

**DEEP LEARNING FOR COMPUTER VISION
MUBARAK SHAH
CENTER FOR RESEARCH IN COMPUTER VISION (CRCV)**

shah@crcv.ucf.edu
<http://crcv.ucf.edu/>

CONTENTS

PART-I: Deep Learning: A Short Overview

PART II: Computer Vision Employing Deep Learning

PART-I: Deep Learning: A Short Overview

CAP6412
Advanced Computer Vision

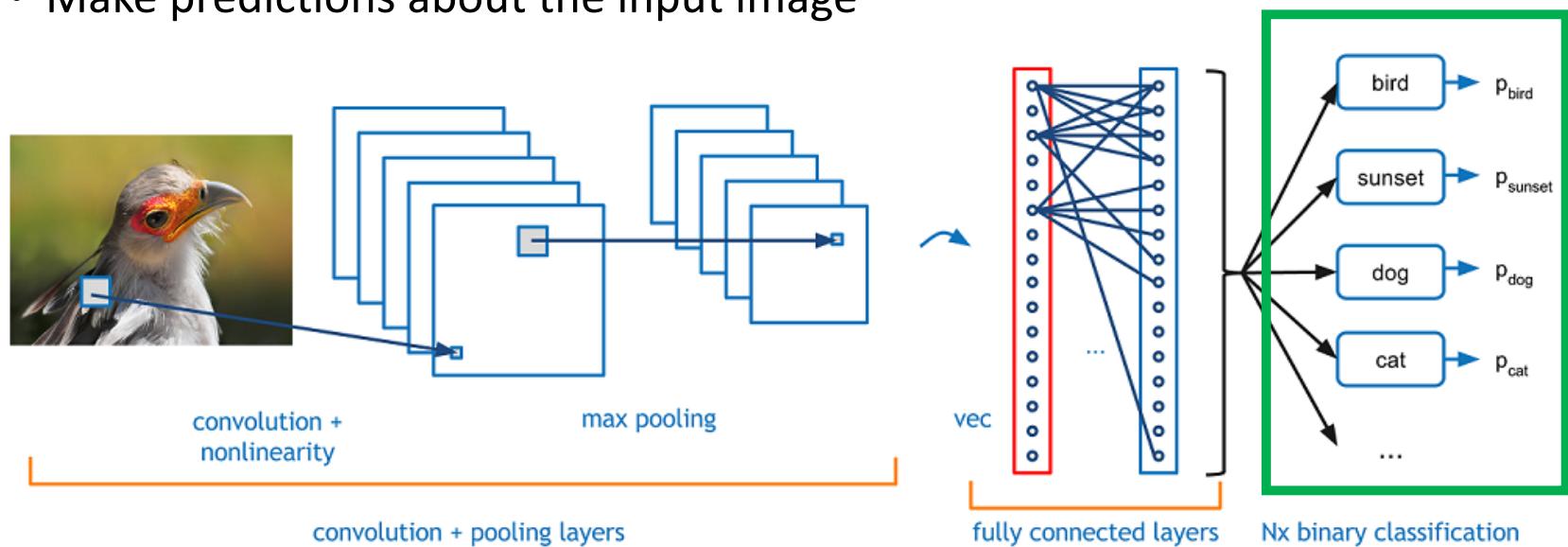
Yogesh S Rawat

What will we cover?

- Convolutional Neural Network
 - Their widespread use in deep learning
- Case study – AlexNet
- Network Training
- Recurrent Neural Networks
 - Working with sequential data

Convolutional Neural Network (CNN)

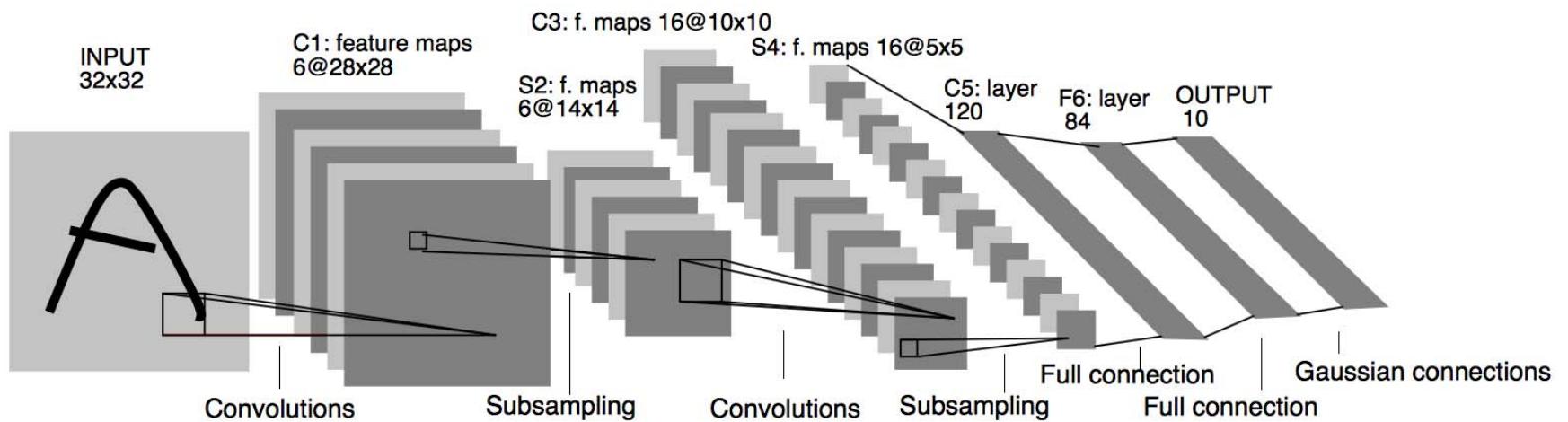
- A class of Neural Networks
 - Takes image as input
 - Make predictions about the input image



Source : <https://adeshpande3.github.io>

History

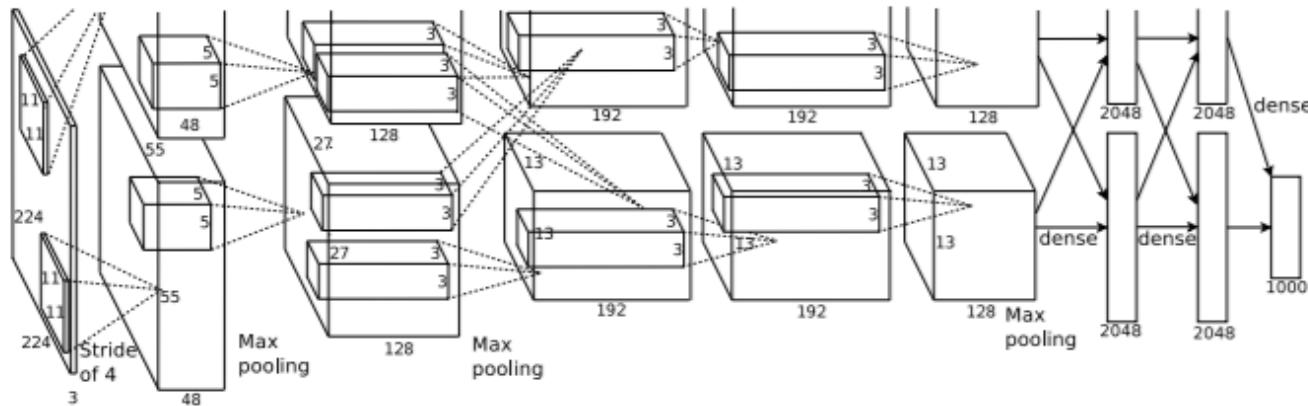
- The LeNet architecture (1990s)



Gradient-based learning applied to document recognition
LeCun Y, Bottou L, Bengio Y, Haffner P. Proceedings of the IEEE. 1998

First Strong Results

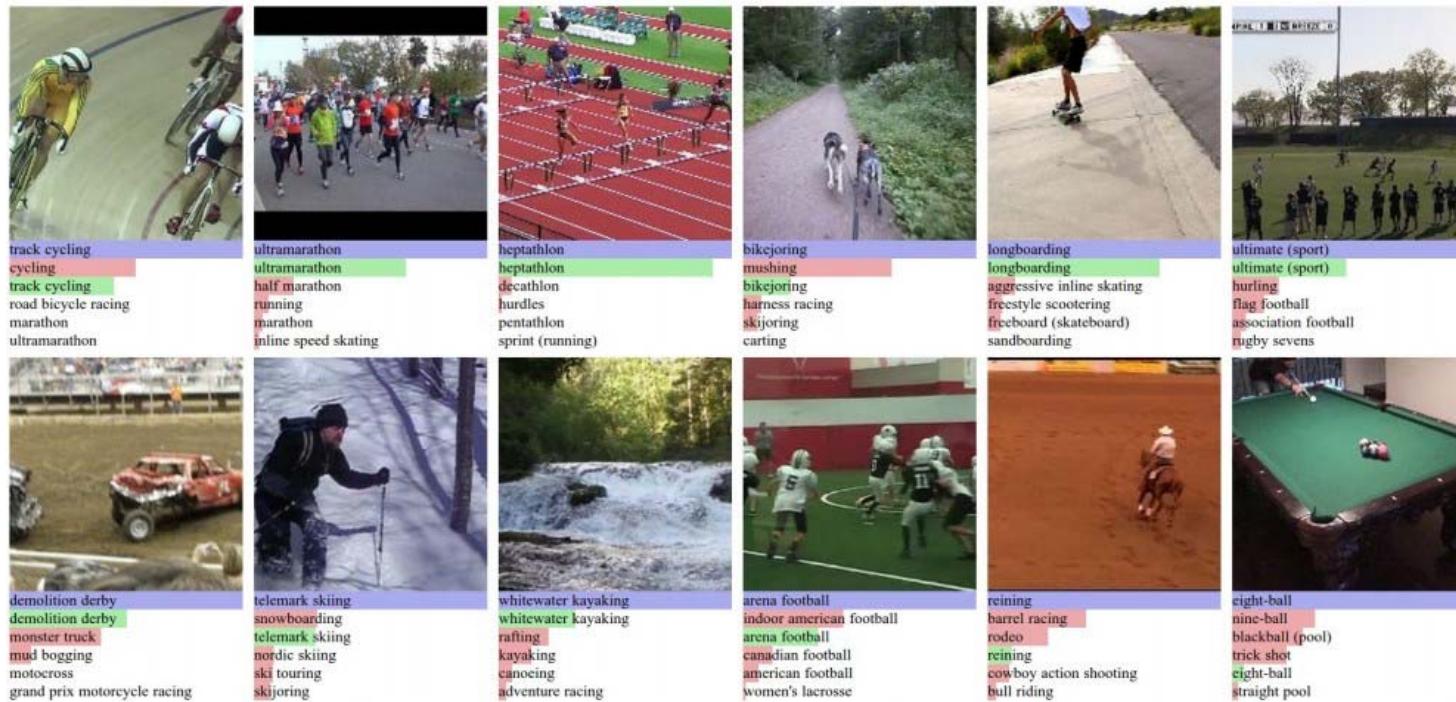
- AlexNet 2012
 - Winner of ImageNet Large-Scale Visual Recognition Challenge (ILSVRC 2012)
 - Error rate – 15.4% (the next best entry was at 26.2%)



Imagenet classification with deep convolutionalneural networks
Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton, 2012

