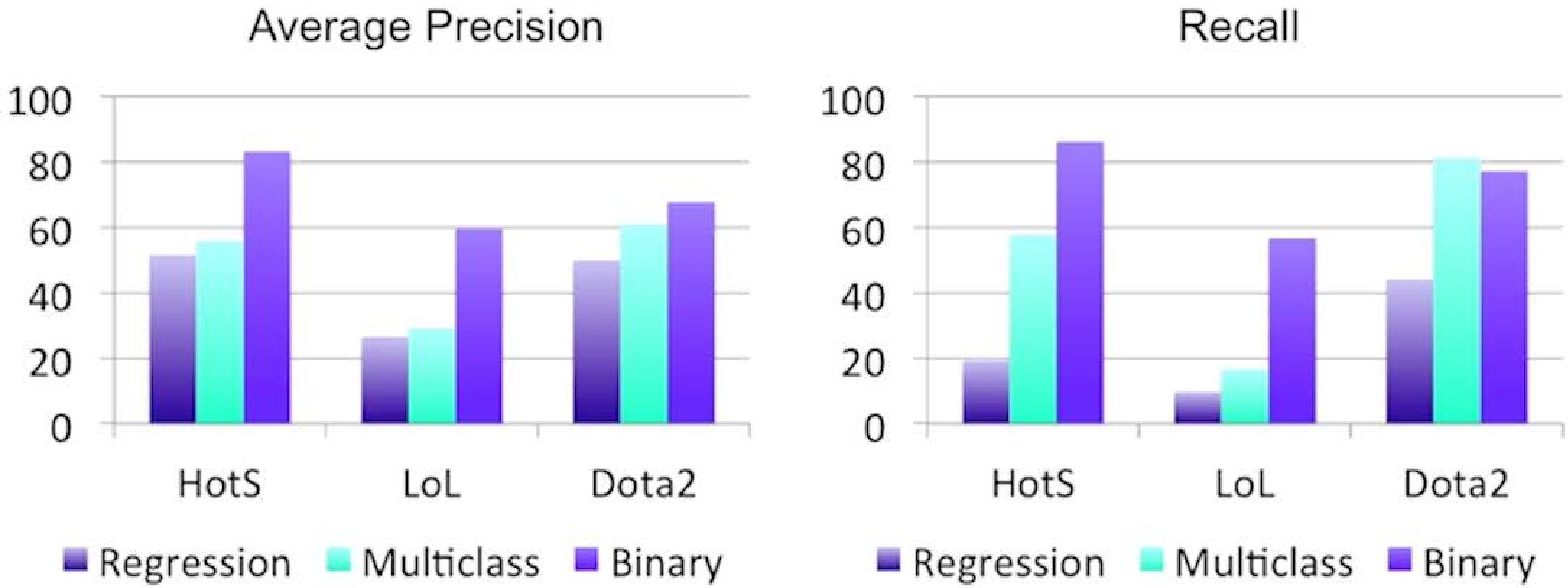
Yahoo Esports Dataset

- Three game titles: HotS, LoL, Dota2
- 300 hours of videos (pro league)
- Frame-level annotation
  - Scene types
  - Highlight scores

Cascaded Architecture is Important



Highlight Detection as Binary Classification



Played in 2x speed

Visualization was created using *Class Activation Mapping*, Zhou et al. CVPR 2016

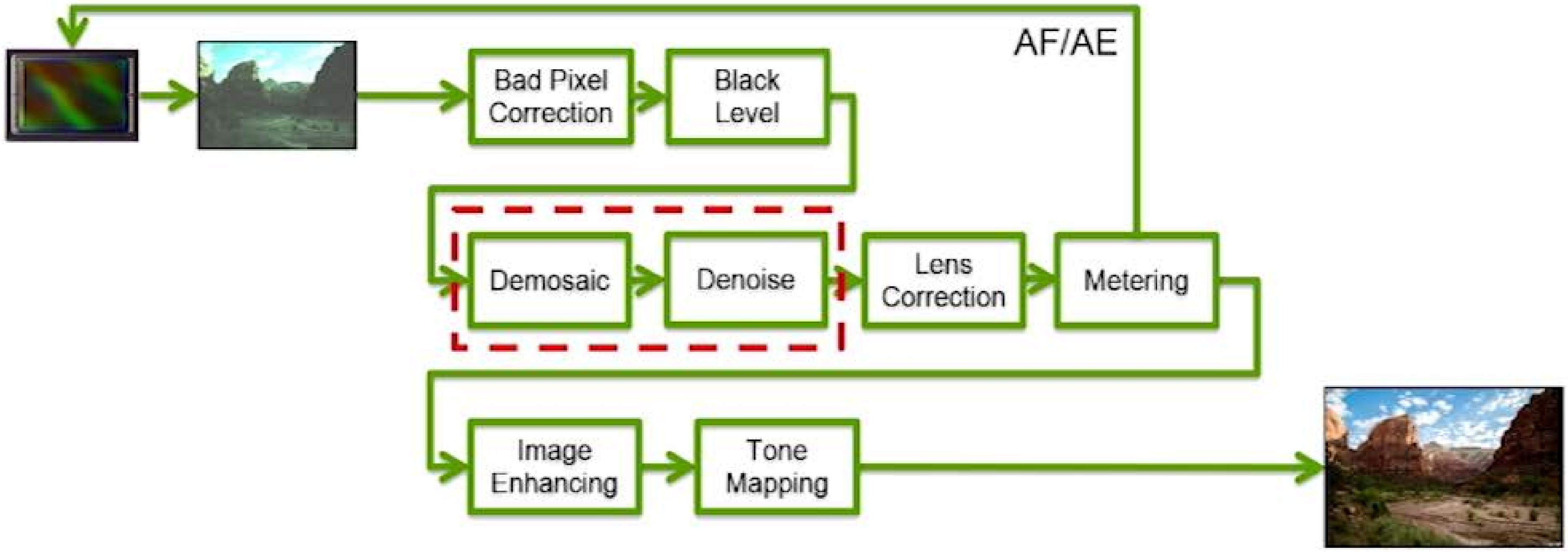
SESSION 6

IMAGE RESTORATION WITH NEURAL NETWORKS

Orazio Gallo, NVIDIA

MOTIVATION

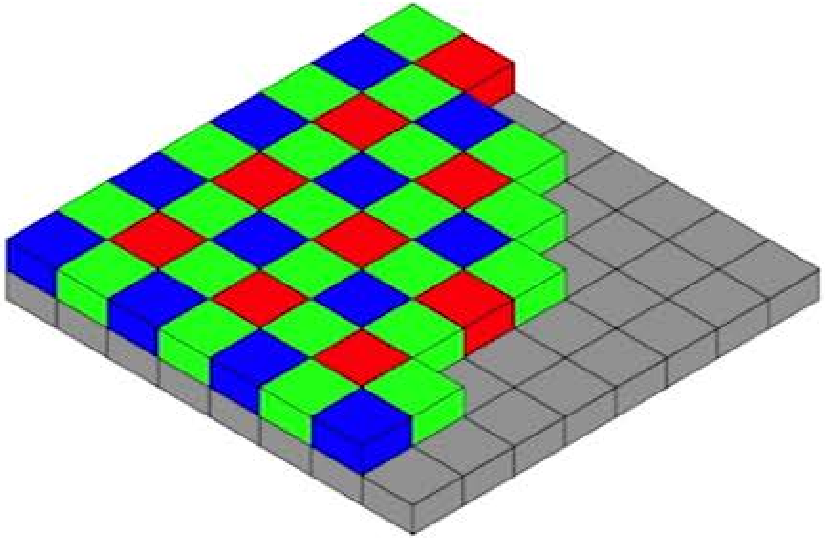
The long path of images…



DEMOSAICING

colors by interpolation

Image credit: Wikipedia



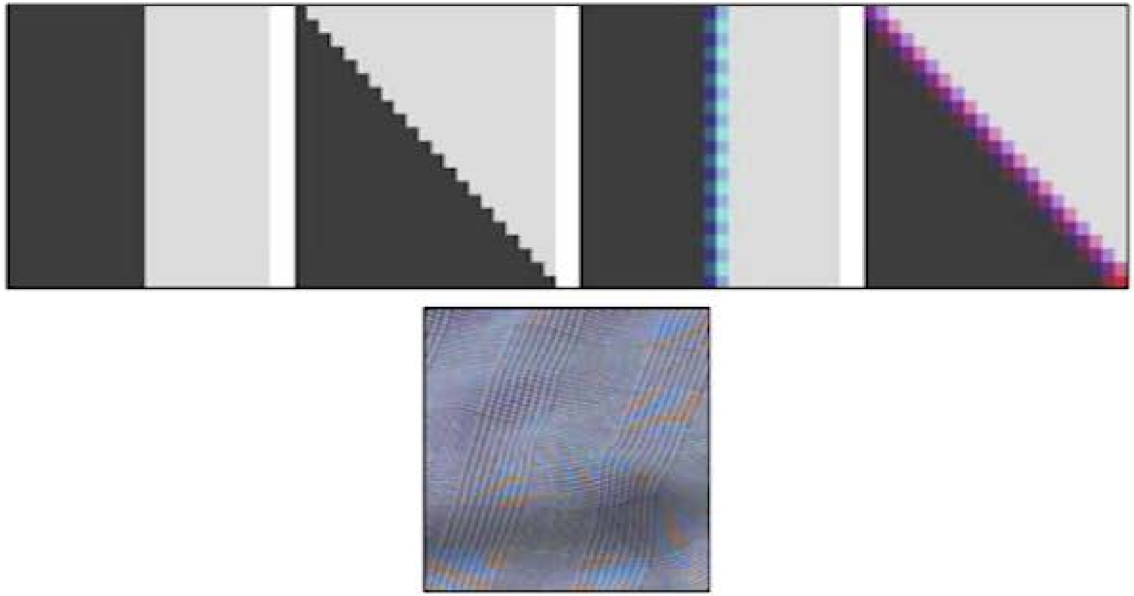


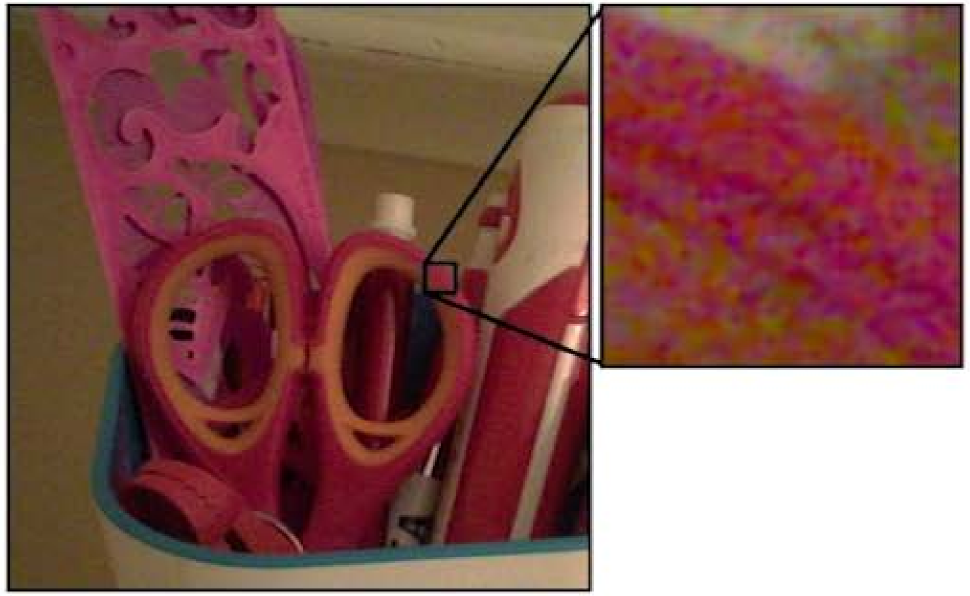
Image credit: Marc Levoy

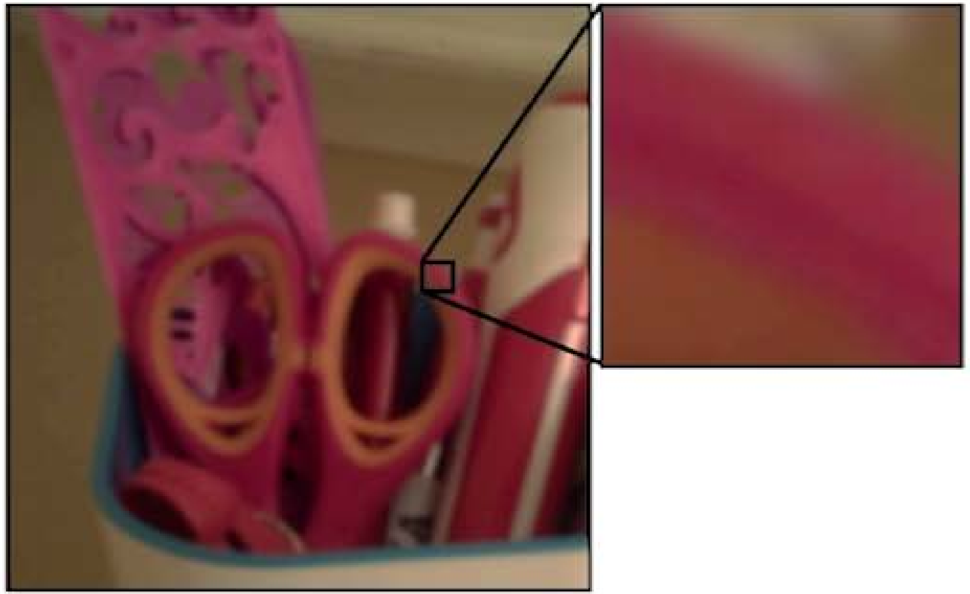
DENOISING

Several types of noise involved in the image formation:

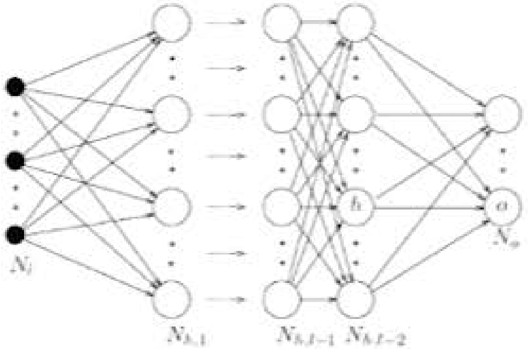
- Photon shot noise
- Dark current (AKA thermal noise)
- Photo-response non-uniformity
- Vignetting
- Readout noise:
  - Reset noise (charge-to-voltage transfer)
  - White noise (during voltage amplification amplification)
  - Quantization noise (ADC)

DENOISING





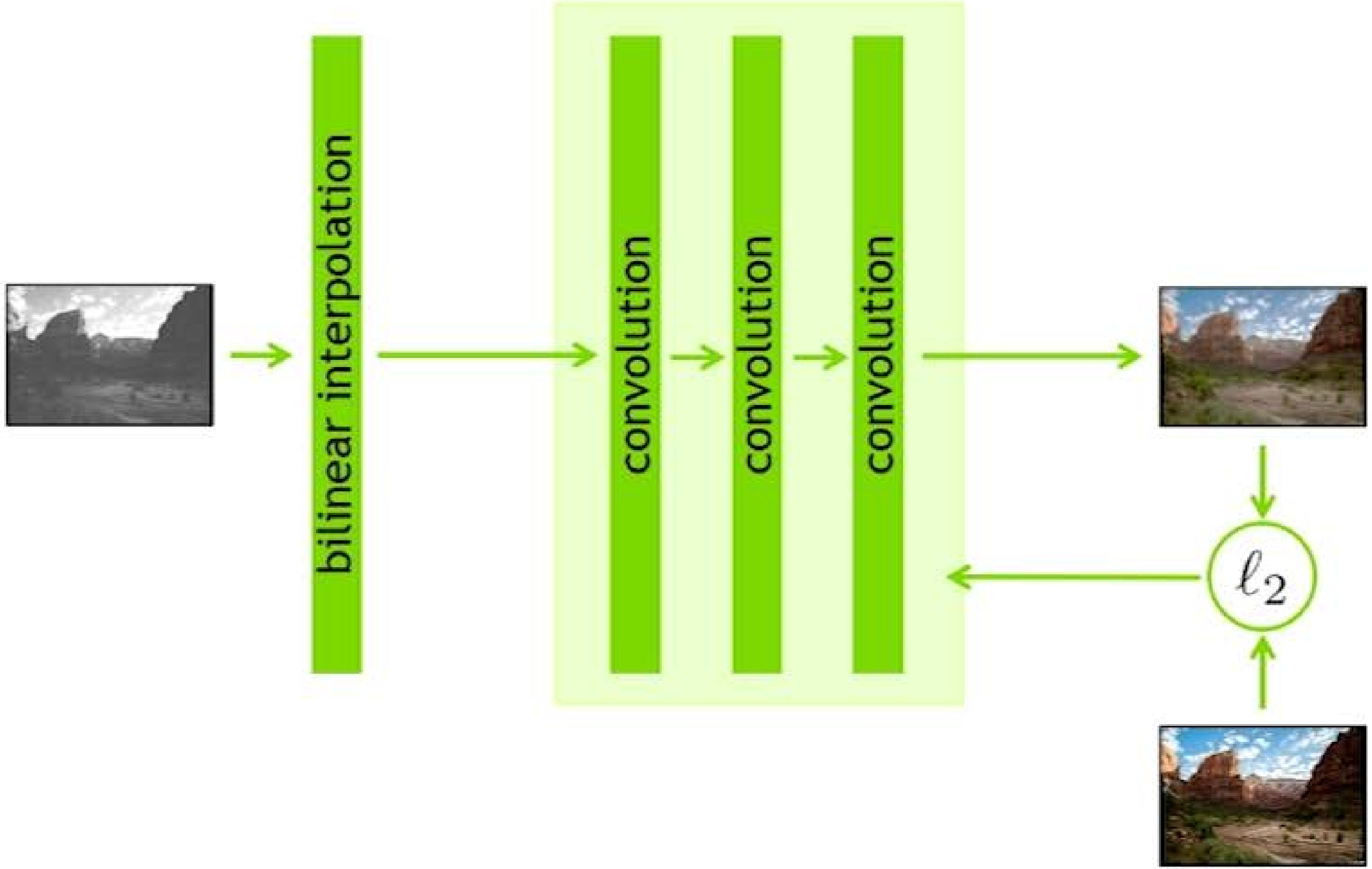
CAN WE DO IT WITH A NEURAL NETWORK?



NVIDIA

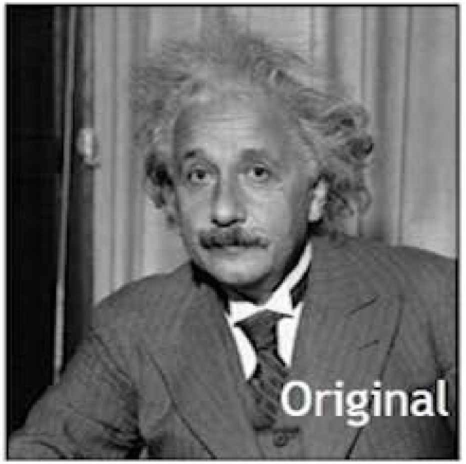
JOINT DEMOSAICING AND DENOISING

Network architecture

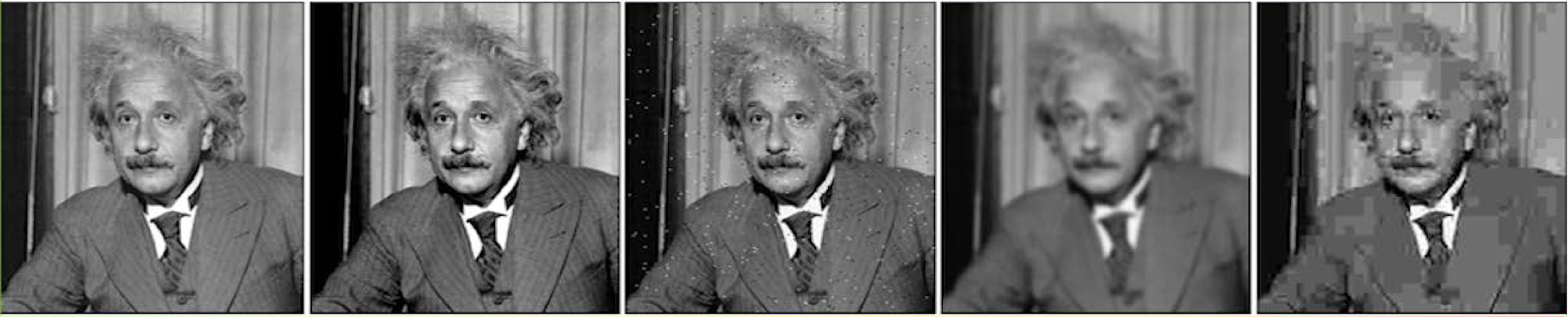


NVIDIA

MEASURING IMAGE QUALITY



ℓ_2



0.988

SSIM

0.662

Image adapted from https://ece.uwaterloo.ca/~z70wang/research/ssim/

MEASURING IMAGE QUALITY

Higher sensitivity to errors in texture-less regions!

$$\ell_1(p) = |I_1(p) - I_2(p)|$$

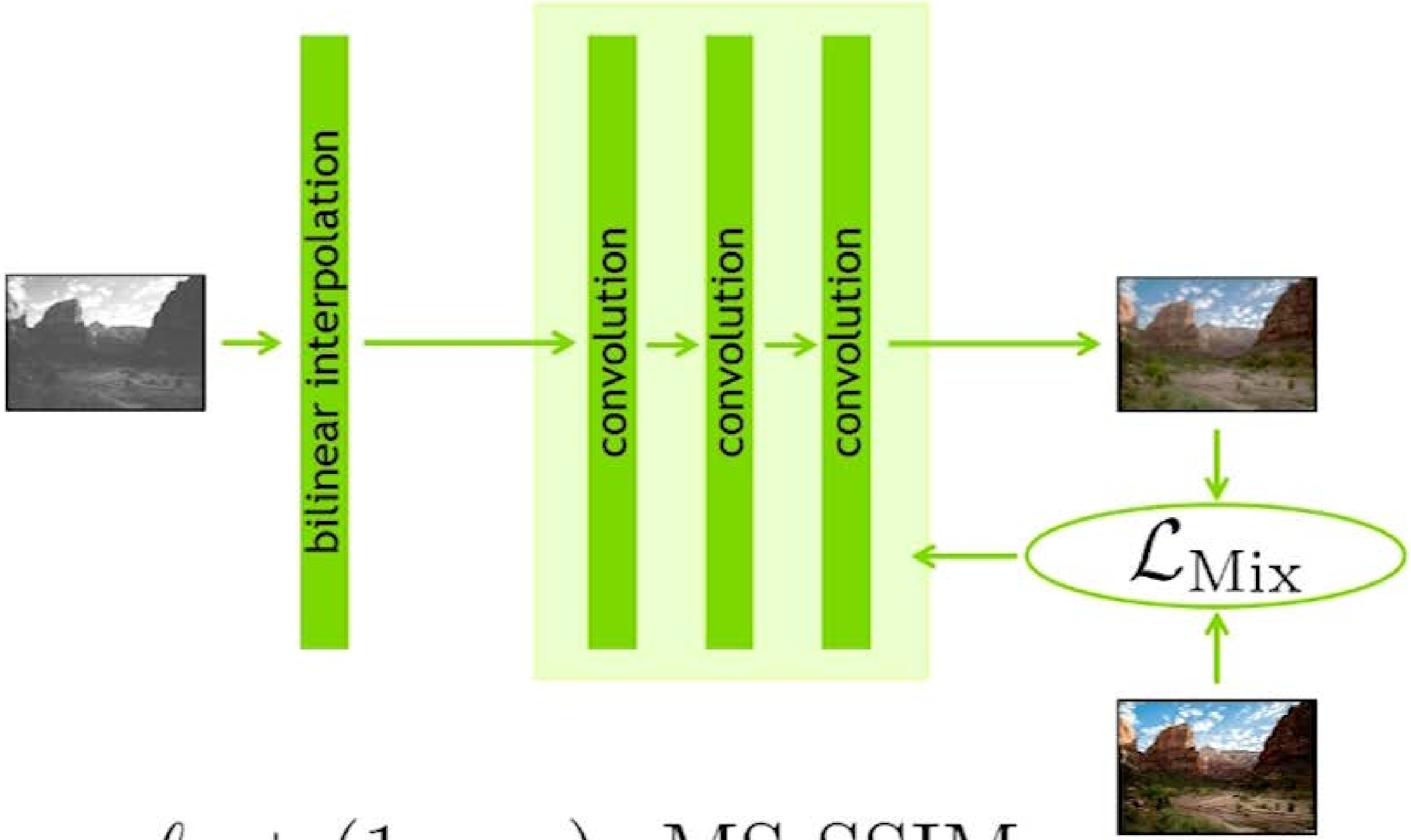
$$\ell_2(p) = \sqrt{I_1^2(p) - I_2^2(p)}$$

$$\text{SSIM}(I_1, I_2) = l(I_1, I_2) \cdot c(I_1, I_2) \cdot s(I_1, I_2)$$

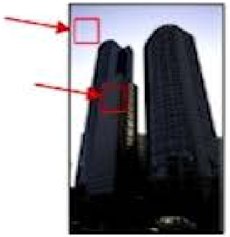
$$\text{MS-SSIM}(I_1, I_2) = \text{Multiscale}(\text{SSIM}(I_1, I_2))$$