Today: CNNs are everywhere

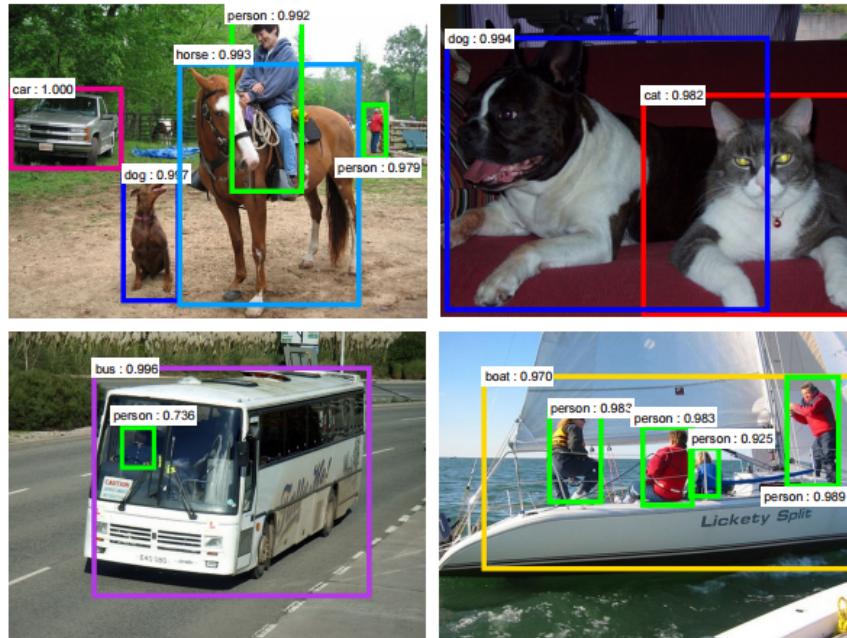
Classification



Source : <http://karpathy.github.io>

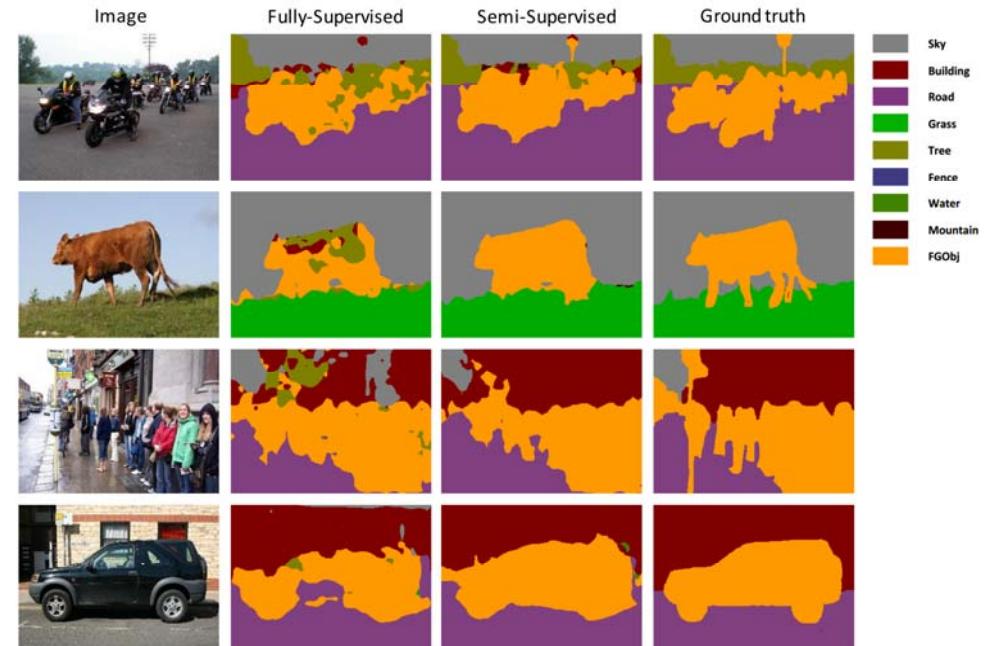
Today: CNNs are everywhere

Object detection



Faster R-CNN: Ren, He, Girshick, Sun 2015

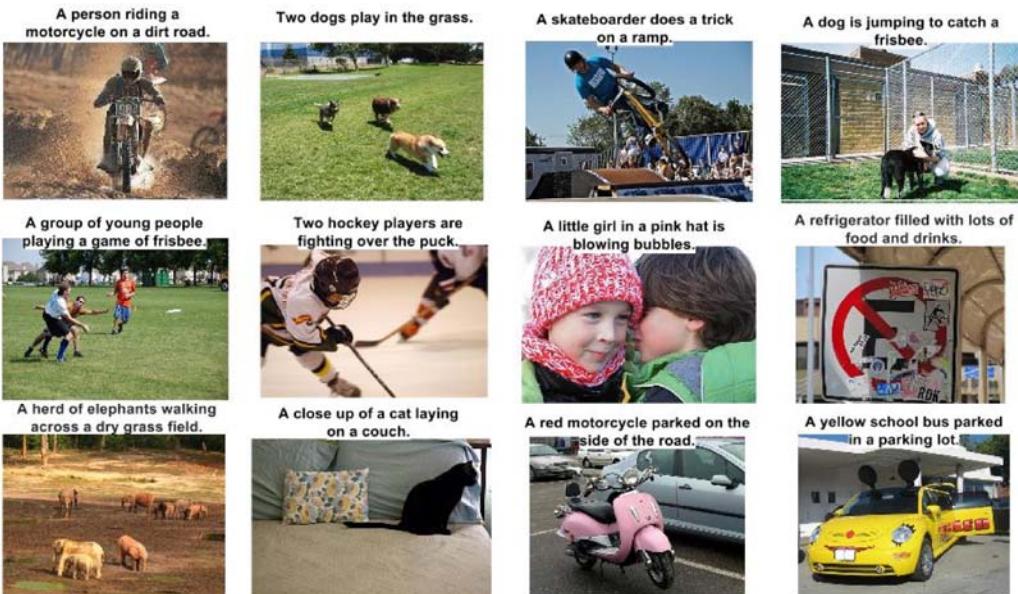
Semantic Segmentation



Semantic Segmentation Using GAN,
Nasim, Concetto, and Mubarak, 2017.

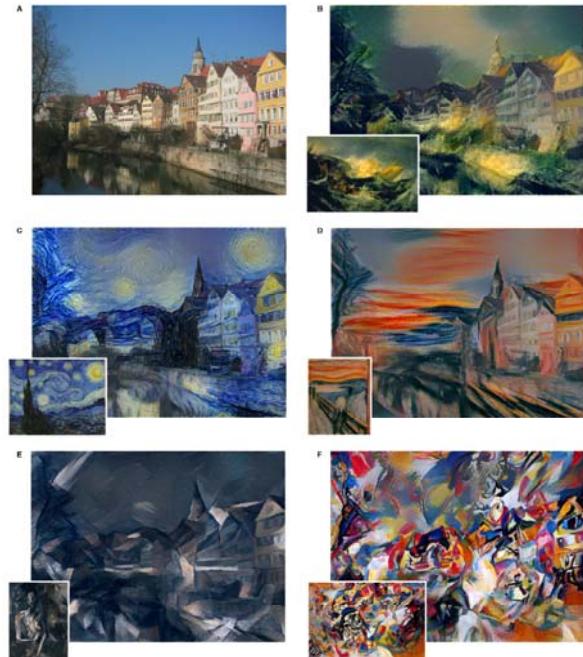
Today: CNNs are everywhere

Image captioning



"Show and tell: A neural image caption generator."
Vinyals, Oriol, et al. CVPR 2015.

Style transfer



A Neural Algorithm of Artistic Style
L. Gatys et al. 2015).

CNN – Not just images

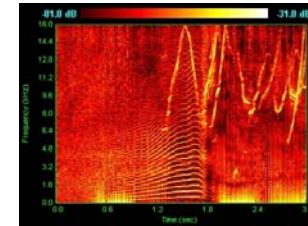
- Natural Language Processing (NLP)

- Text classification
 - Word to vector

```
I = [0 1 0 1 1 0]
Love = [1 0 1 0 0 0]
NLP = [0 1 0 1 0 0]
And = [1 0 1 0 0 0]
Like = [1 0 0 0 0 1]
Dogs = [0 0 0 0 1 0]
```

- Audio Research

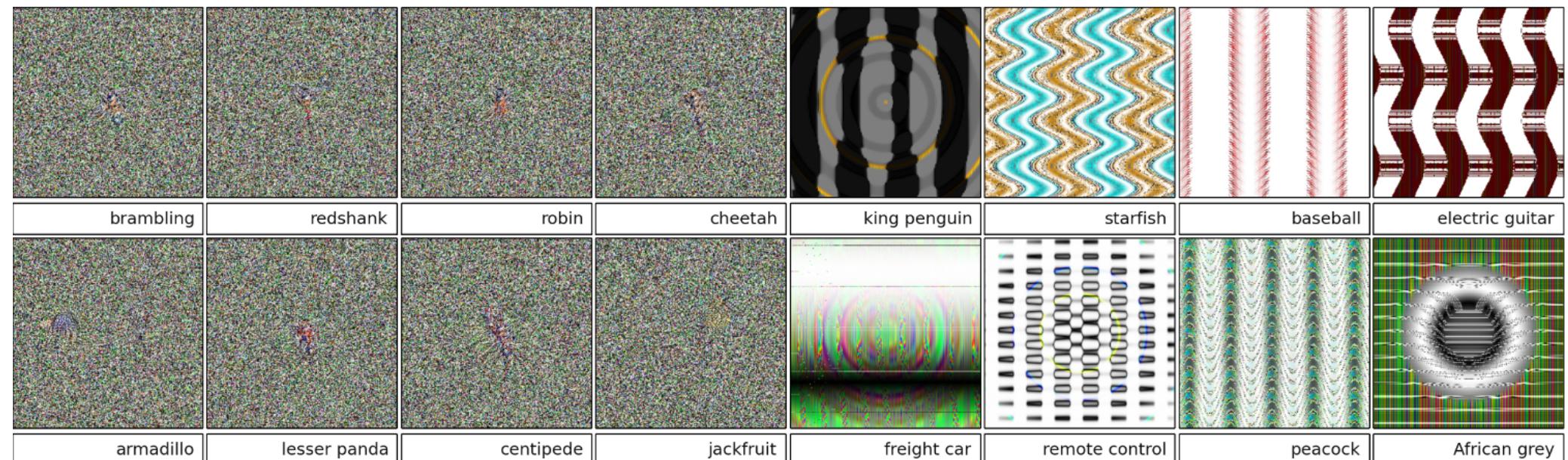
- Speech recognition
 - Can be represented as spectrograms



- Converting data to a matrix (2-D) format

- 1D convolution – Audio, EEG, etc.
 - 3D convolution - Videos

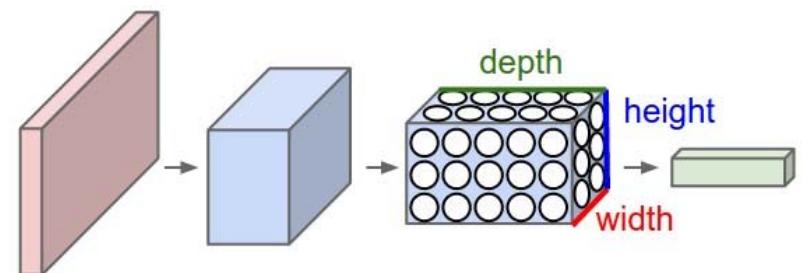
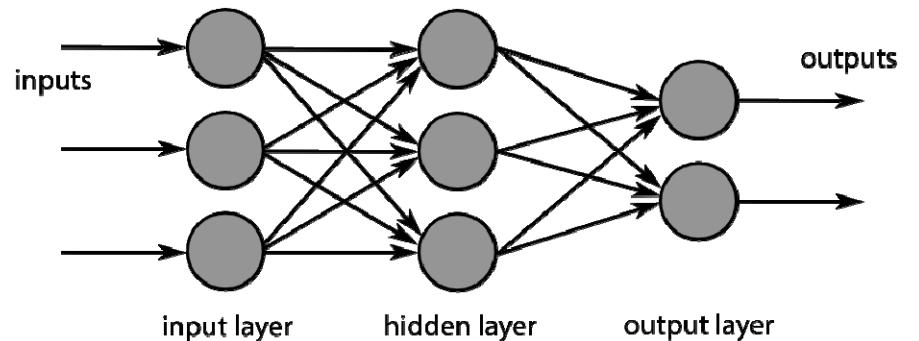
CNN – Is it perfect?



Convolutional Neural Network

Neural Network vs CNN

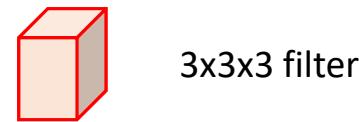
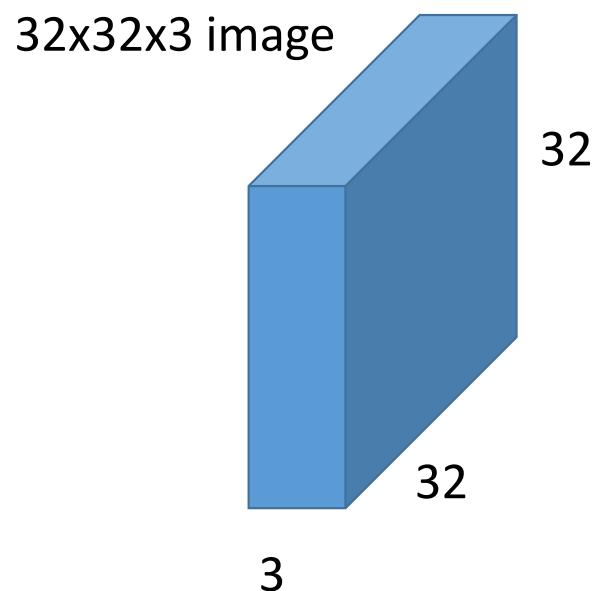
- Neural Network
 - Fully connected layers
- Image as input in neural network
 - Size of feature vector = $H \times W \times C$
 - For 256x256 RGB image
 - 196,608 dimensions
- CNN - Special type of neural network
 - Operate with volume of data
 - Weight sharing in form of kernels



Source: <http://cs231n.github.io>

Convolution

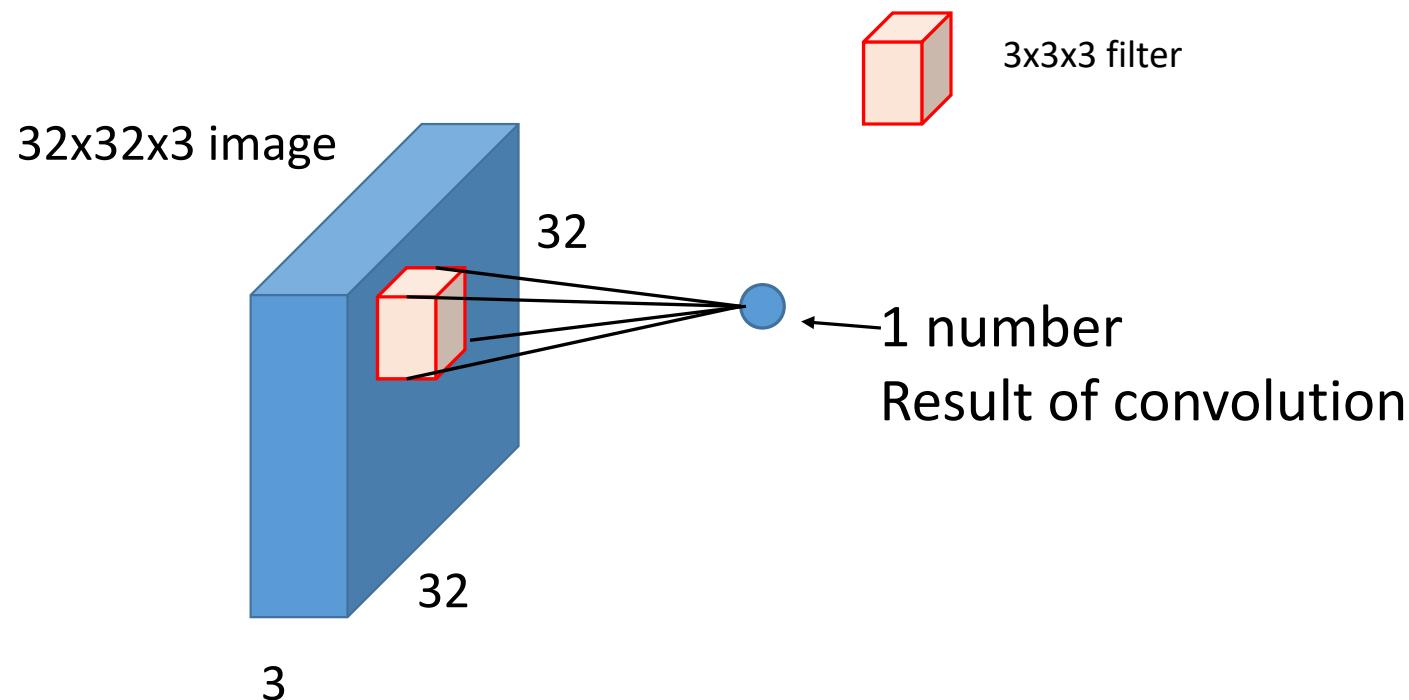
- Core building block of a CNN
 - Spatial structure of image is preserved