JOINT DEMOSAICING AND DENOISING

Network architecture

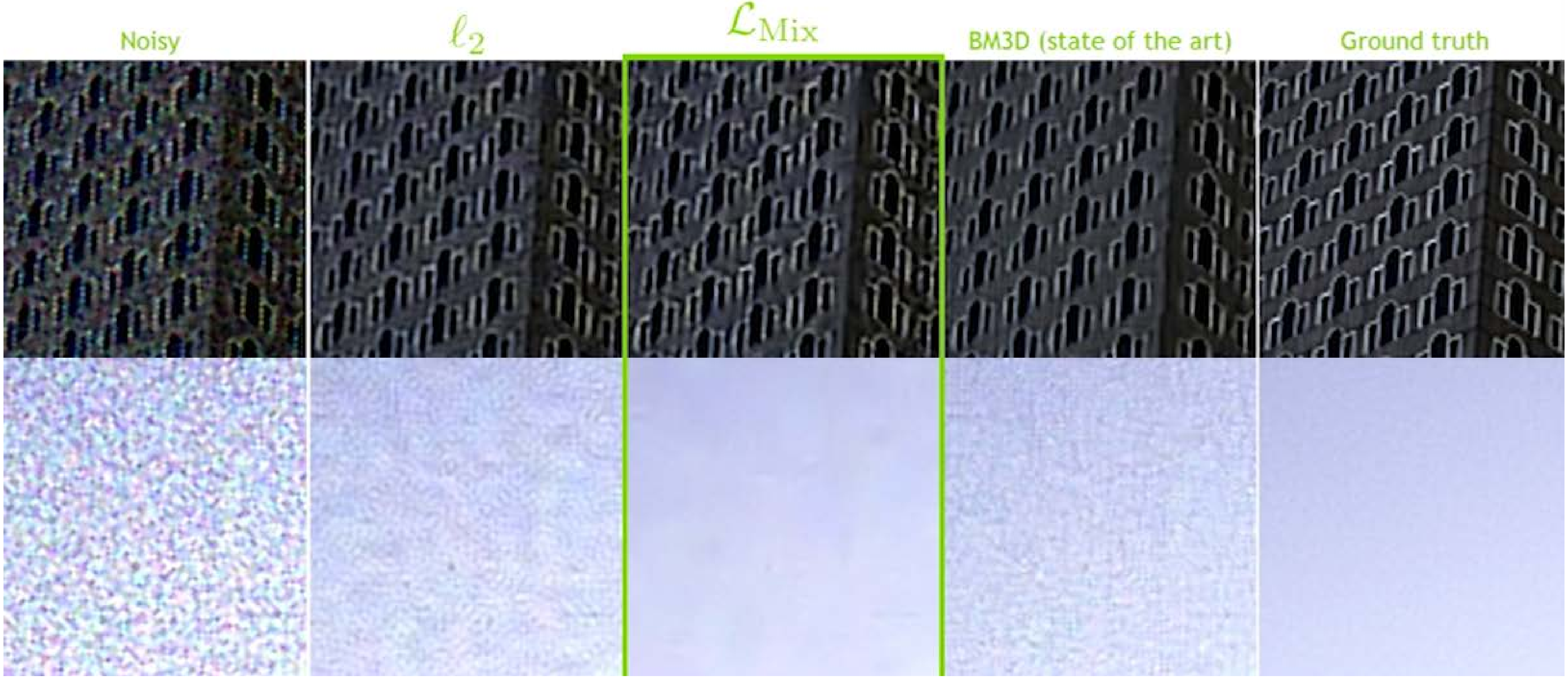


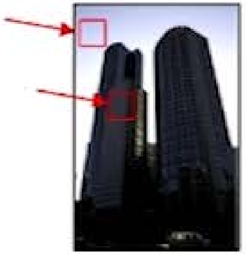
$$\mathcal{L}_{\text{Mix}} = \alpha \cdot \ell_1 + (1 - \alpha) \cdot \text{MS-SSIM}$$



RESULTS

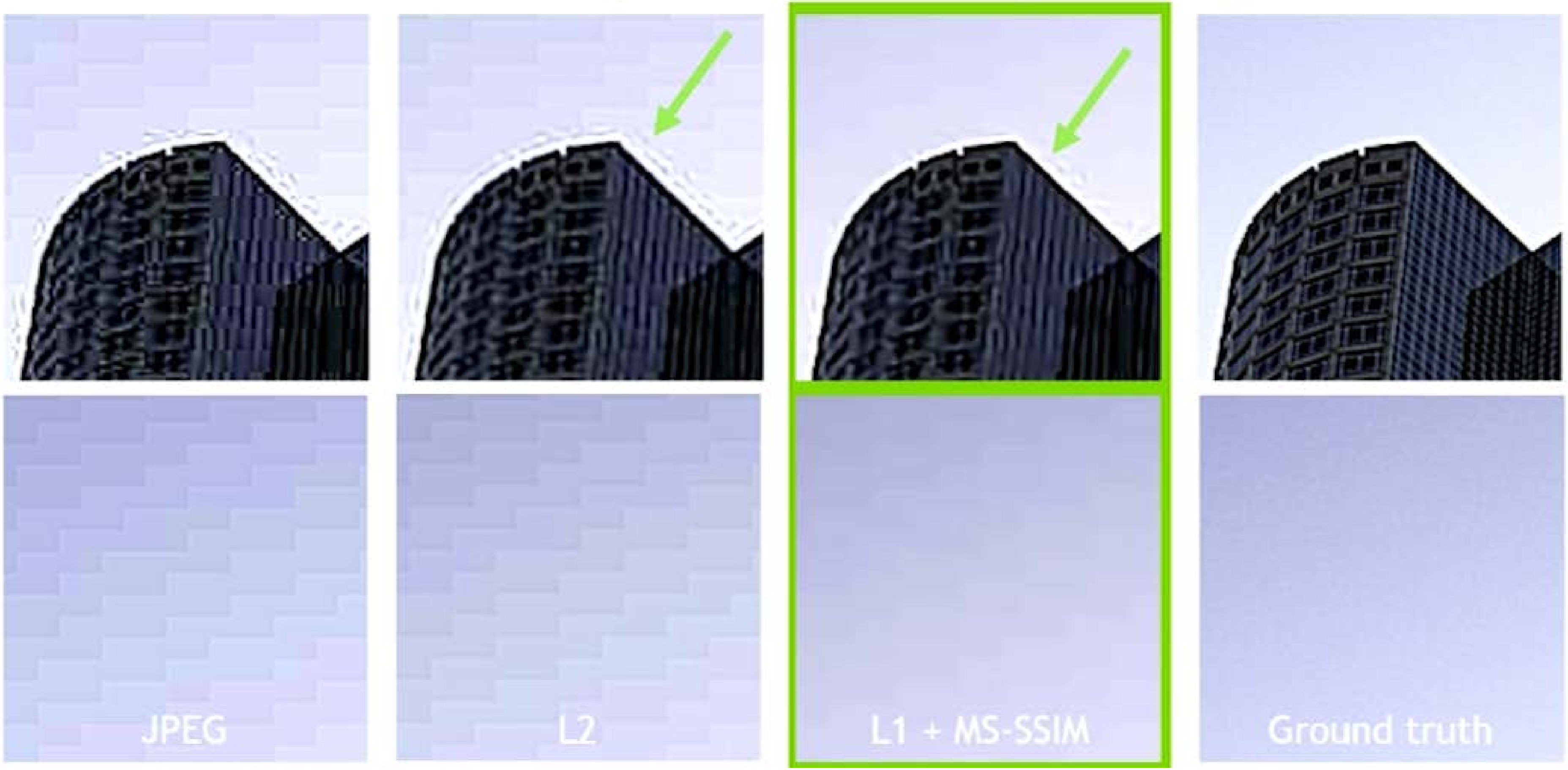
Visual comparison (+ unsharp masking)





JPEG ARTIFACT REMOVAL: RESULTS

Visual comparison (+ unsharp masking)



SUPER-RESOLUTION: RESULTS

Visual comparison (+ unsharp masking)

SESSION 7

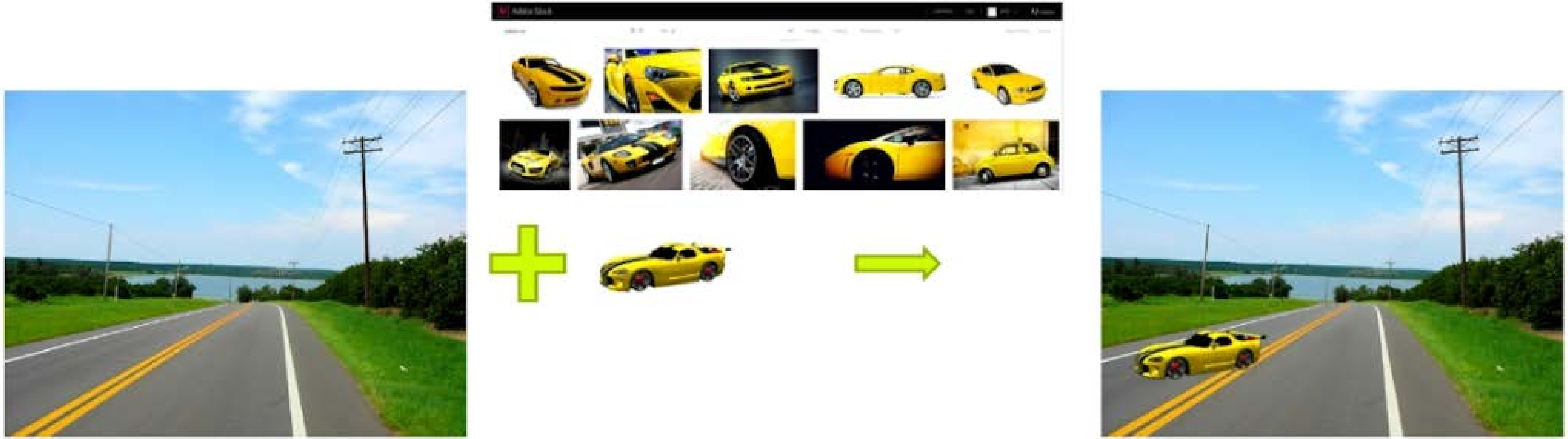
NOVEL 3D VIEW SYNTHESIS FROM A SINGLE IMAGE

Jimei Yang - Research Scientist, Adobe

Synthesizing Object Images from Novel Viewpoints

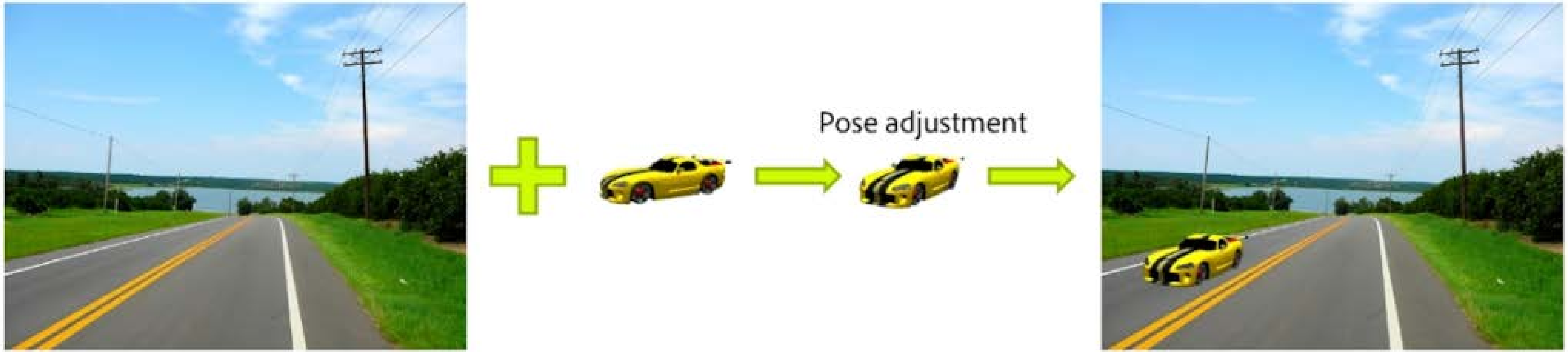
Input Image | Synthesized Views | Input Image | Synthesized Views | Input Image | Synthesized Views