A filter/kernel is **convolved** with the image

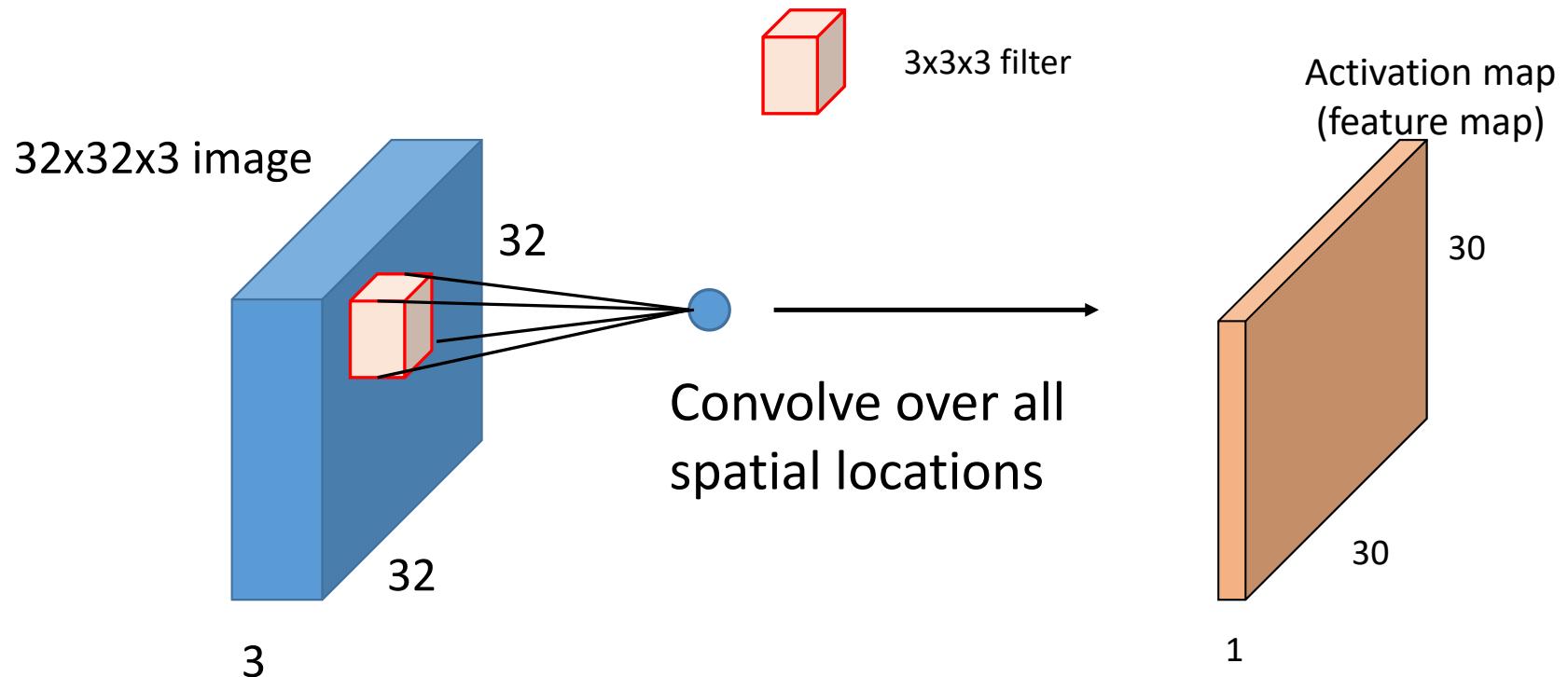
Convolution

- Convolution at one spatial location



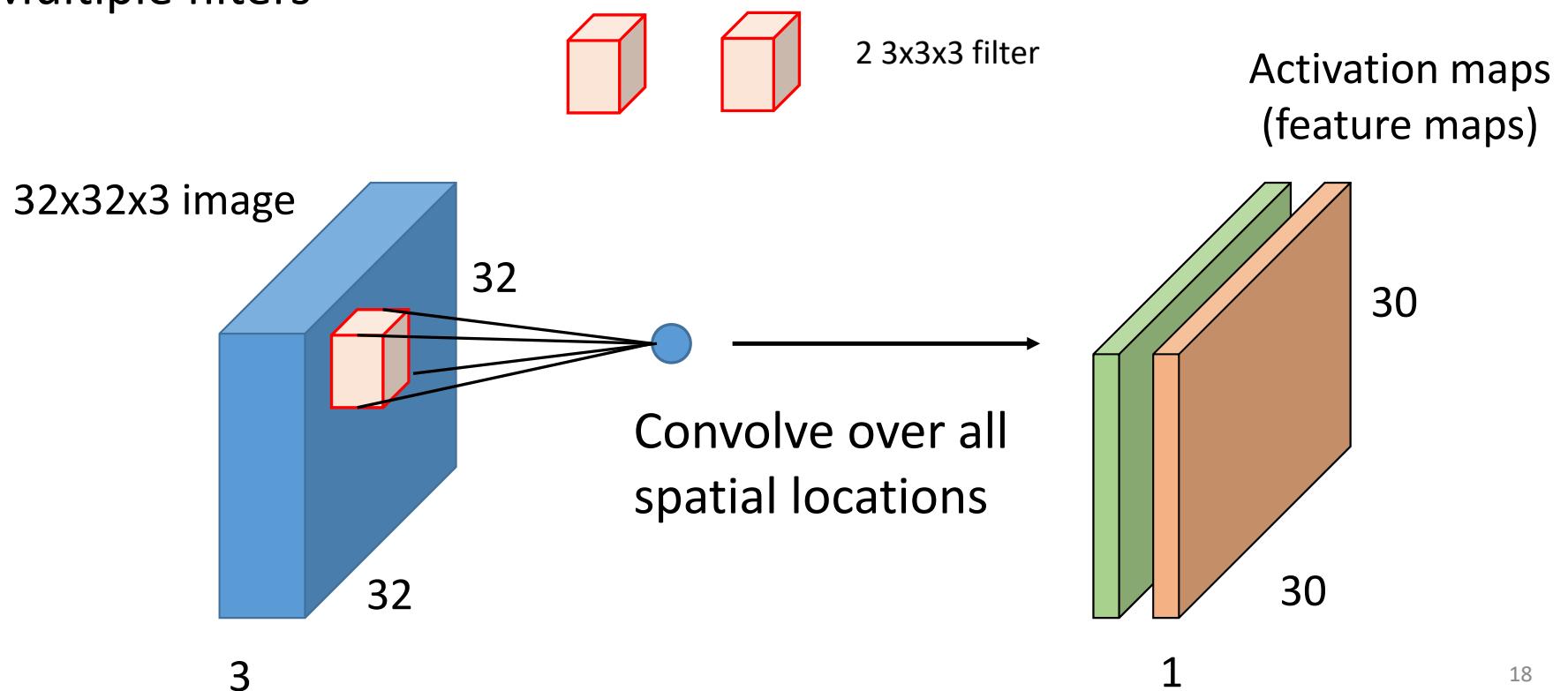
Convolution

- Convolution over whole image



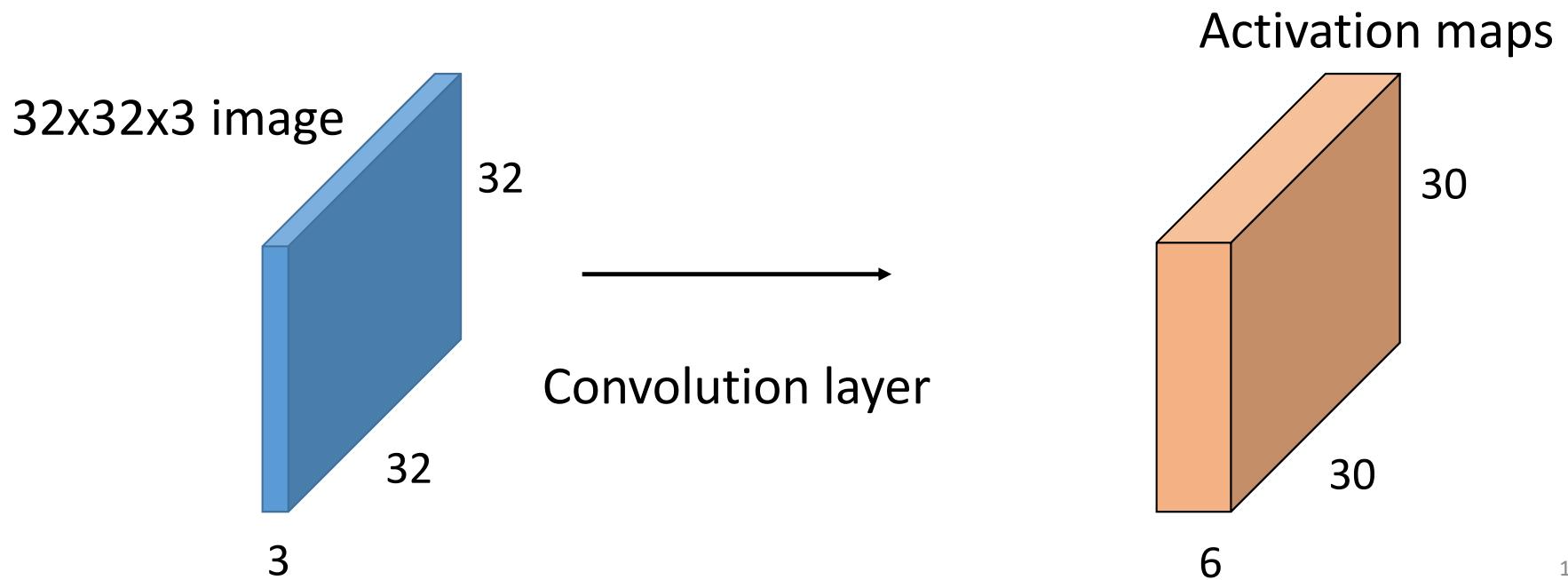
Convolution

- Multiple filters



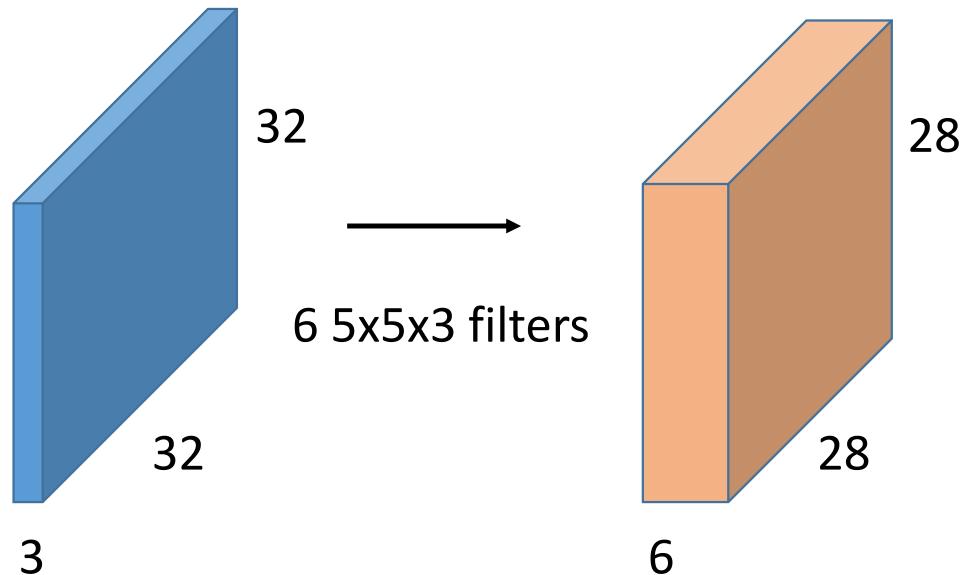
Convolution

- One convolution layer
 - 6 $3 \times 3 \times 3$ kernels



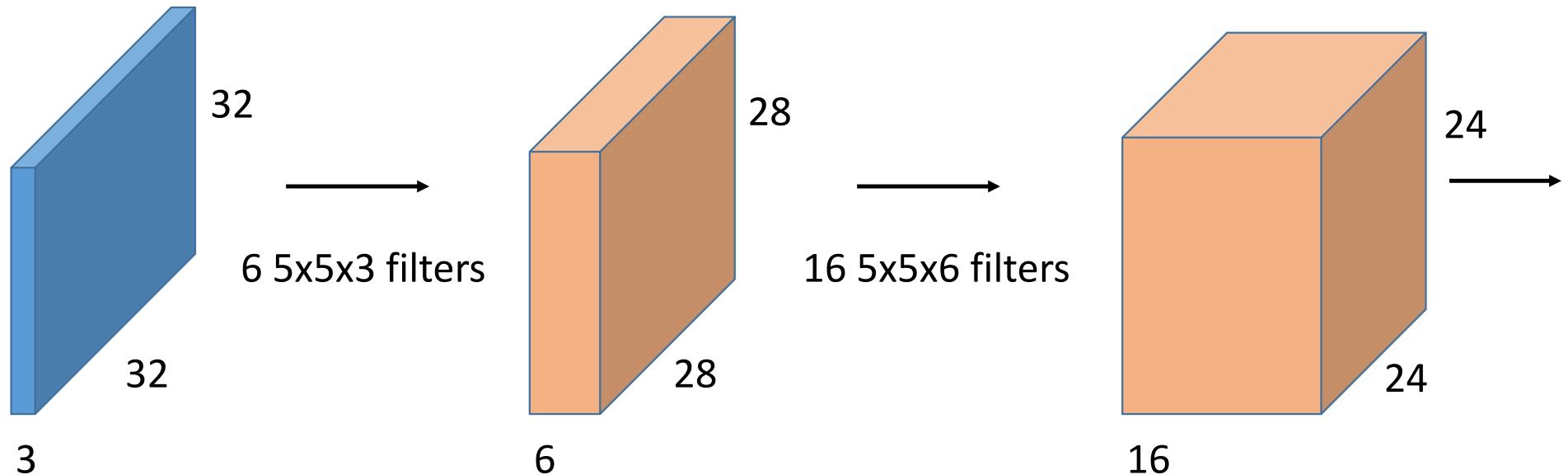
Convolutional Network

- Convolution network is a sequence of these layers



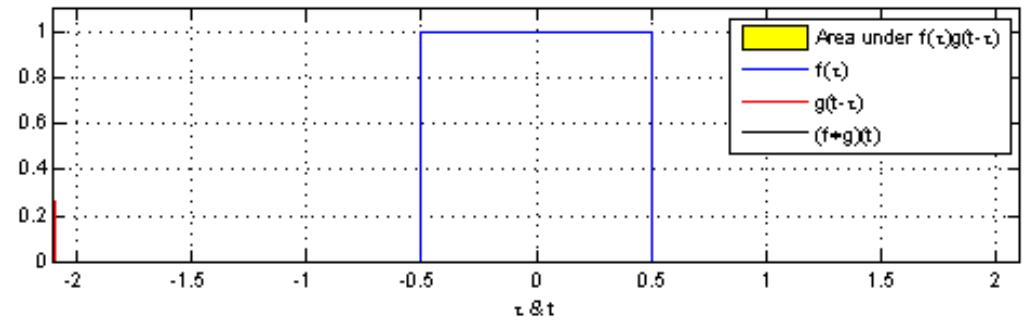
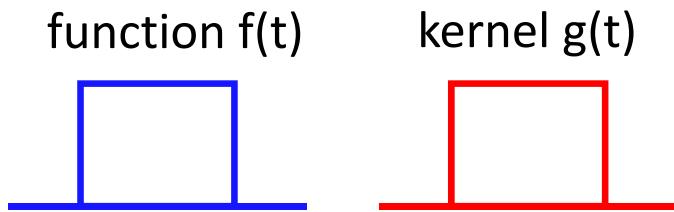
Convolutional Network

- Convolution network is a sequence of these layers



Convolution Operation

- Convolution of two functions f and g



$$(f * g)(t) = \int_{-\infty}^{+\infty} f(\tau)g(t - \tau)d\tau$$

In CNN we use 2D convolutions

Demo

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	0	0	0

Input image

filter

1	0	1
0	1	0
1	0	1

4

output

Demo

1	1	1	0	0
0	1	1	1	0
0	0	1	1	1
0	0	1	1	0
0	1	0	0	0

Input image

filter

1	0	1
0	1	0
1	0	1

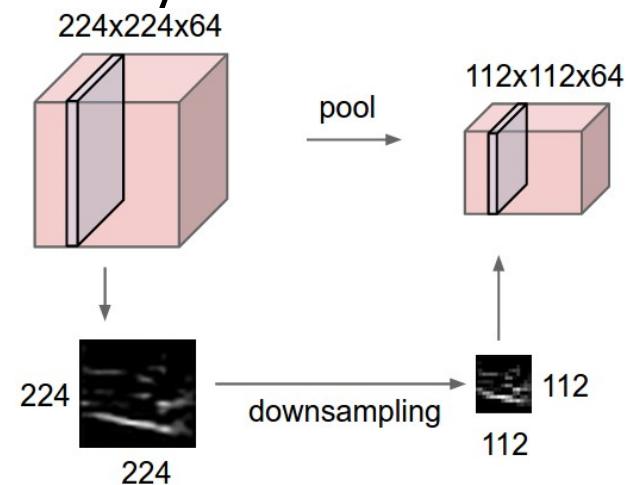
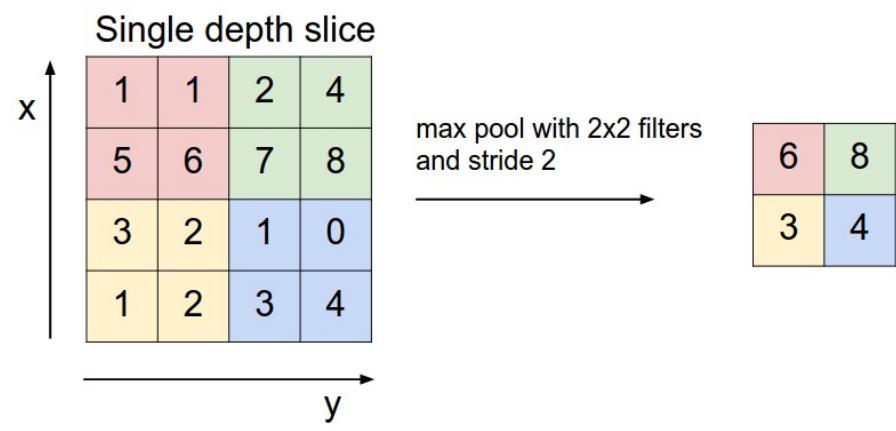
4 3

4	3	

output

Pooling

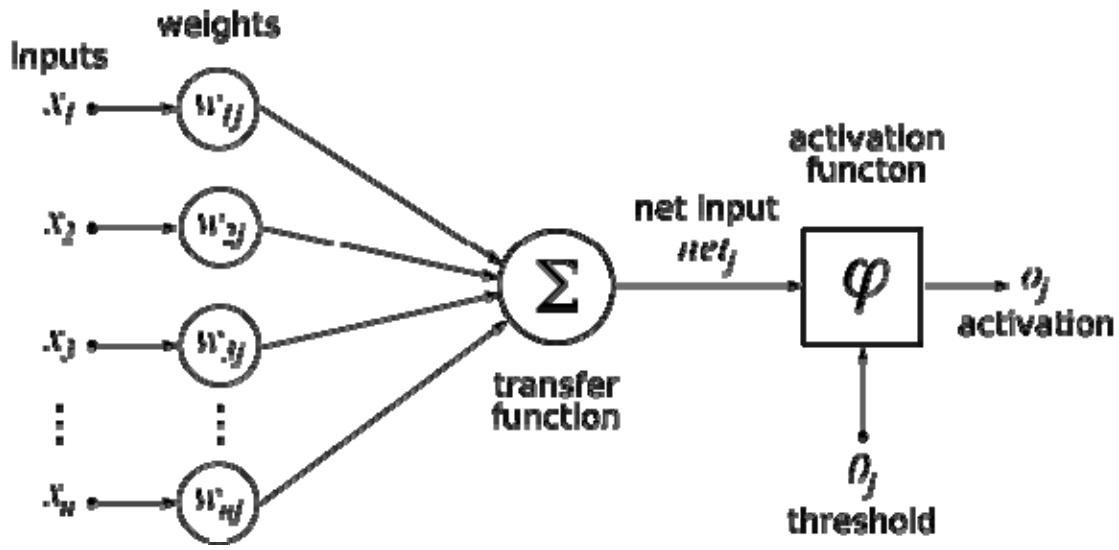
- Introduce translation invariance
- Makes the representations smaller
- Operates over each activation map independently



Source : <http://cs231n.github.io>

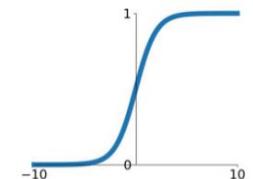
Activation Functions

- Introduces Non-Linearity



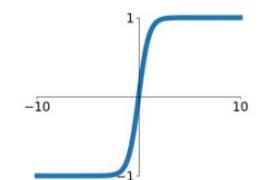
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



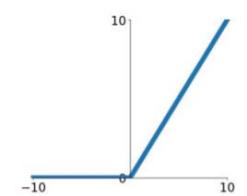
tanh

$$\tanh(x)$$



ReLU

$$\max(0, x)$$

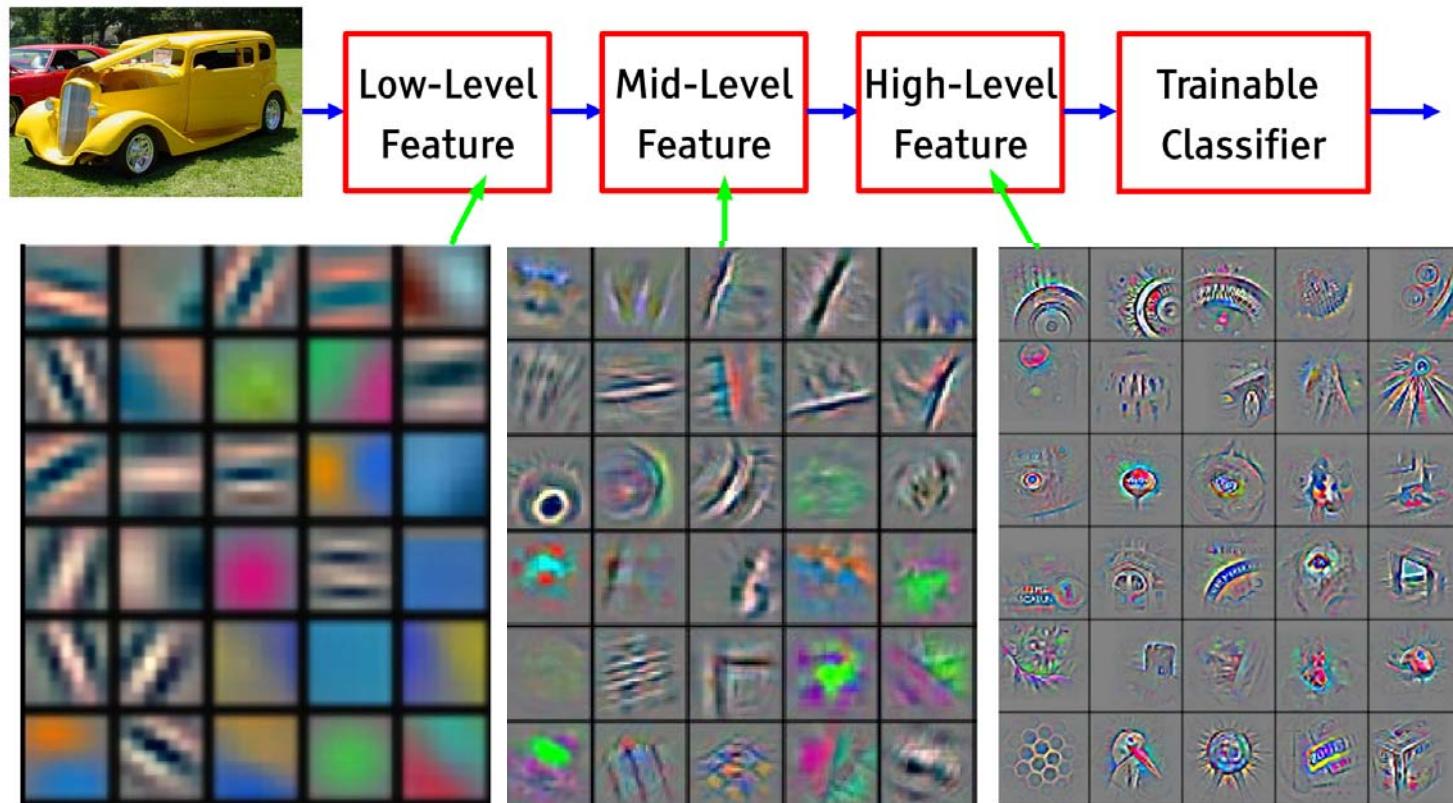


Convolution - Intuition



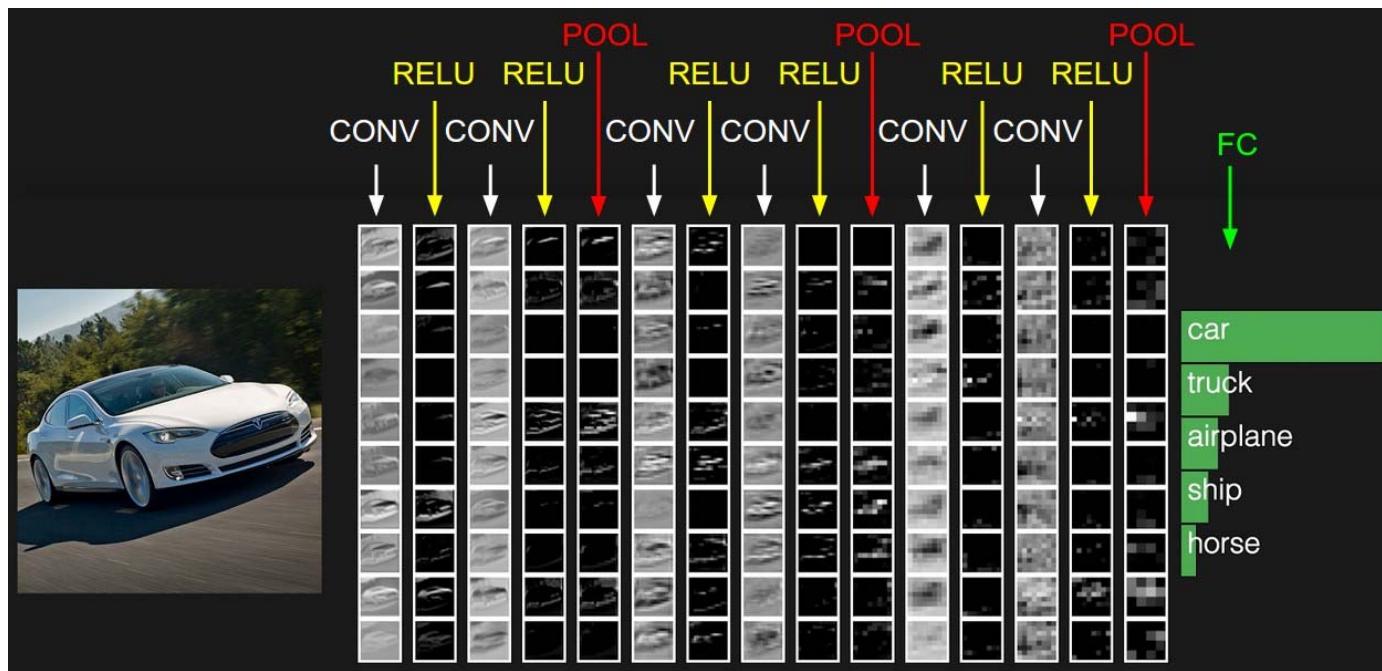
Source : https://cs.nyu.edu/~fergus/tutorials/deep_learning_cvpr12/

Visualizing Convolution



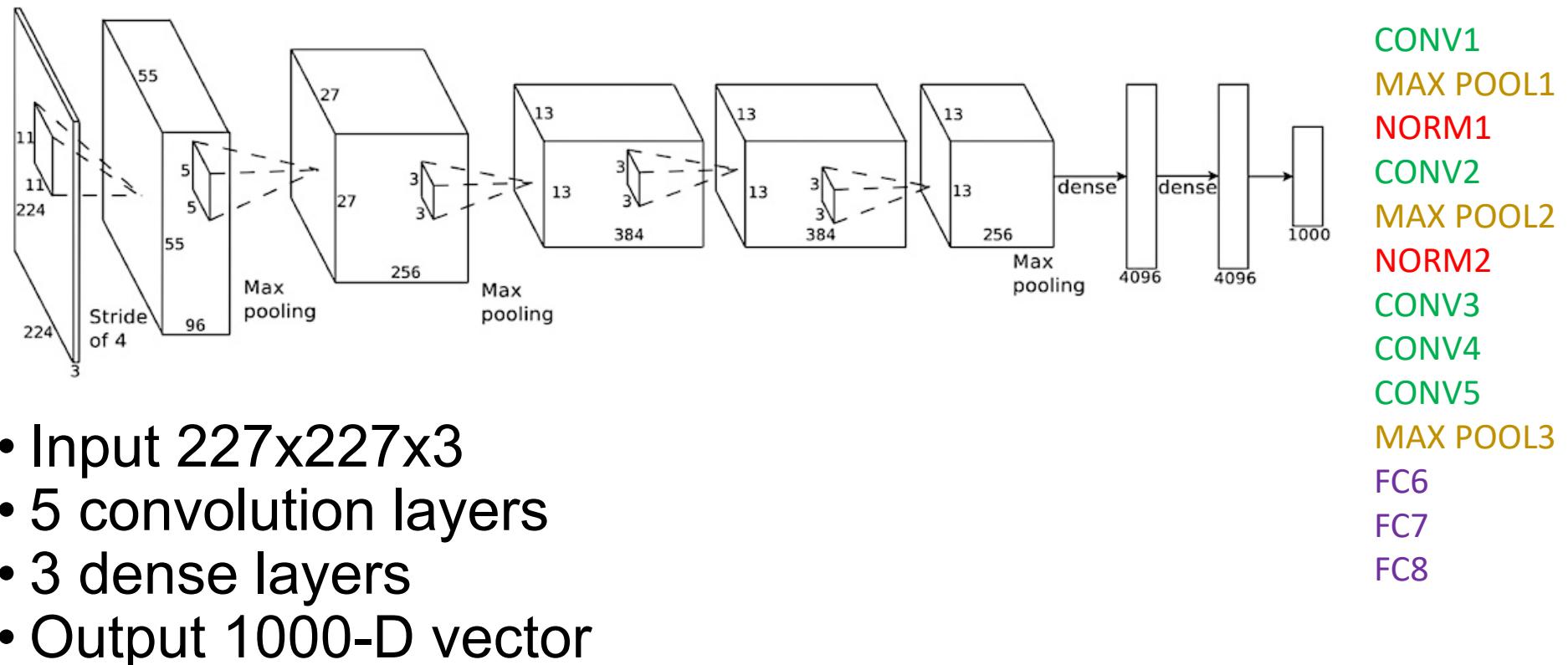
Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Visualizing CNN

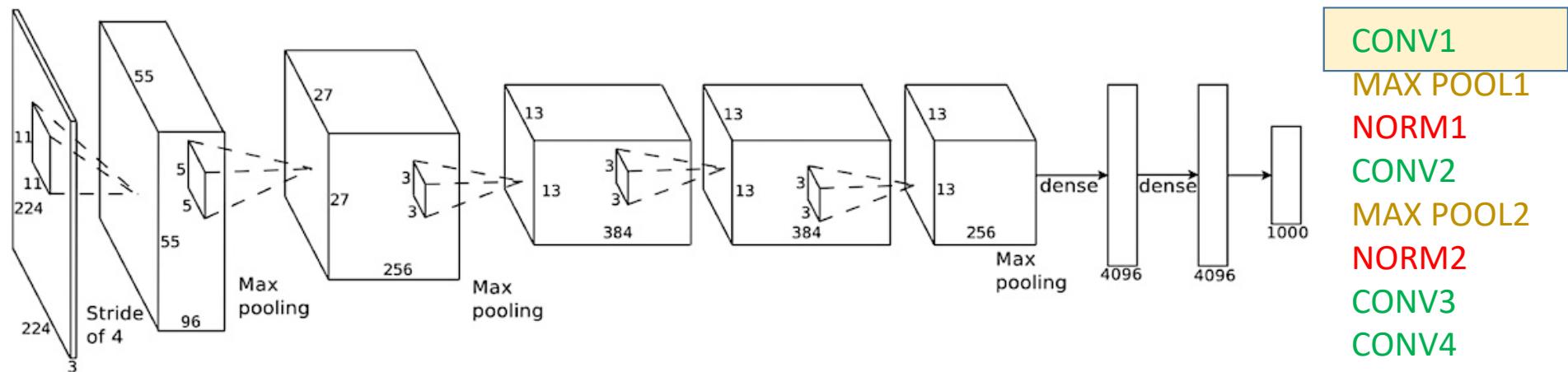


Source : <http://cs231n.github.io>

AlexNet : Network Size

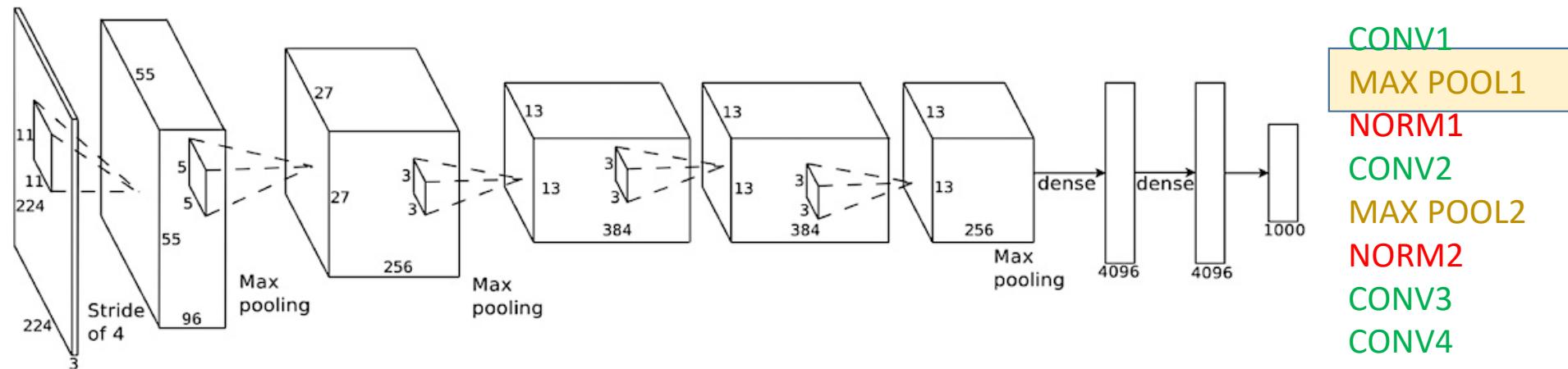


AlexNet : Network Size



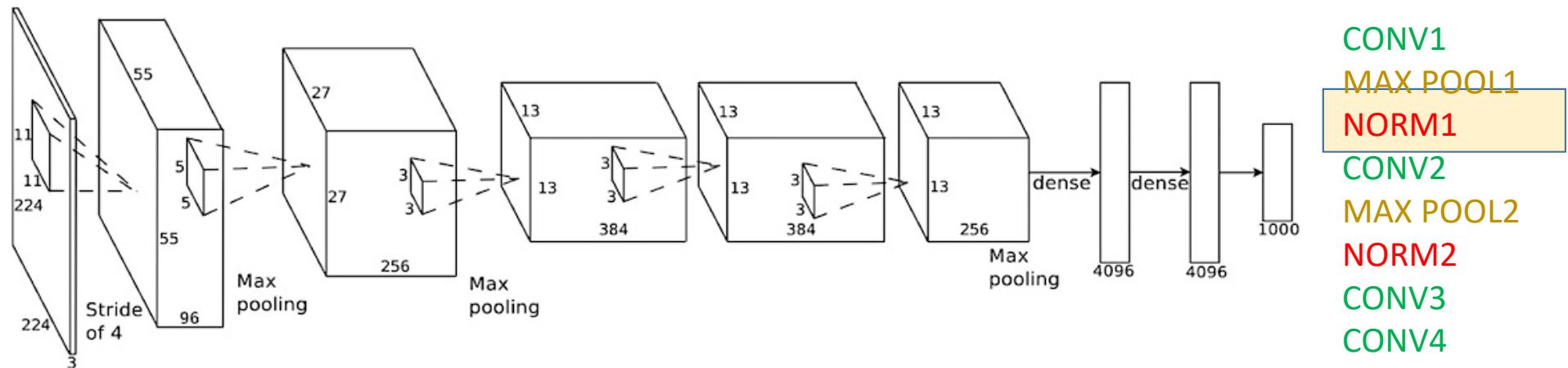
- Input: 227x227x3 images
- First layer (CONV1): 96 11x11 filters applied at stride 4
- What is the output volume size? $(227-11)/4+1 = 55$
- What is the number of parameters? $11 \times 11 \times 3 \times 96 = 35K$