Image Composition

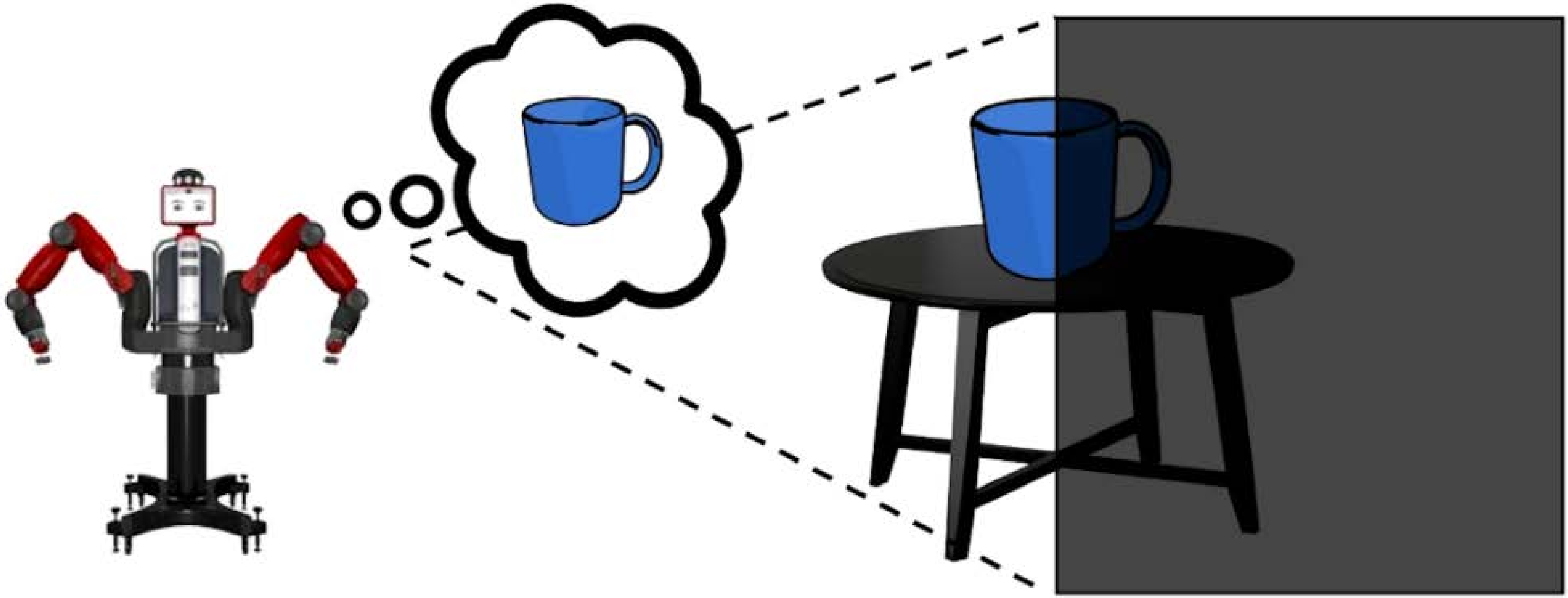


DEEP LEARNING INSTITUTE

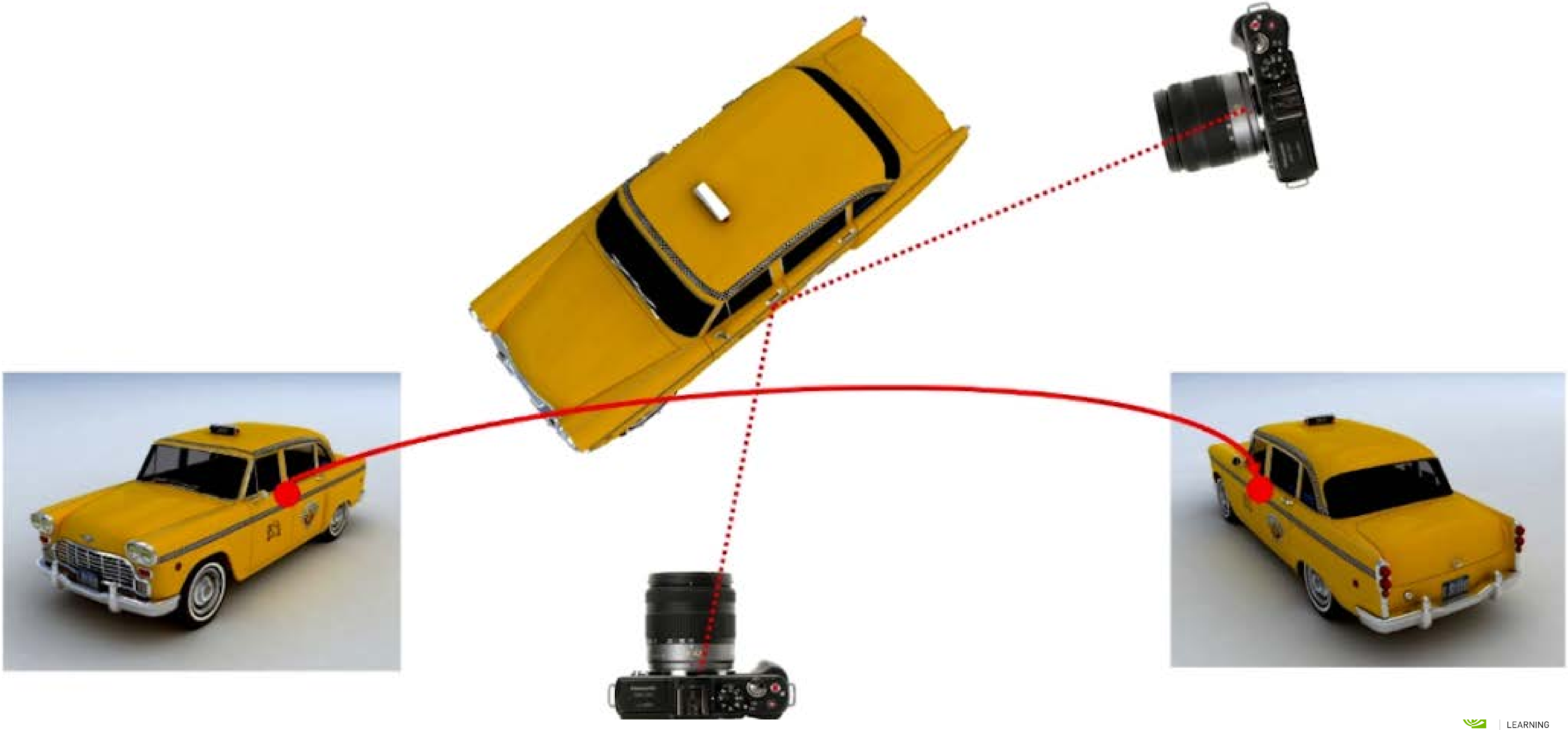
Image Composition



Robot Grasp Planning

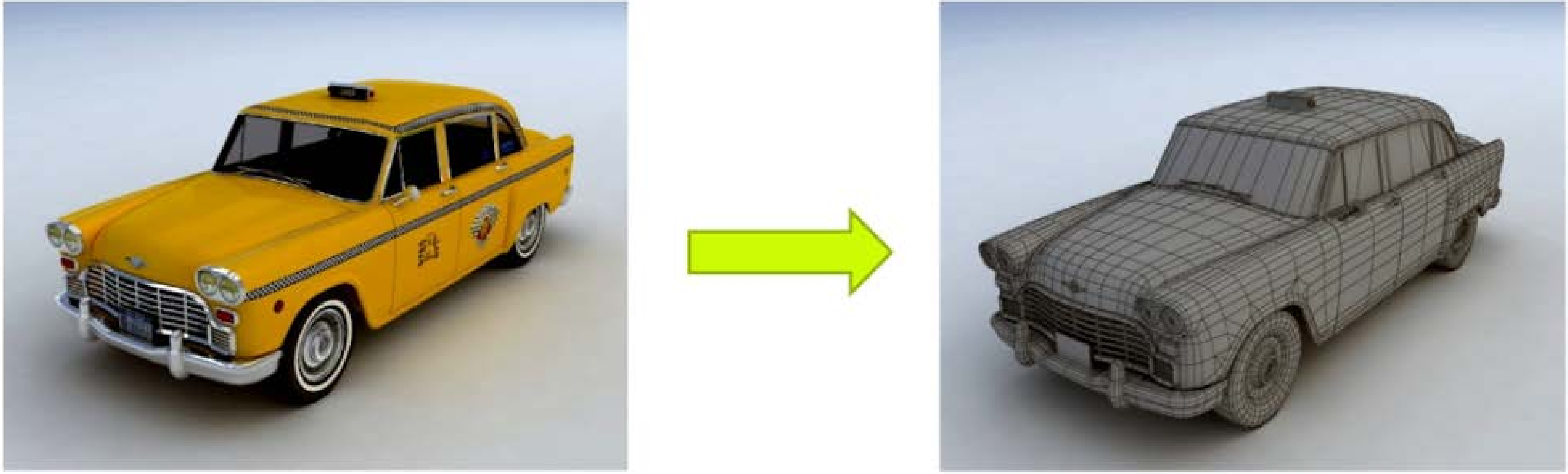


View Synthesis
as Simulating a New Camera Looking at the 3D Object

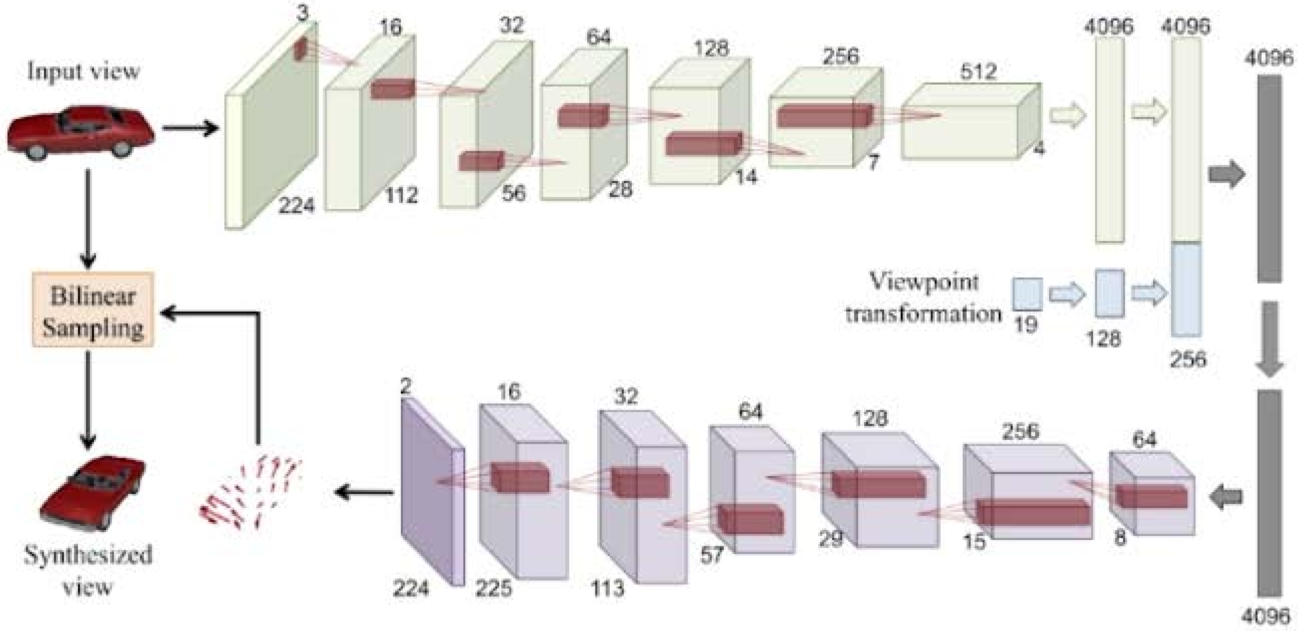


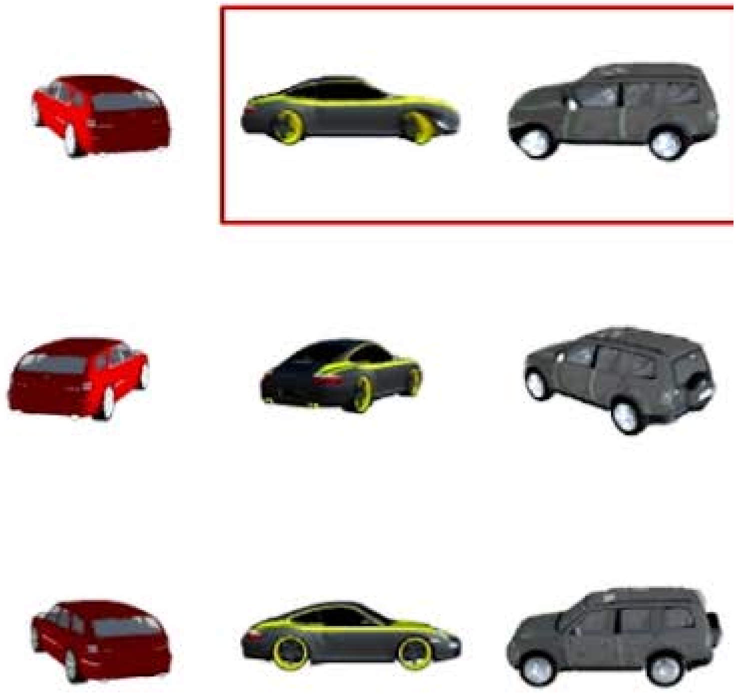
NVIDIA. LEARNING INSTITUTE

First Challenge: Recovering the 3D Structure



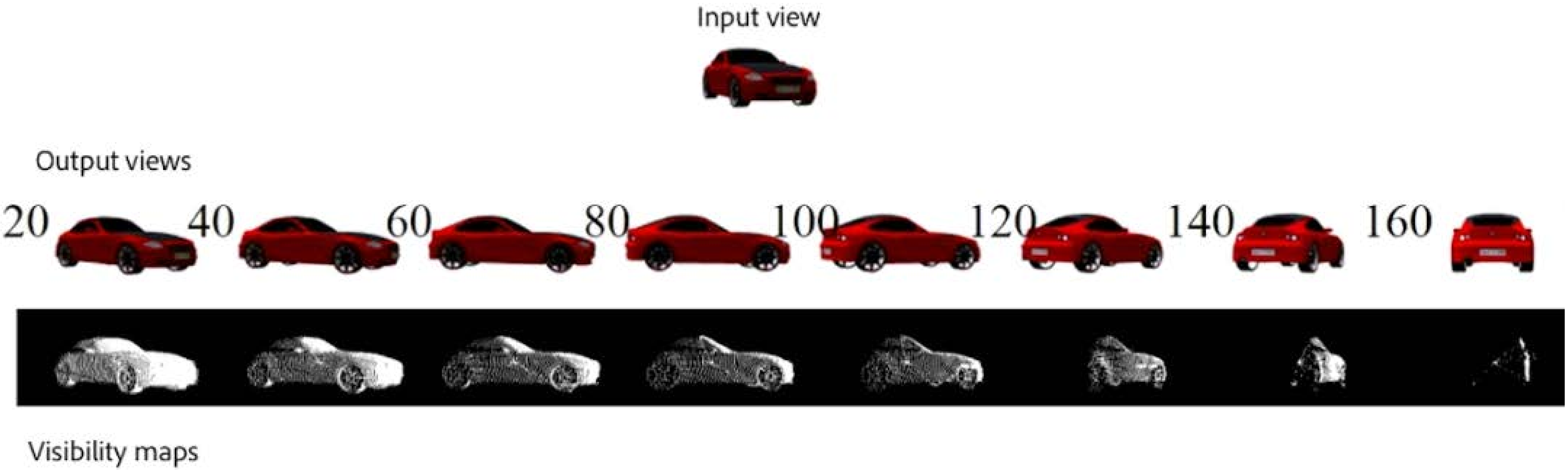
Learning the Relation Between Any Two Views



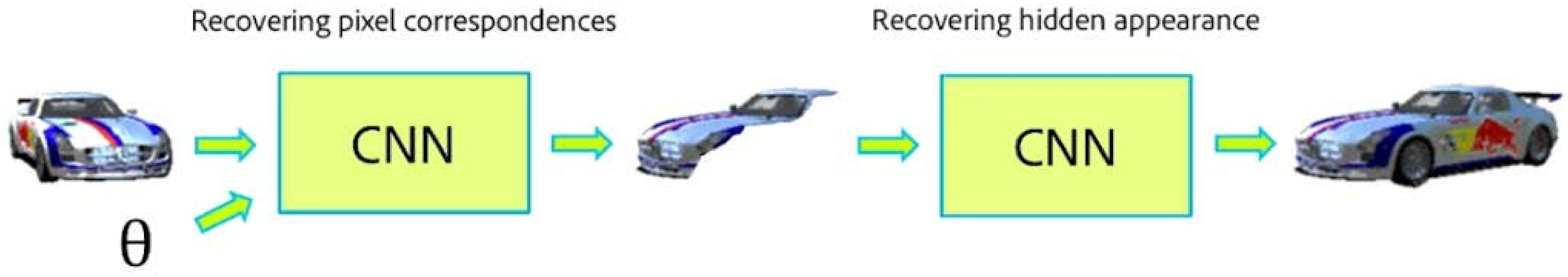


T. Zhou, et al. ECCV 2016

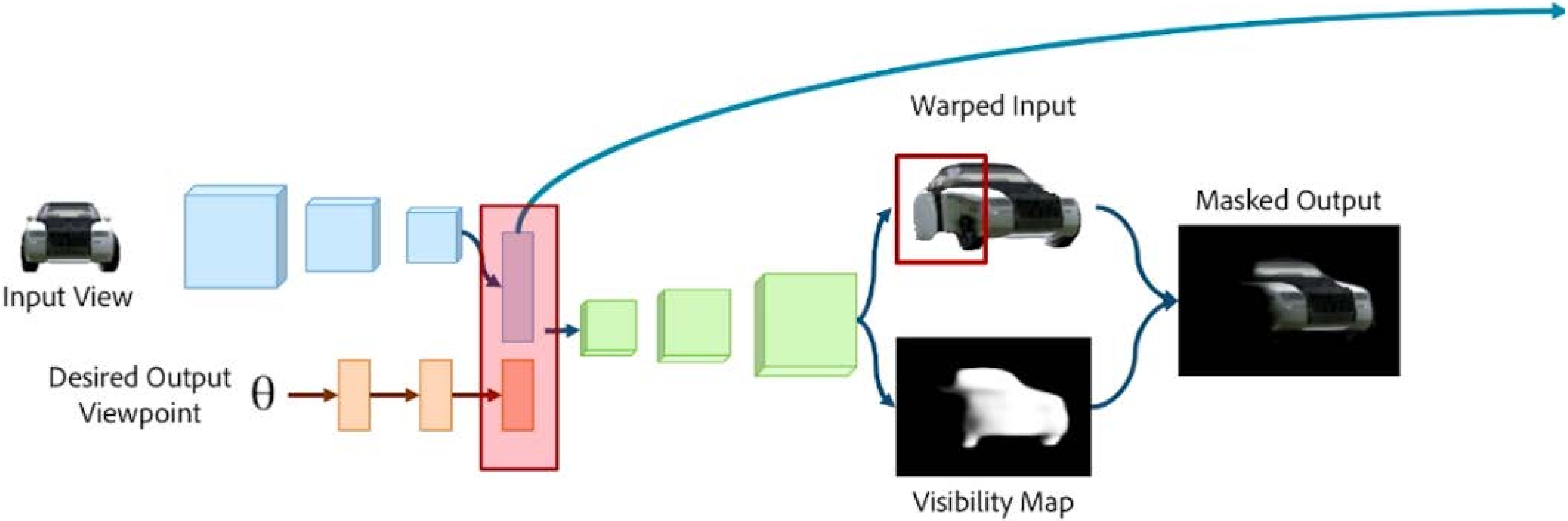
Second Challenge: Recovering Hidden Appearance



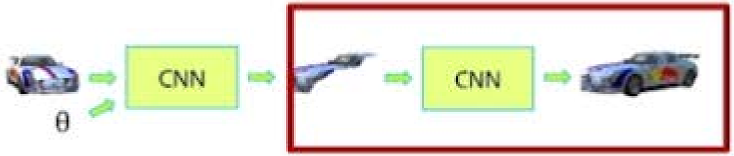
An End-to-End Deep Learning Approach

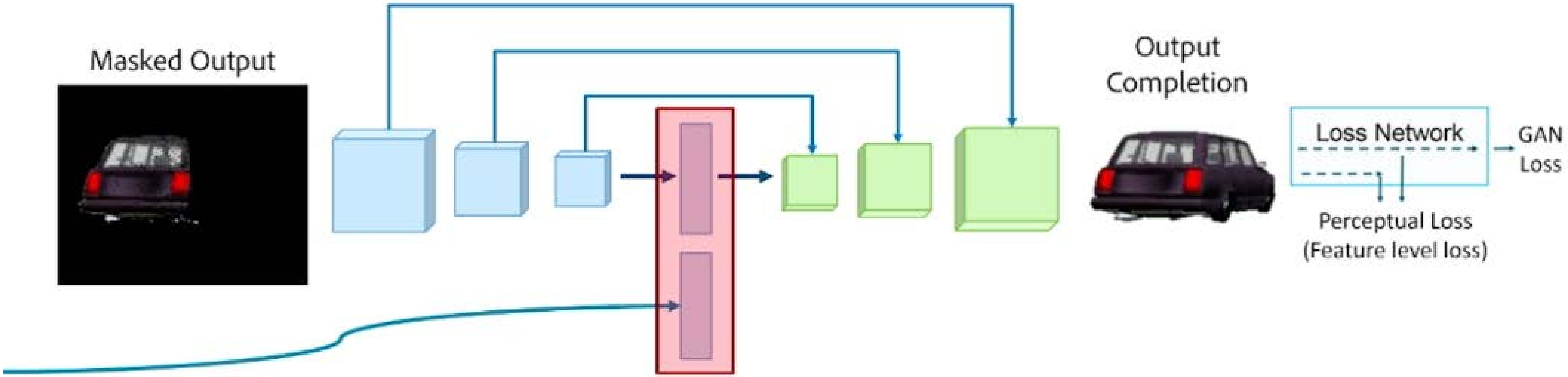


Disocclusion-aware Appearance Flow Network



Completion Network using GANs

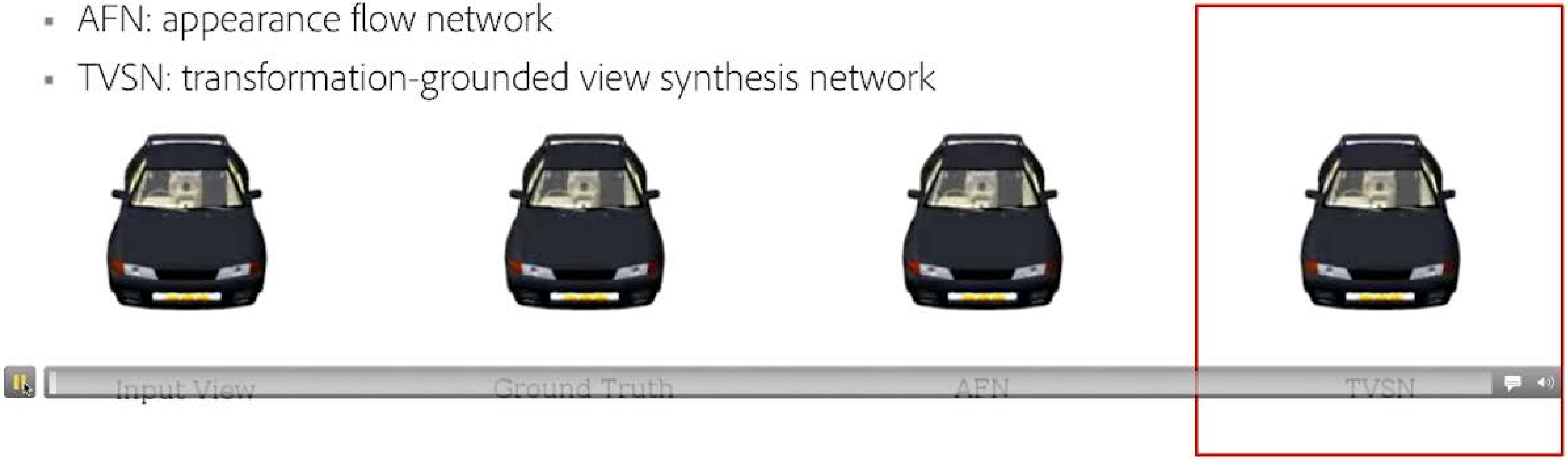




High-level object identity features from DOAFN

View Synthesis Results

- AFN: appearance flow network
- TVSN: transformation-grounded view synthesis network



SESSION 8

IMPROVING CONSUMER COMPLIANCE THROUGH BETTER PRODUCT RECOMMENDATION- NEW SKIN ADVISOR TOOL POWERED BY AI

Matthew L. Barker, Ph.D. - Principal Data Scientist, Procter & Gamble

Development Overview

Deep Learning Algorithm
Visible skin age prediction with aging area identification.