AlexNet : Network Size



- After CONV1: 55x55x96
- Second layer (POOL1): 3x3 filters applied at stride 2
- What is the output volume size? $(55-3)/2+1 = 27$
- What is the number of parameters in this layer? 0

AlexNet : Network Size



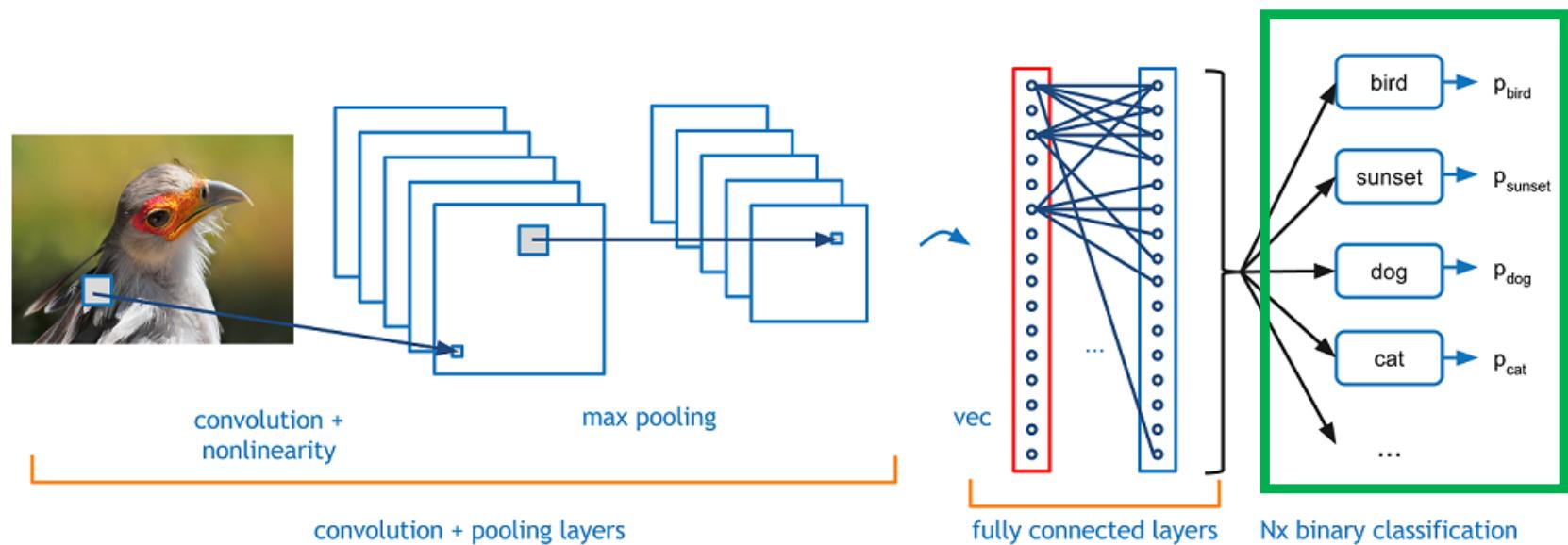
- After POOL1: 27x27x96
- Third layer (NORM1): Normalization
- What is the output volume size? 27x27x96

AlexNet : Network Size

1.	[227x227x3] INPUT		
2.	[55x55x96] CONV1: 96 11x11 filters at stride 4, pad 0	CONV1	35K
3.	[27x27x96] MAX POOL1: 3x3 filters at stride 2	MAX POOL1	
4.	[27x27x96] NORM1: Normalization layer	NORM1	
5.	[27x27x256] CONV2: 256 5x5 filters at stride 1, pad 2	CONV2	307K
6.	[13x13x256] MAX POOL2: 3x3 filters at stride 2	MAX POOL2	
7.	[13x13x256] NORM2: Normalization layer	NORM2	
8.	[13x13x384] CONV3: 384 3x3 filters at stride 1, pad 1	CONV3	884K
9.	[13x13x384] CONV4: 384 3x3 filters at stride 1, pad 1	CONV4	1.3M
10.	[13x13x256] CONV5: 256 3x3 filters at stride 1, pad 1	CONV5	442K
11.	[6x6x256] MAX POOL3: 3x3 filters at stride 2	MAX POOL3	
12.	[4096] FC6: 4096 neurons	FC6	37M
13.	[4096] FC7: 4096 neurons	FC7	16M
14.	[1000] FC8: 1000 neurons (class scores)	FC8	4M

Network Training

Convolutional Neural Network (CNN)



Source : <https://adeshpande3.github.io>

Loss Function

- Way to define how good the network is performing
 - In terms of prediction
- Network training (Optimization)
 - Find the best network parameters to minimize the loss

$$L(W) = \frac{1}{N} \sum_{i=1}^N L_i(f(x_i, W), y_i)$$

The diagram shows the loss function equation with red arrows pointing from labels to its components:

- An arrow points from "Loss function" to the summation symbol (\sum).
- An arrow points from "input" to the term $f(x_i, W)$.
- An arrow points from "Ground truth" to the term y_i .
- An arrow points from "Network parameters" to the term W .
- An arrow points from "network" to the term $f(x_i, W)$.

Loss Functions

- Cross entropy

$$-\frac{1}{N} \sum_{i=1}^N (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))$$

Ground-truth Predicted value



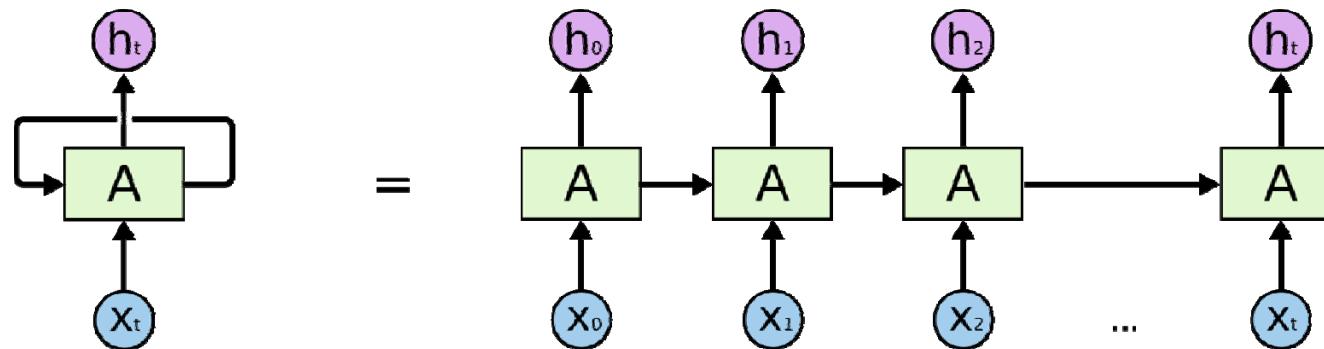
- Kullback–Leibler divergence (KL divergence)

$$KL(P||Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right)$$

Recurrent Neural Network

Recurrent Neural Network

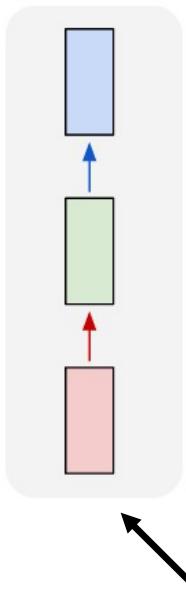
- Process sequences
- Feedback from previous time step



An unrolled recurrent neural network.

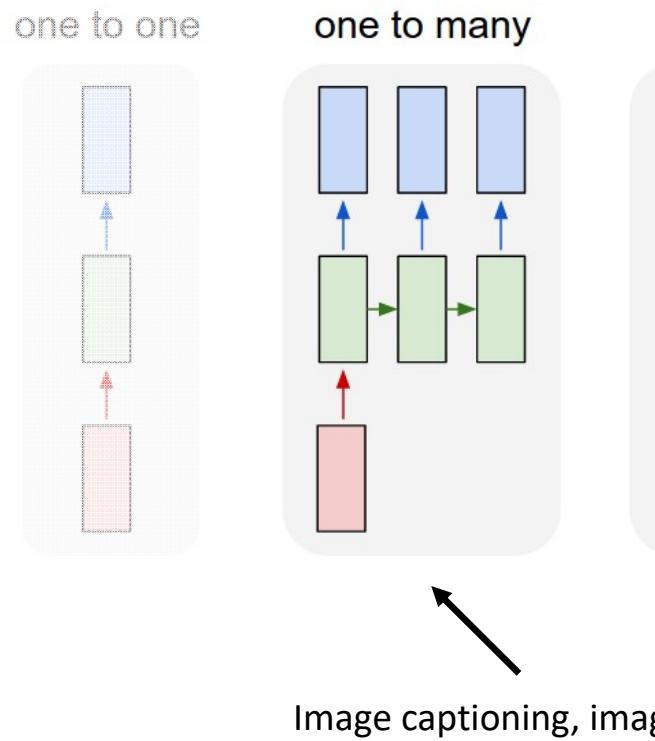
Recurrent Neural Network

one to one



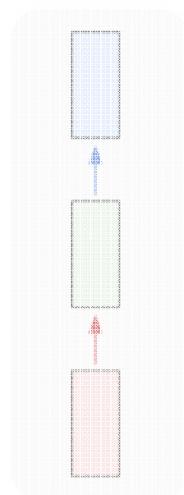
Vanilla Neural Networks

Recurrent Neural Network

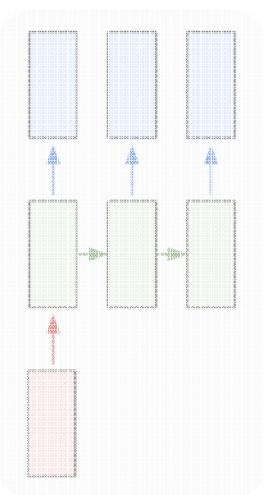


Recurrent Neural Network

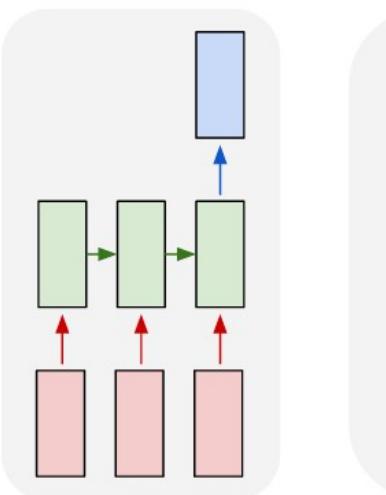
one to one



one to many



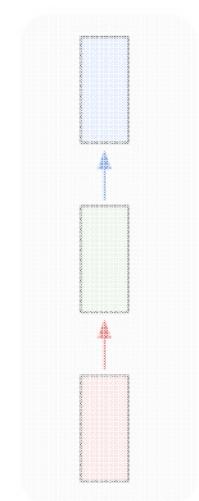
many to one



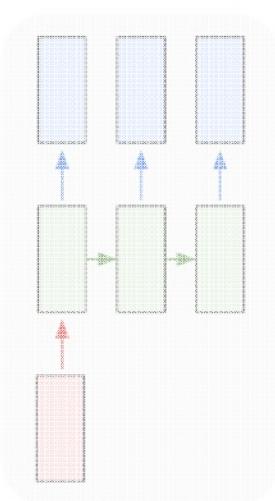
Sentiment Classification
sequence of words -> sentiment