→

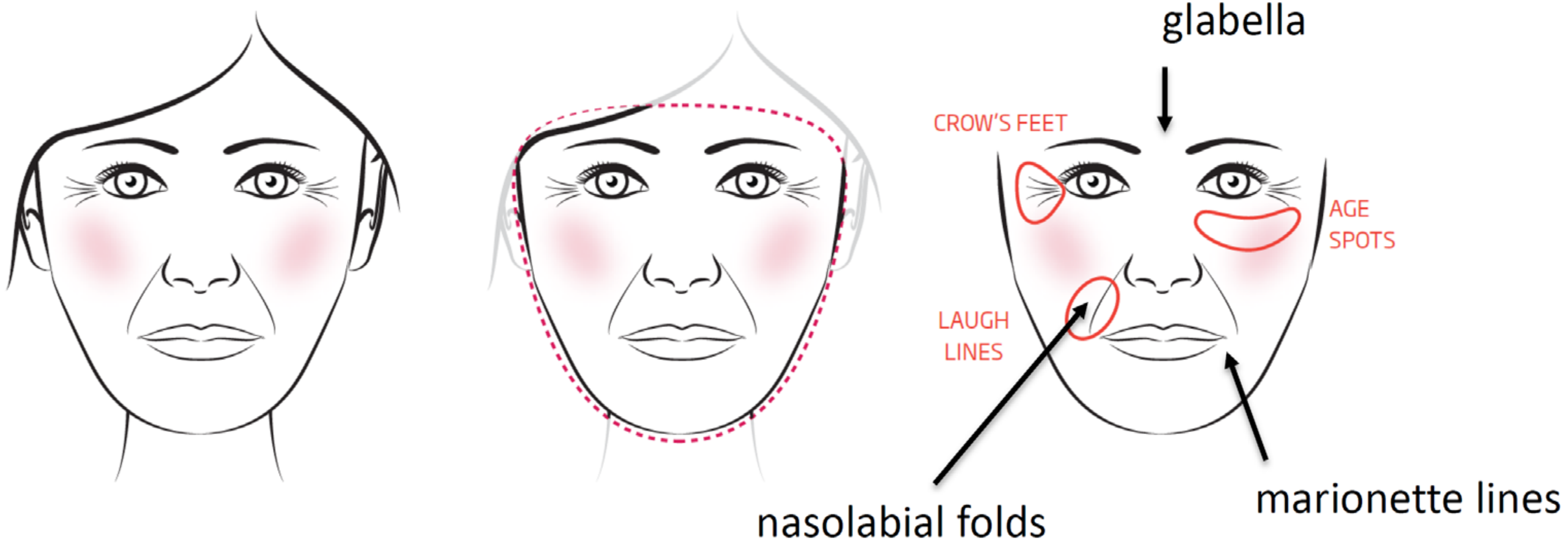
Visible Skin Age Validation
Predictions compared to expert.

Aging Area Insights
Facial Mapping Study informs how appearance of aging areas change with chronological age.

→

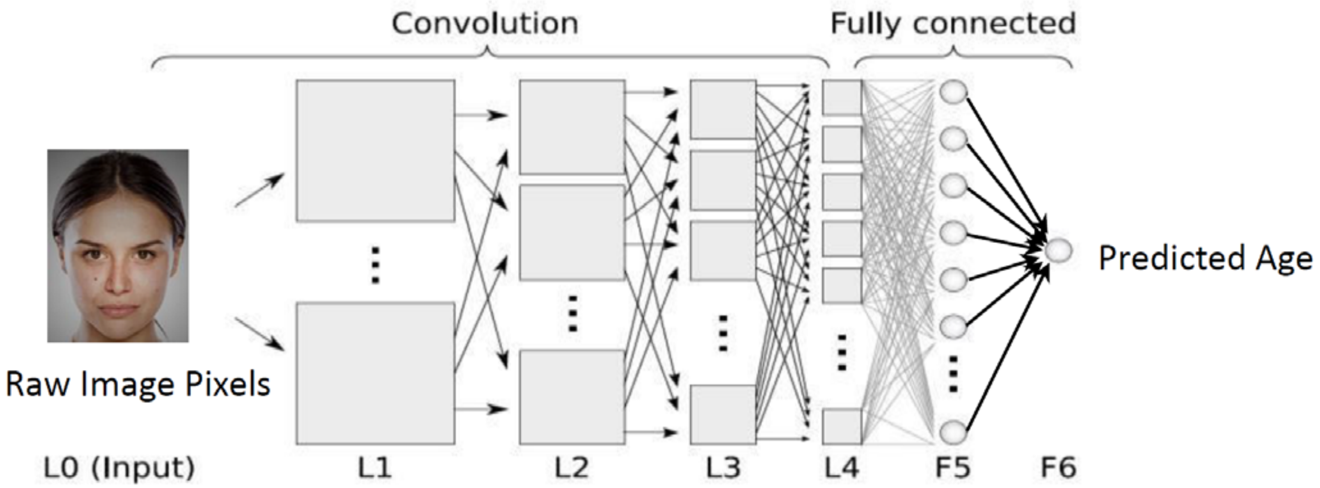
Compliance Verification
Proving skin advisor with deep learning algorithm, visible aging insights and consumer preferences drives compliance.

Facial Features & Aging



Deep Neural Network application

- The skin advisor uses convolutional neural networks trained using NVIDIA graphics processors to perform trillions of calculations per second. The model was trained on 50,000 images with chronological age data tags.
- When an image of a user is received, the model is used to determine the visible skin age based on the pixels in the image, further a two-dimensional heat map is generated that identifies a region of the image that contributes to the visible skin age.



Data Setup

- Face detection & alignment performed using dlib: rotated, scaled & cropped to a standard size.
- Spatial augmentation was applied: random horizontal flipping, rotation, scaling, zoom cropping causing slight translation.
- HSV Color augmentation: random changes to saturation & exposure.
- Oval Mask, global contrast normalization GCN, reapply Oval Mask.

Gradient Heat Map for Visualization

- After training, with fixed model parameters. A gradient heat map was created in order to localize pixel differences of a subject's image relative to younger than their predicted age.
- An input image was forward propagated through the model to obtain a predicted age. Then a target of predicted age minus 10 years was set and the gradients were propagated back through the network to the input image. A heat map was created by summing absolute values of the RGB gradients for each pixel and rescaling from 0 to 1 for display purposes.
- The gradient heat map was then blended with the original image to visualize areas that were different from their younger predicted age.

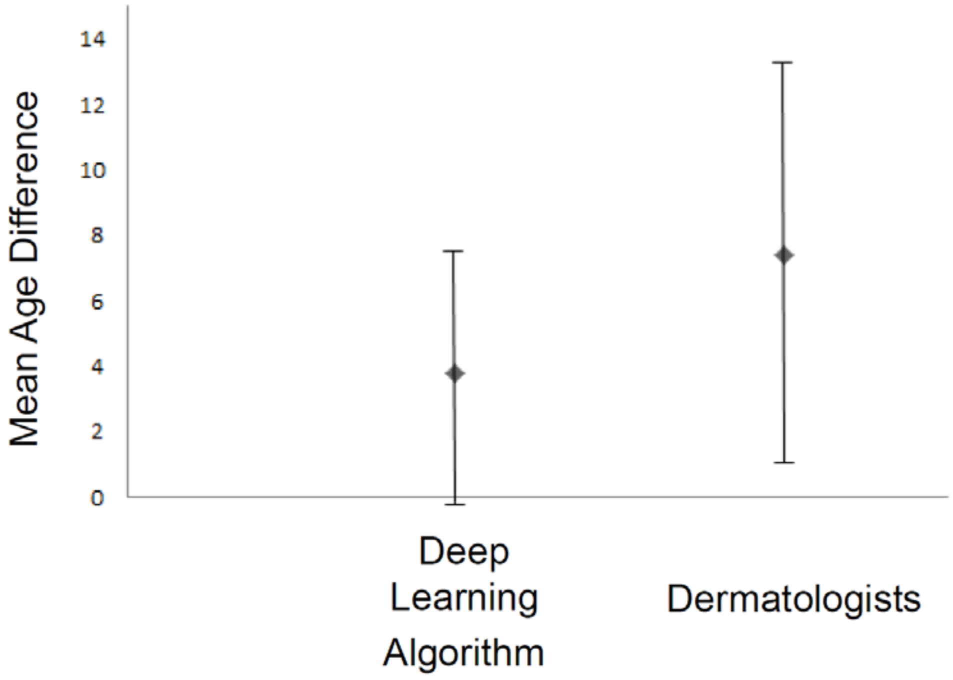
Visible Skin Age Validation

Evaluate robustness of the visible skin age algorithm by comparing output to a gold standard dermatologist assessment.

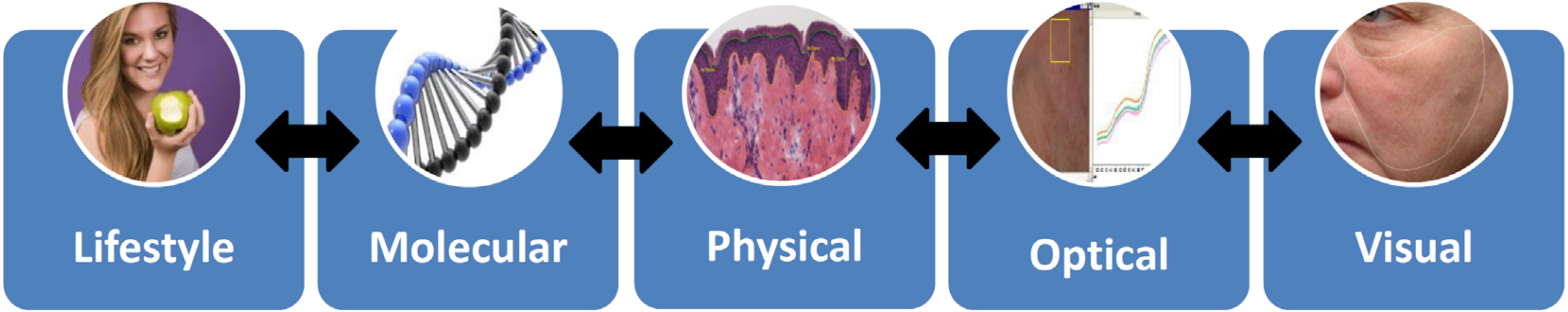
1. A validation set of 630 selfie images representing the general US female population were obtained.
2. These images were presented to 615 dermatologists, who represent the gold standard in visible skin evaluation, in a randomized order in sets of 8 images. Each dermatologist evaluated images.
3. The dermatologists were asked to input the perceived age of each image.

Validation Results

The mean difference of the predicted visible skin age versus the chronological age using the skin advisor deep learning algorithm was comparable to the mean difference of the perceived age versus the chronological age by dermatologists.



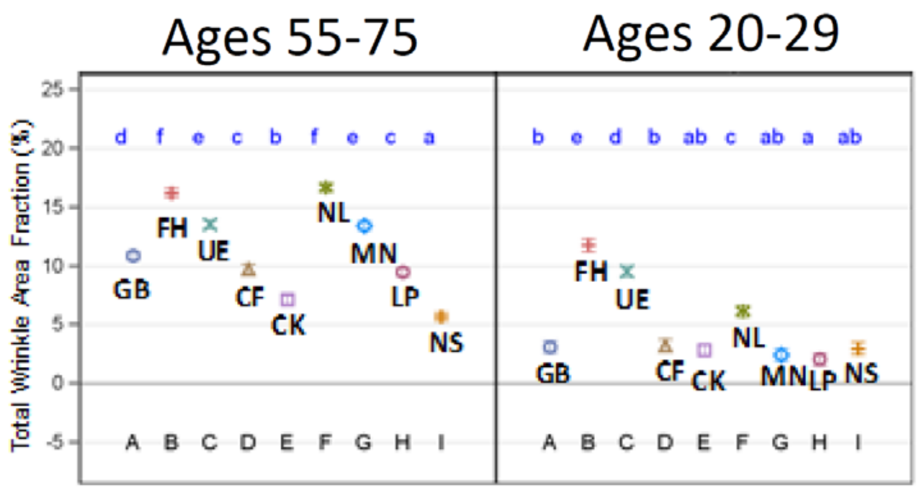
Facial Area Insights – Mapping Study



- To build a fundamental understanding of the underlying mechanisms of facial aging across different facial sites, a clinical Facial Mapping Study enrolling over 150 subjects
- Study assessed facial skin genomics, image analysis parameters, lifestyle factors, and skin measurements in two groups of female subjects: a younger ages (20-29 years) and an older ages (55-75 years). Study did not assess applying cosmetics.
- Facial locations analyzed included the forehead, crow's feet area, under eye, nasolabial fold, cheek, glabella, marionette lines, above mouth, and nose regions.

Facial Mapping Study - Results

- The Skin Advisor Tool shares the best aging area and the area that needs improvement based on the deep learning algorithm. Key educational information about how those areas age is also given.
- Insights from the facial mapping study were used to inform how visible aging areas change with chronological age.
- Quantitative assessment of wrinkles revealed distinct visible topography feature presentation across facial zones and with aging.



GB = Glabella
FH = Forehead
UE = Under Eye
CF = Crow's Feet
CK = Cheek
NL = Nasolabial Fold
MN = Marionette
LP = Above Lip
NS = Nose

Figure 2. Total Wrinkle Area Fraction (%) – Facial Site Comparison by Age Group. Analysis of Variance (ANOVA) was used for each age group (Younger and Older). Same grouping letters indicates no significant difference at 0.10 (2-sided).

Compliance Verification

- 100 US women, age 25-65, facial moisturizer users, were enrolled in a 4-week online consumer test.
- Group 1 (n=50) received a product regimen based on the skin advisor deep learning algorithm and preferences and Group 2 (n=50) self-selected a product regimen.
- Self-assessment questions were completed pre-use and post-4 weeks product use.

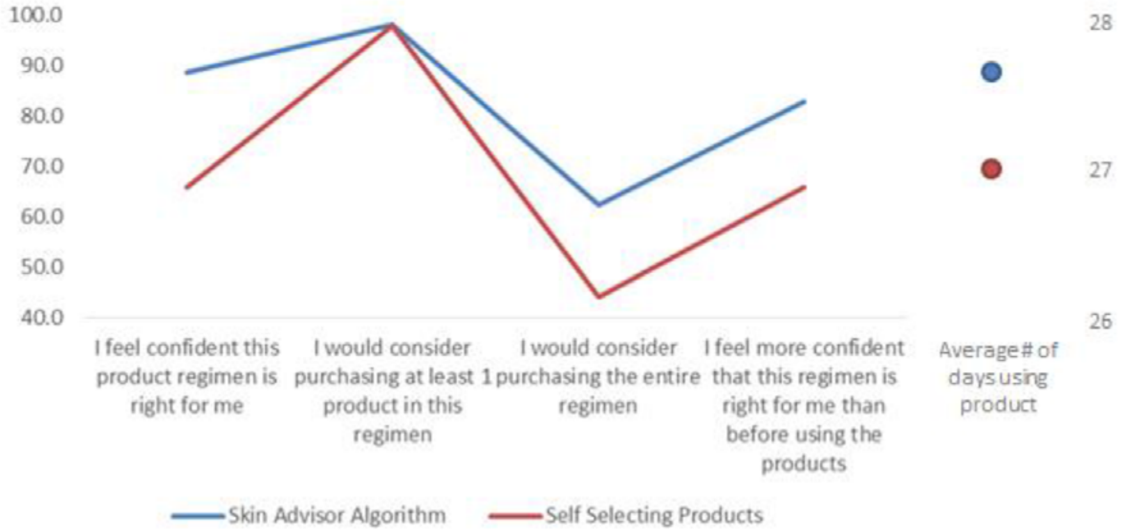
Compliance Results

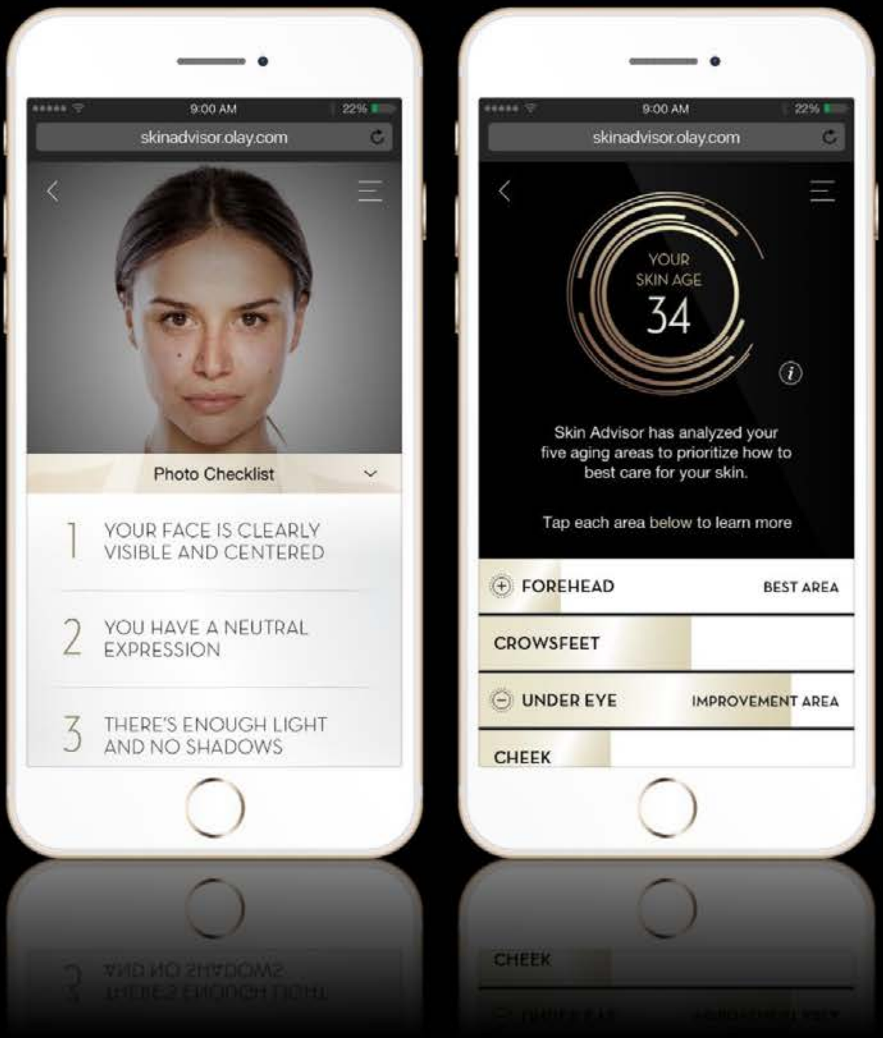
Figure 4

Pre-product use indicates satisfaction with the skin advisor product recommendation.

Figure 5

Post 4 weeks product use indicates satisfaction with the skin advisor product recommendation and improved consumer compliance with longer product use.





OLAY® SKIN ADVISOR

THE SCIENCE BEHIND OLAY SKIN ADVISOR

skinadvisor.olay.com

SESSION 9

FACIAL EXPRESSION AND EMOTION DETECTION FOR MOBILE

Jay Turcot - Director of Applied AI, Affectiva


What if technology could **identify emotions as humans can**?
:) Affectiva
@affectiva

Task: Facial expression recognition

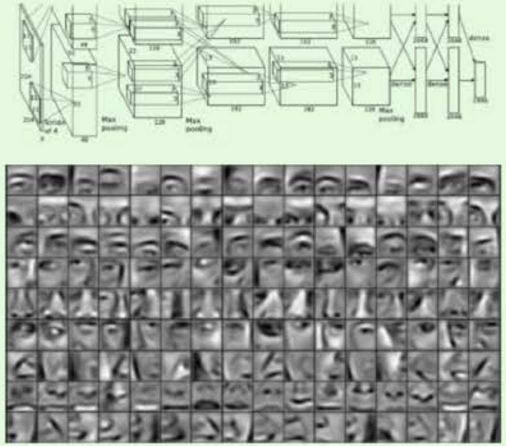
Brow raise

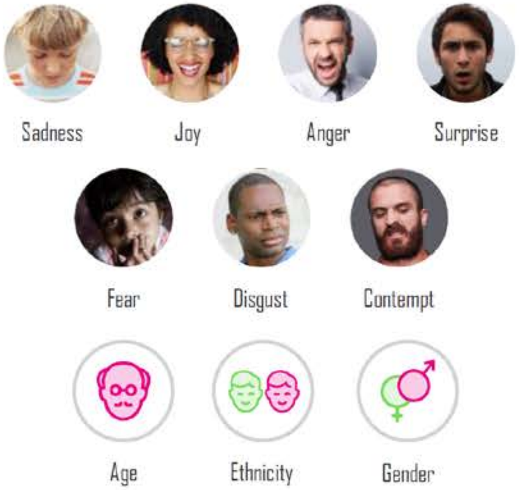
Brow furrow

Smile

- Multi-attribute classification (~20+ classes)
  - Upright, fixed-size, grayscale
- Fast enough to run on-device!

Emotion AI platform built on deep learning





Input:
Labeled and unlabeled videos (+voice) data. Meta data. Latest training used 1M+ images.

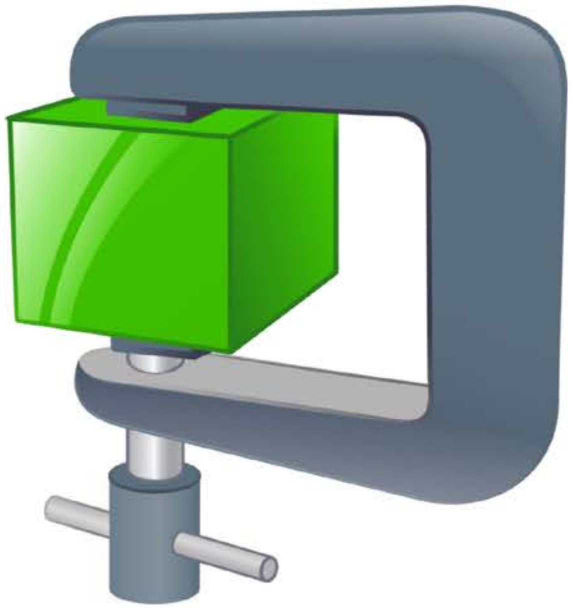
Convolutional Neural Networks

Output:
11 Facial expressions
Gender

Speeding up deep learning models

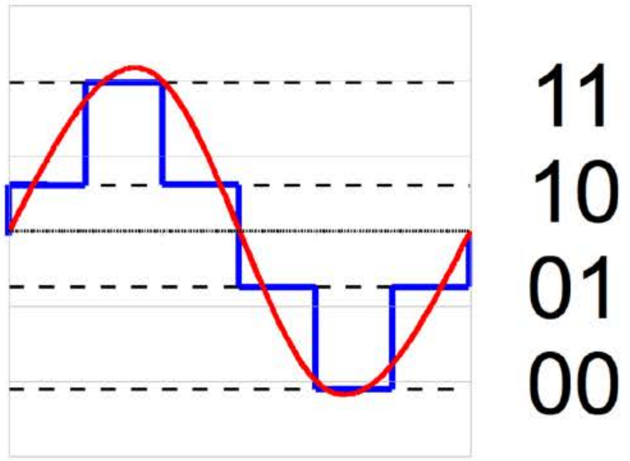
Several approaches are used for speeding up models

Model Compression



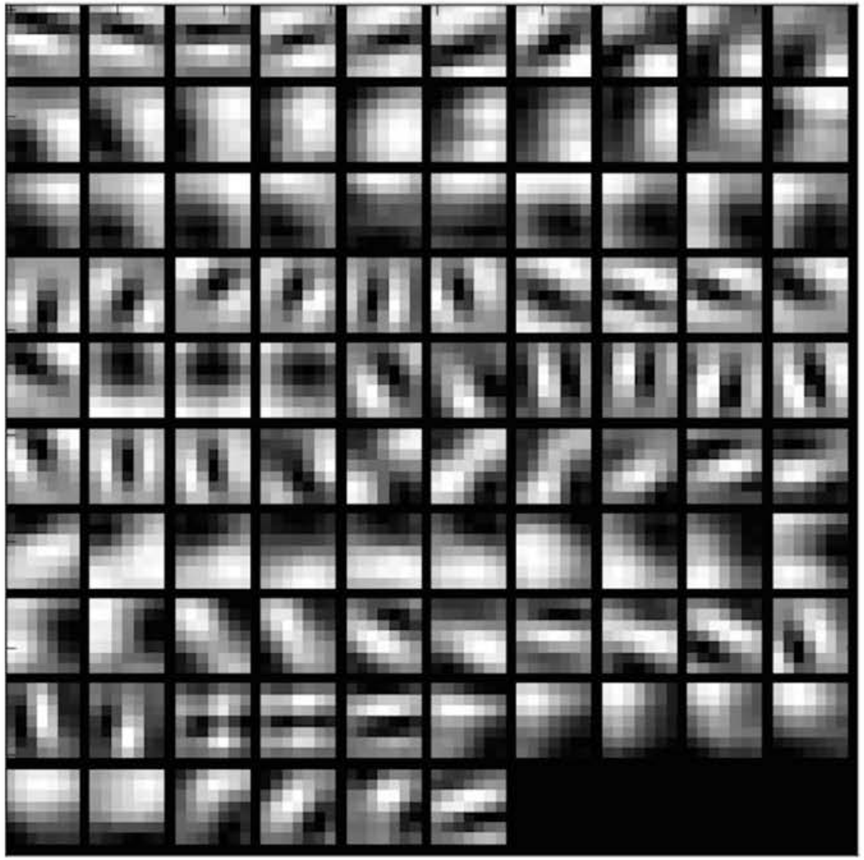
Model Pruning

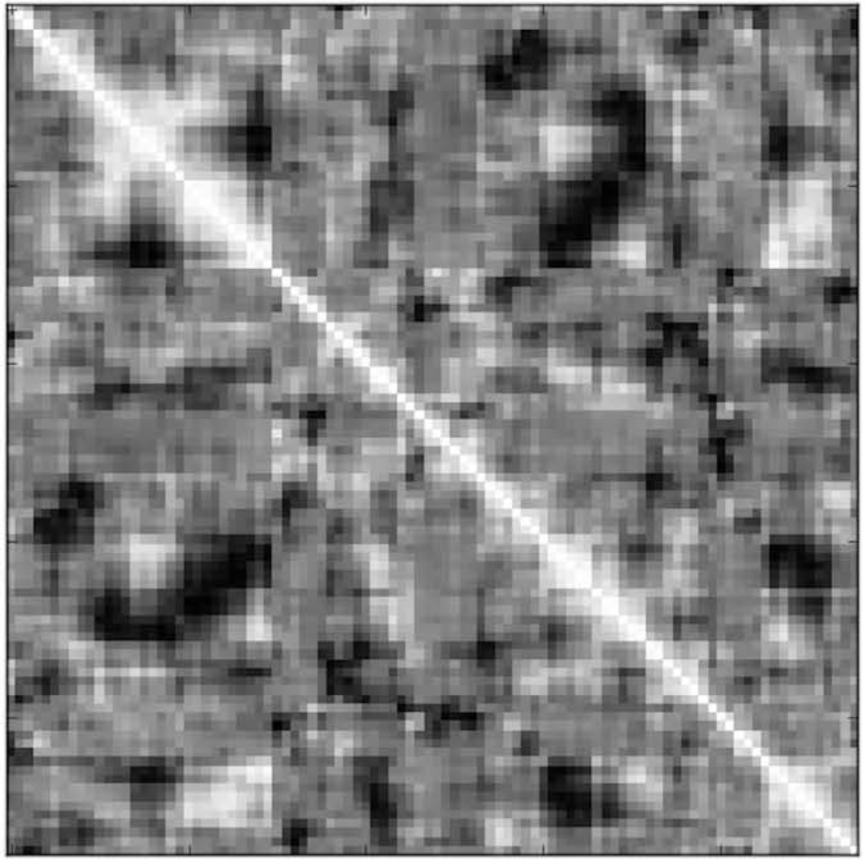
Model Quantization



Lots of big filters are **expensive**!