Recurrent Neural Network

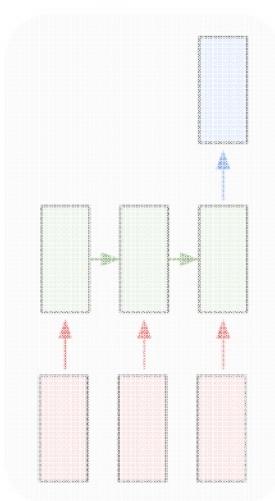
one to one



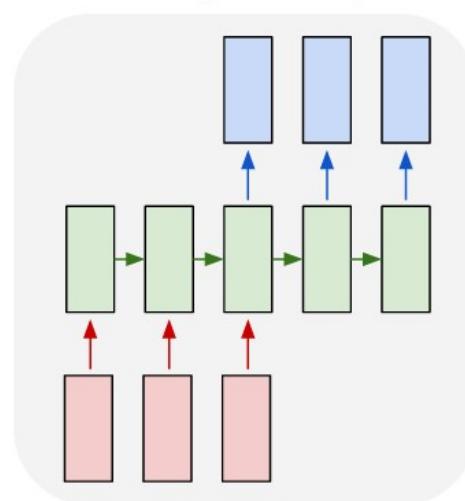
one to many



many to one

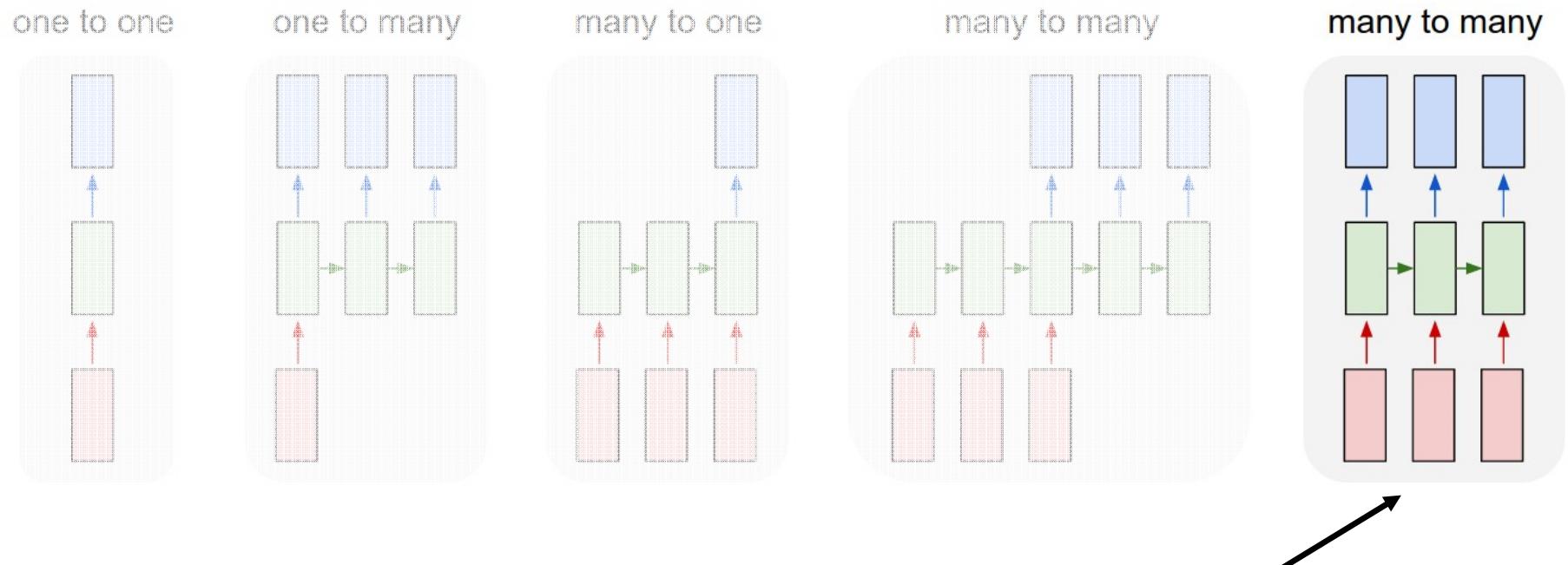


many to many



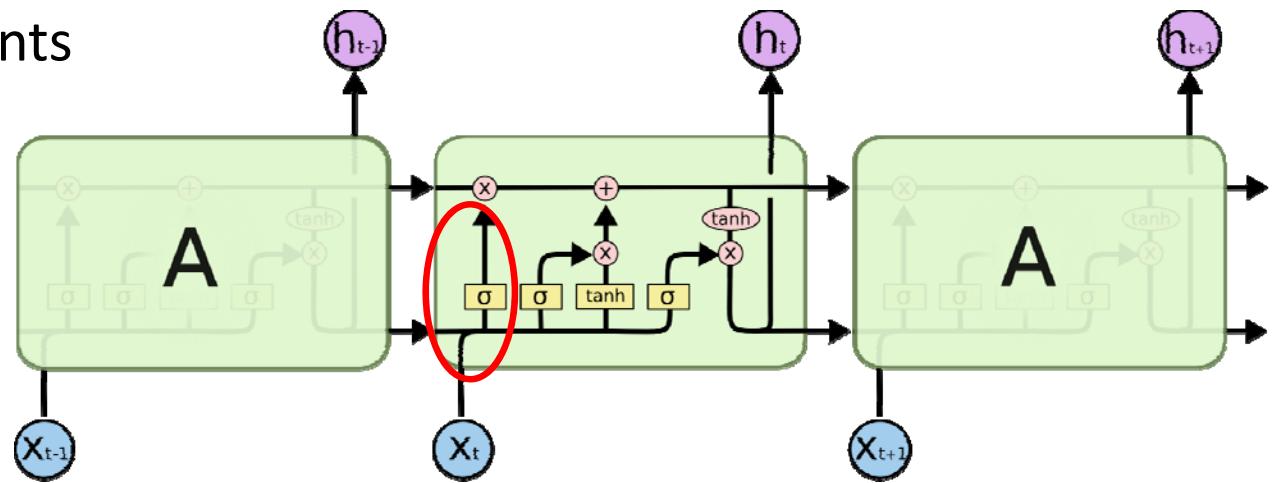
Machine Translation
seq of words \rightarrow seq of words

Recurrent Neural Network



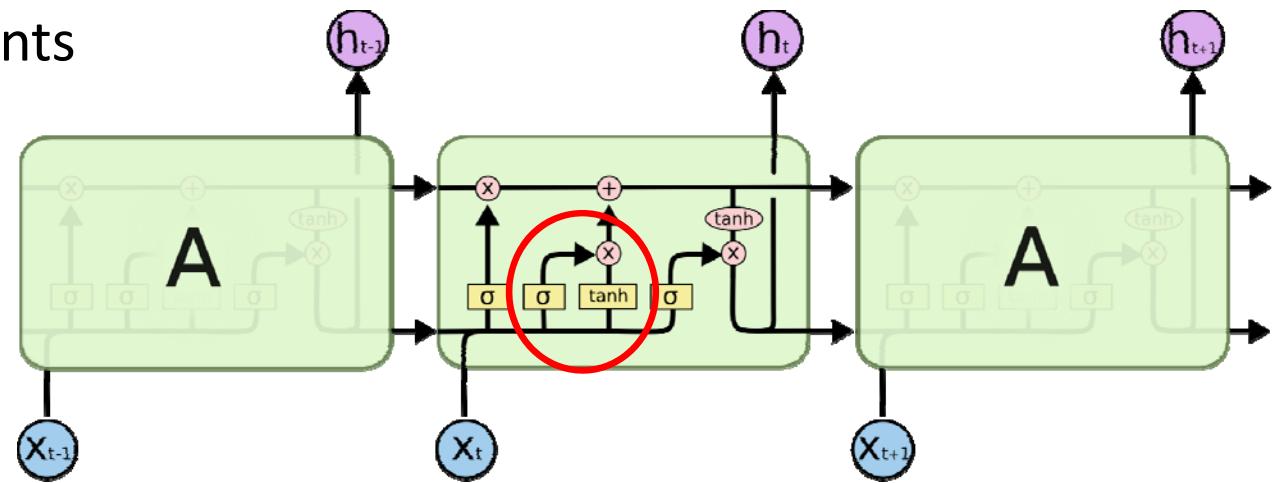
Long-Short Term Memory (LSTM)

- Three main components
 - Forget gate layer



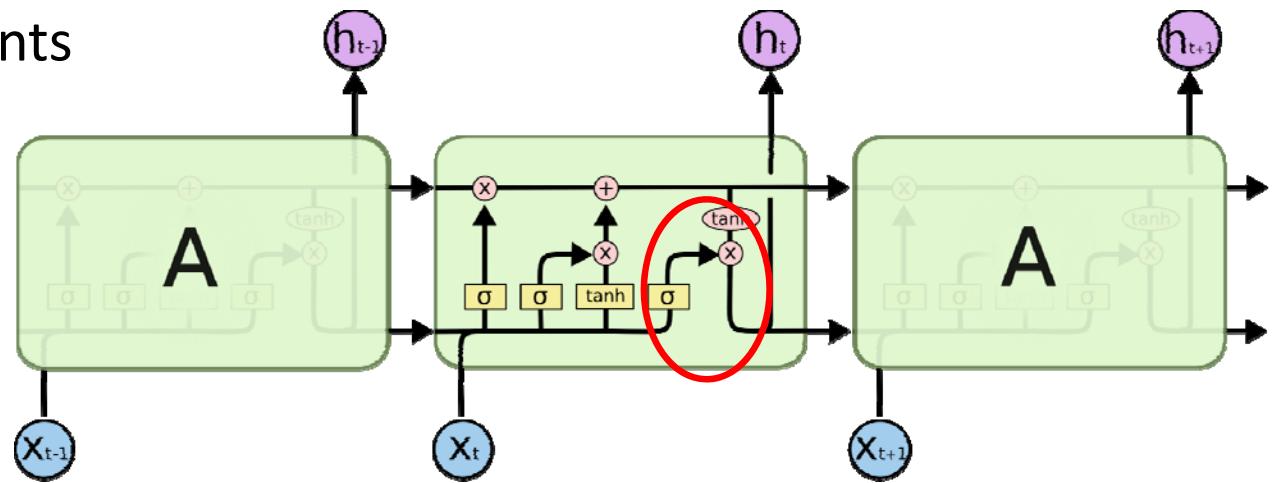
Long-Short Term Memory (LSTM)

- Three main components
 - Forget gate layer
 - Input gate layer

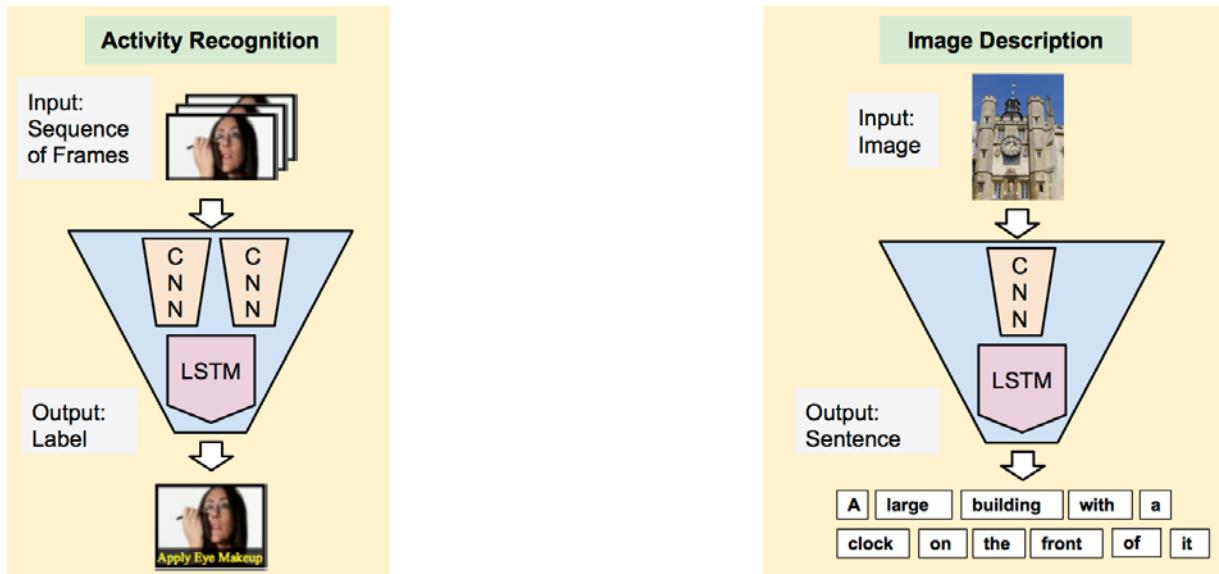


Long-Short Term Memory (LSTM)

- Three main components
 - Forget gate layer
 - Input gate layer
 - Output gate layer