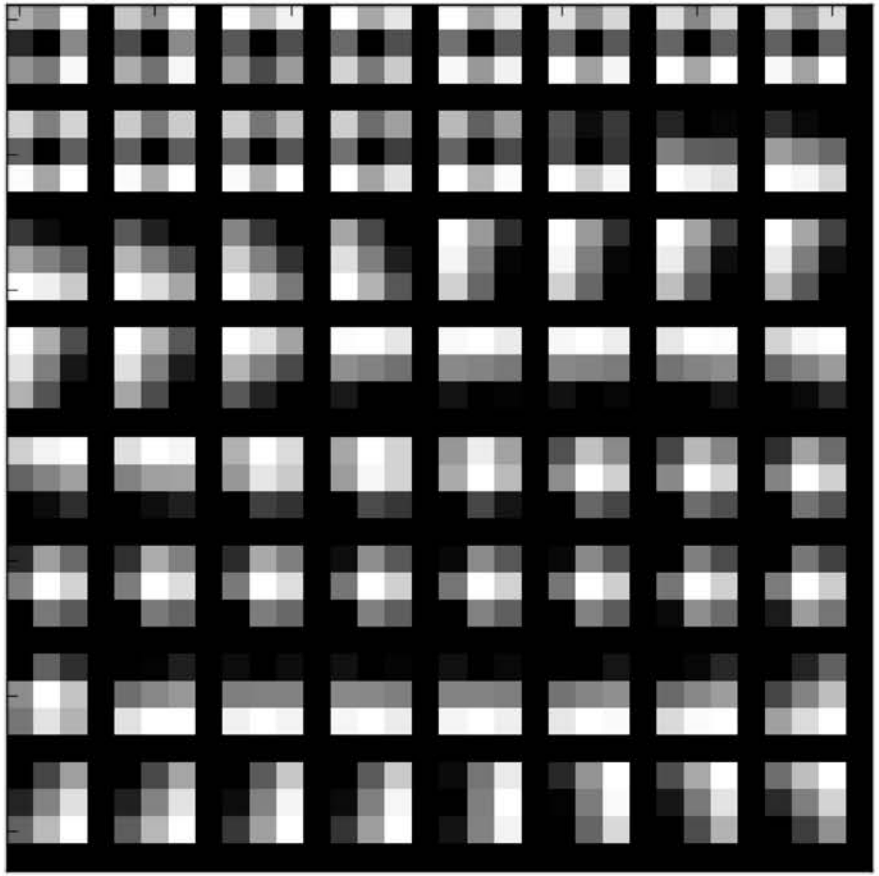
Use smaller filters to condense information

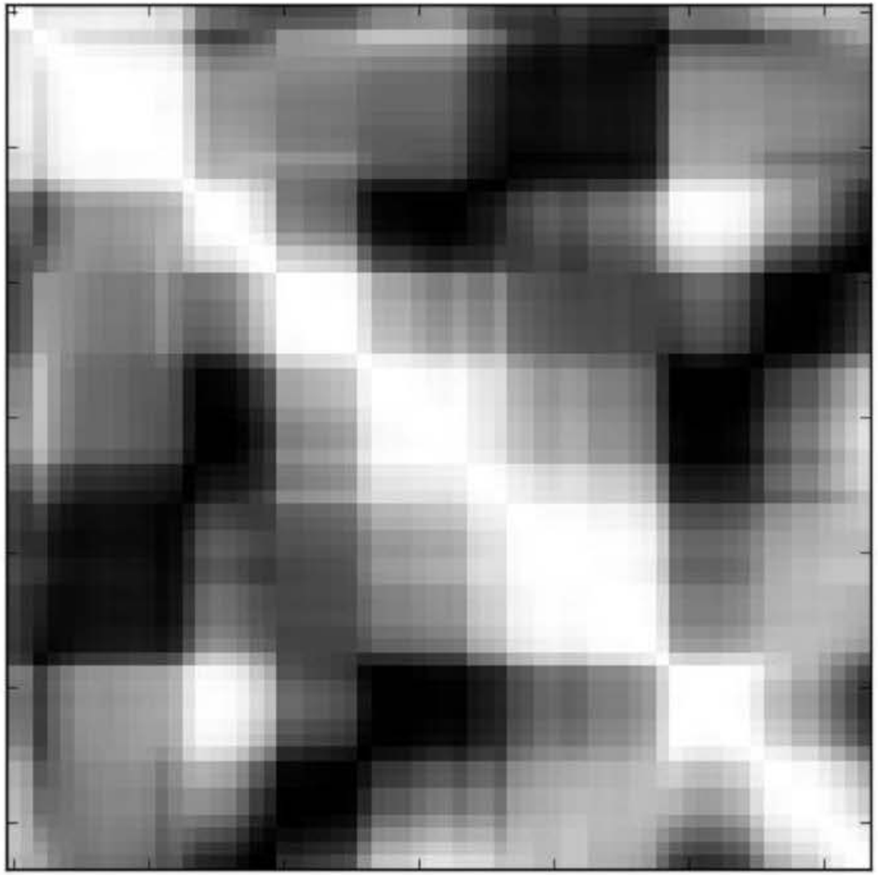




Look for **redundancy** in your layers

Small filters are faster… but can be highly correlated





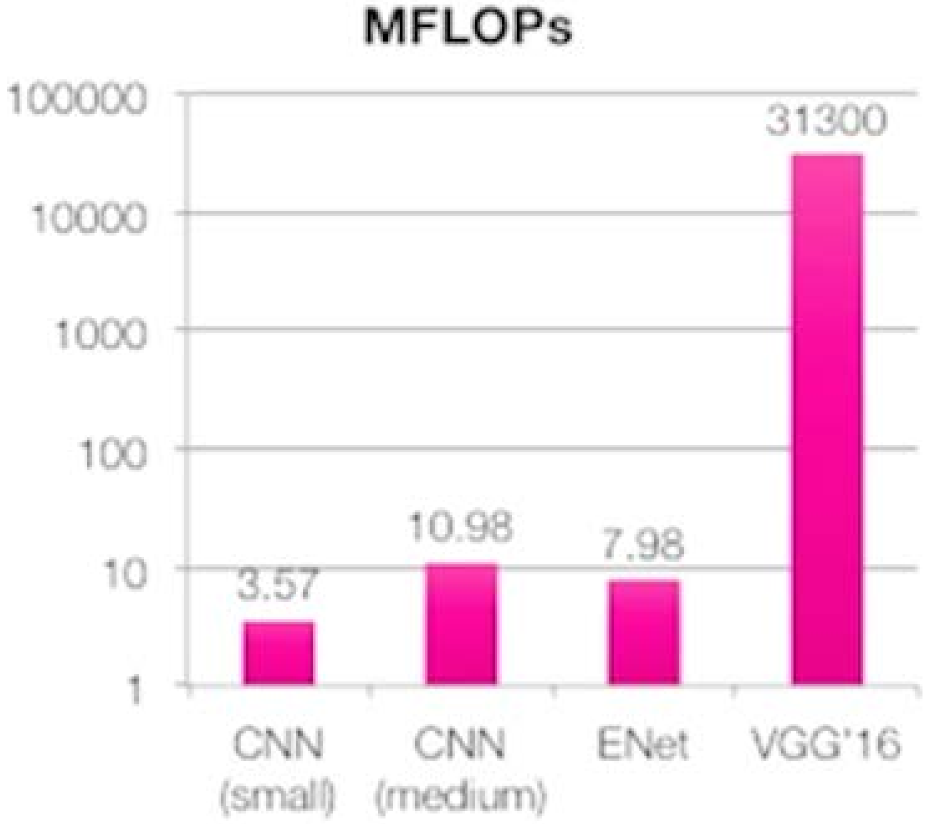
Match architecture to the problem

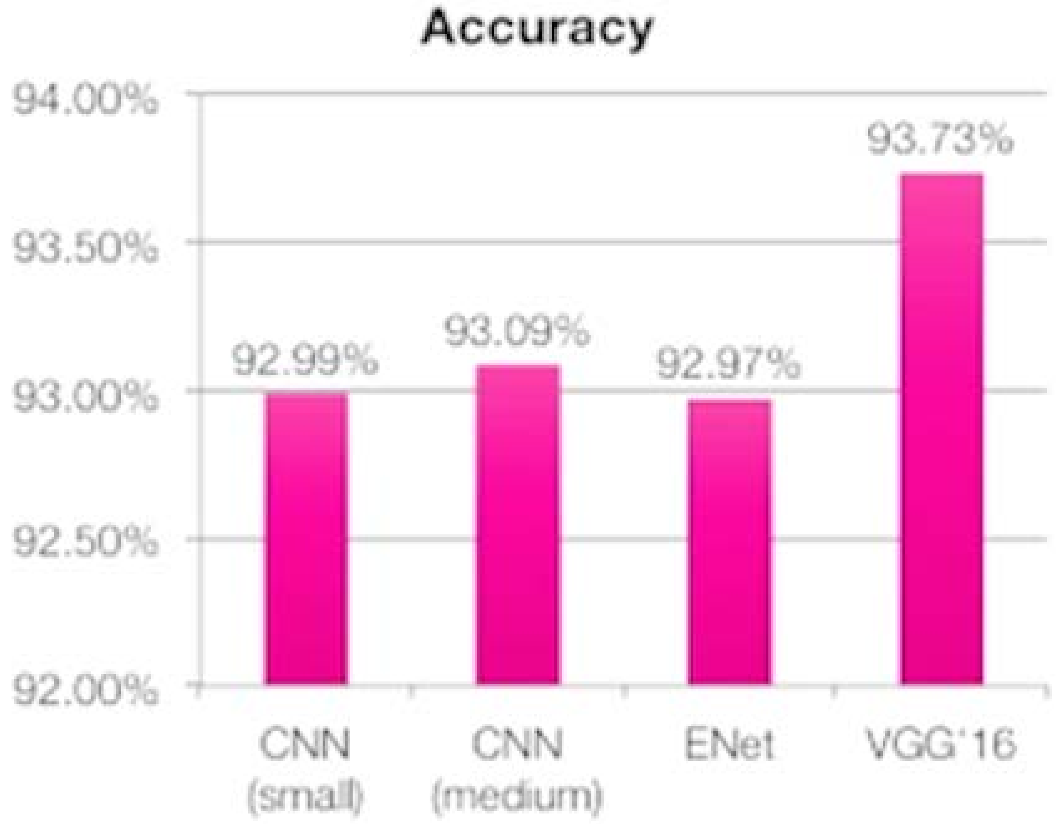
Avoid network architecture that is larger than needed

Problem	Object detection (& classification) IMAGENET	Facial action & attribute classification
Details	1000 classes ~224x224 pixels, color Objects with arbitrary scales / positions / orientations	20+ classes ~100x100 pixels, grayscale Faces only, upright & registered
Architectures	VGG'16 [1] - 16 layers (~30.9 GOP/image) ResNet [2] - 152 layers (~22.6 GOP/image) Others: Inception v4, E-Net	?

Small networks still work very well…

… and are sufficiently small for on device processing





ディープラーニング相談室

コンサルティング、システムインテグレーションなど各種ご相談に応じます

- ディープラーニングのシステム開発にお困りでしたら
- DL-HELP@nvidia.com
- までお問い合わせください。
- 内容に応じ、各種パートナー企業様をご紹介します。

DEEP
LEARNING
INSTITUTE

www.nvidia.com/dli