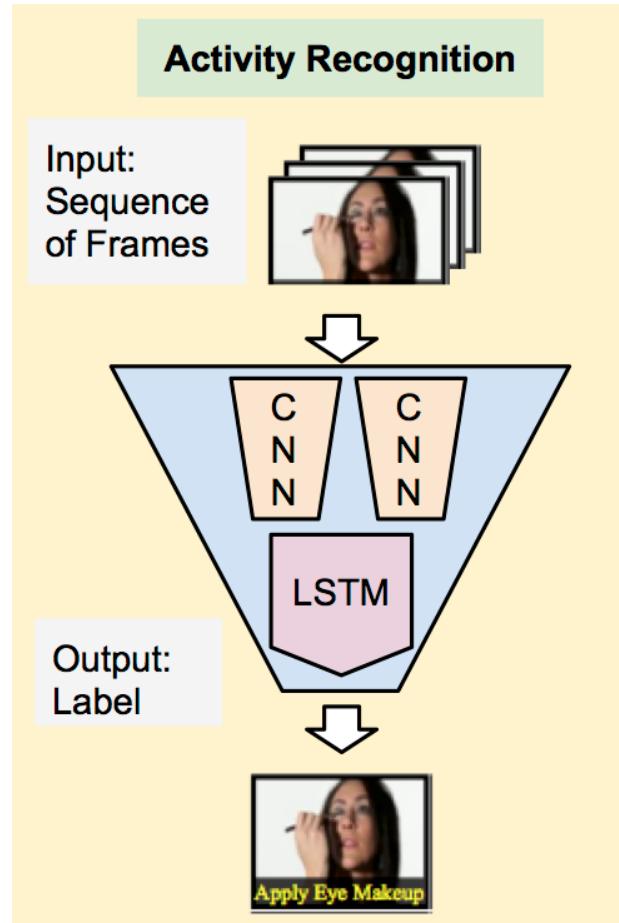
CNN-LSTM



<http://jeffdonahue.com/lrcn/>

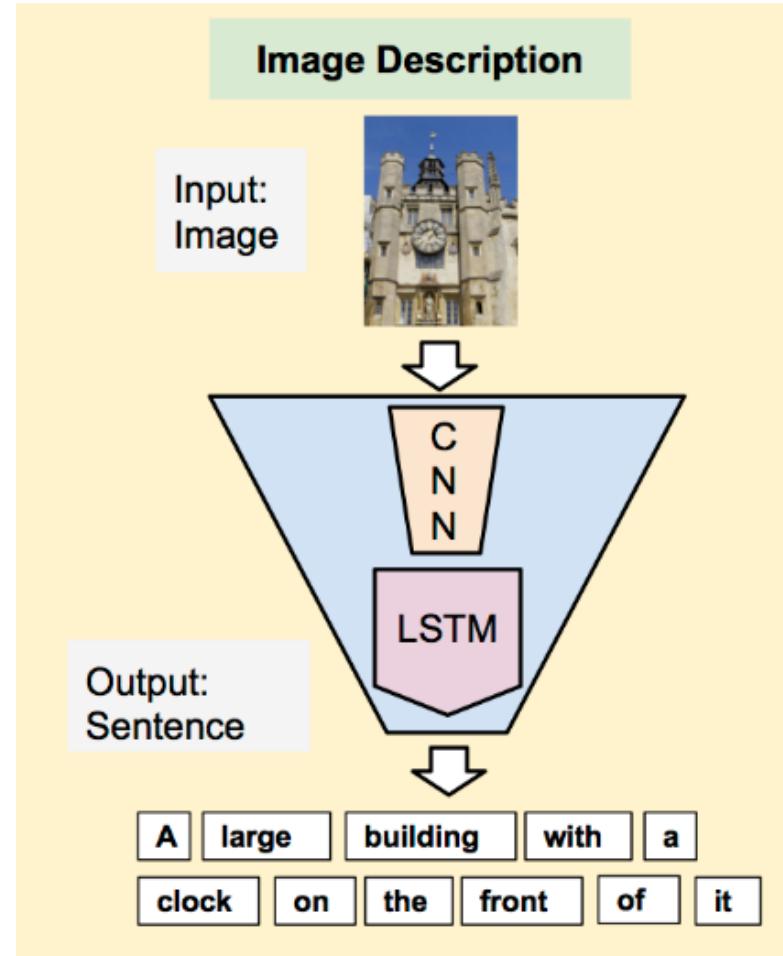
Vinyals et. al. Show and Tell, 2015
Jef et. al. Long-term Recurrent Convolutional Networks 2015

CNN-LSTM



<http://jeffdonahue.com/lrcn/>

CNN-LSTM



<http://jeffdonahue.com/lrcn/>

Vinyals et. al. Show and Tell, 2015
Jef et. al. Long-term Recurrent Convolutional Networks 2015

Summary

- CNN
 - Convolution
 - Pooling
 - RELU
- Case study – AlexNet
- Network Training
 - Loss Function
- Recurrent Neural Networks
 - Working with sequential data

PART-I: Deep Learning: A Short Overview

CAP6412
Advanced Computer Vision

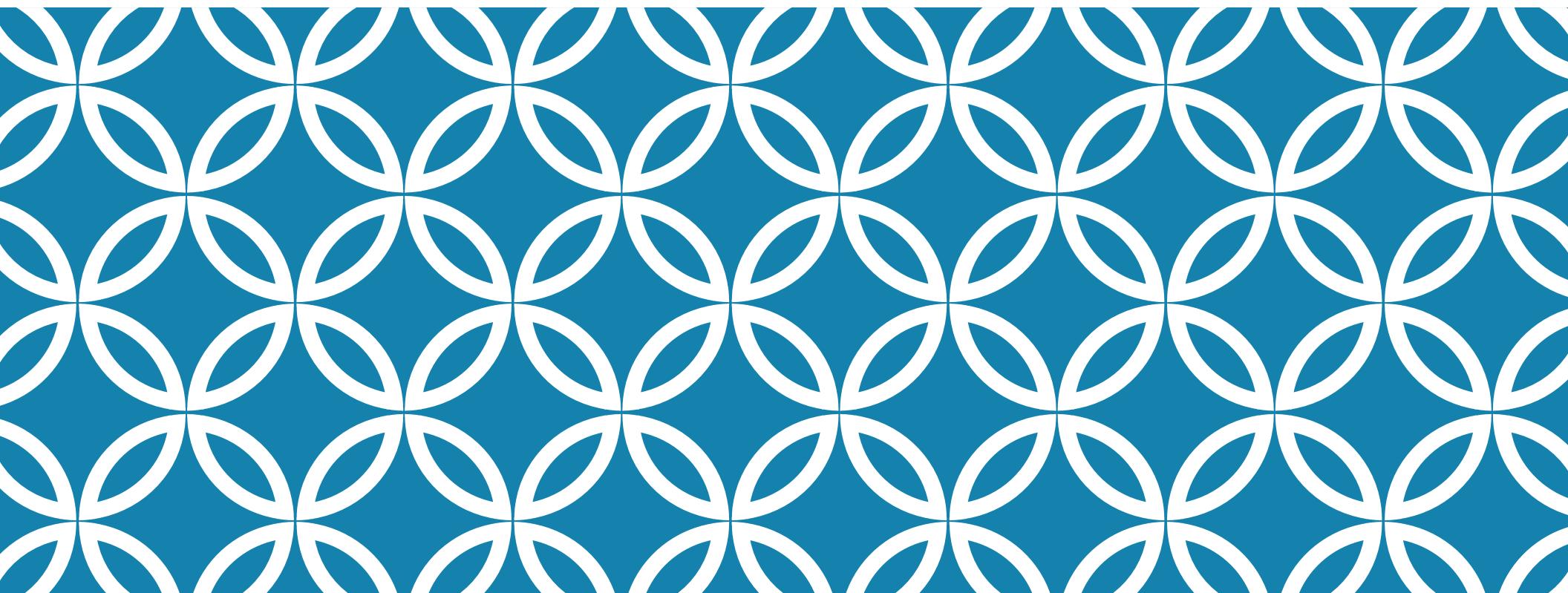
Yogesh S Rawat

Questions?

Contents

- PART-I: Deep Learning: A Short Overview
- **PART II: Computer Vision Employing Deep Learning**

PART II: Computer Vision Employing Deep Learning



**DEEP LEARNING FOR COMPUTER VISION
MUBARAK SHAH
CENTER FOR RESEARCH IN COMPUTER VISION (CRCV)**

shah@crcv.ucf.edu
<http://crcv.ucf.edu/>



CAP6412

Advanced Computer Vision

DEEP LEARNING | A SHORT OVERVIEW

Yogesh S Rawat

CLASSICAL COMPUTER VISION VS DEEP LEARNING

Classical Computer Vision

1. Hand Crafted Features
2. Encode Expert Knowledge into constraints
3. Convert constraints into Objective Function
4. Optimize Objective Function

Deep Learning

1. Motivated by human neural networks
2. Real Learning
3. Requires Labeled/Annotated Data
4. Requires Massively Parallel Computations (GPUs)
5. Use simple Stochastic Gradient Descend

MAIN THEMES

GAN: Generative- Adversarial Network

Reinforcement Learning

Transfer Learning/Domain Adaptation

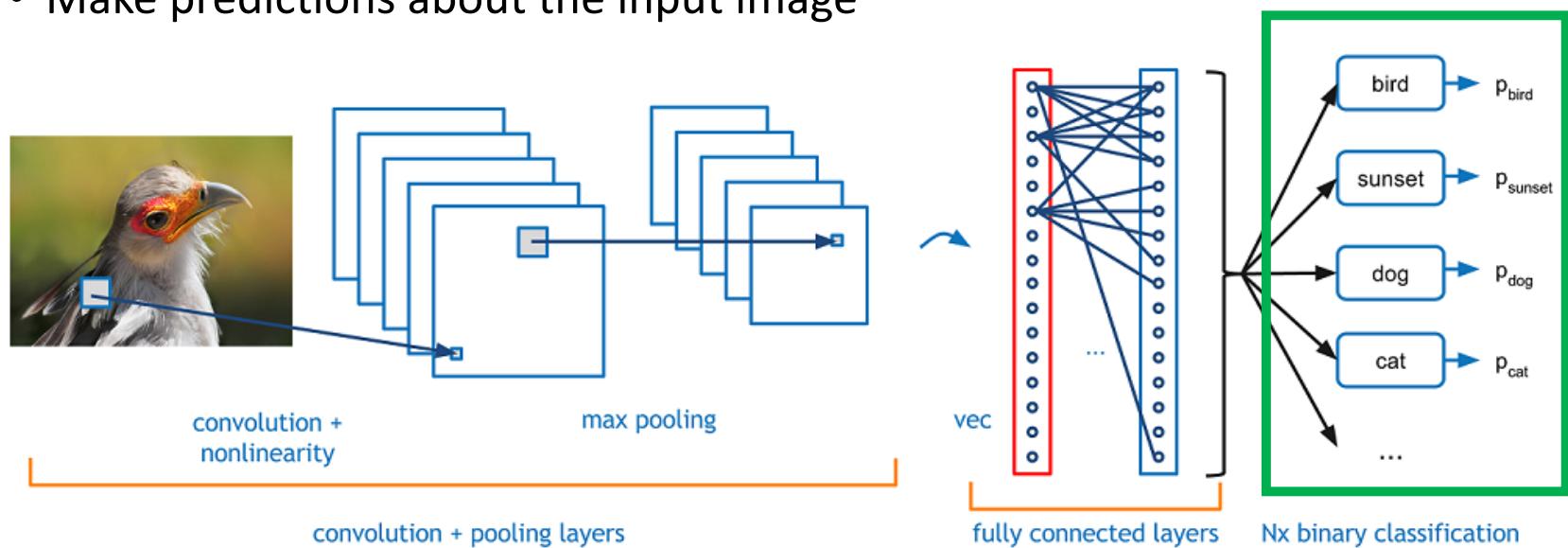
Multi-modal Analysis

End-to-End Learning

Bayesian Deep Learning

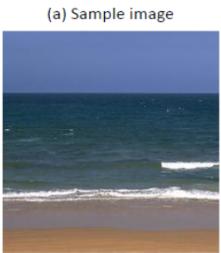
Deep Convolutional Neural Network (DCNN)

- A class of Neural Networks
 - Takes image as input
 - Make predictions about the input image



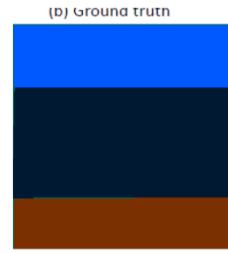
Source : <https://adeshpande3.github.io>

CONTENTS



(a) Sample image

road
sea
sky
sand



(b) Ground truth

Semantic Segmentation



Facial Attributes Detection



Human Re-Identification



Diving

Human Action Localization



Single Blank:

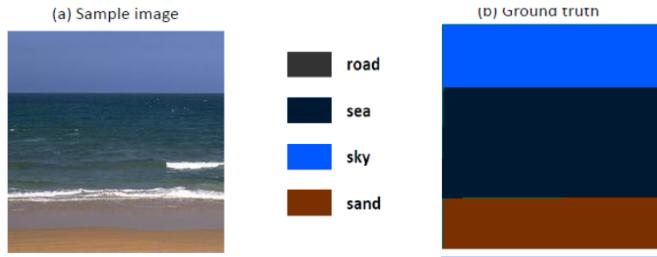
He ___ up the steps of the stand and away. (**Runs**)

Video Fill In The Blank

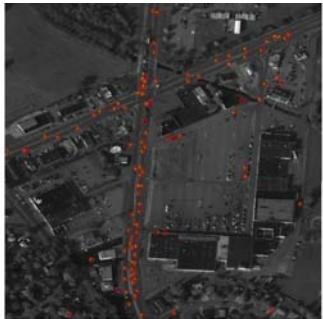


Reading The Mind

CONTENTS



Sematic Segmentation



Target Detection in WAMI



Facial Attributes Detection

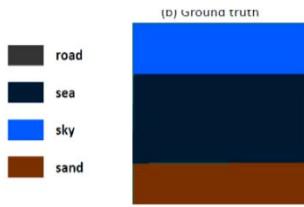


Human Re-Identification



Anomaly Detection

CONTENTS



Semantic Segmentation



Facial Attributes Detection



Human Re-Identification



Target Detection in WAMI



Anomaly Detection

Diving



Human Action Localization



Single Blank:

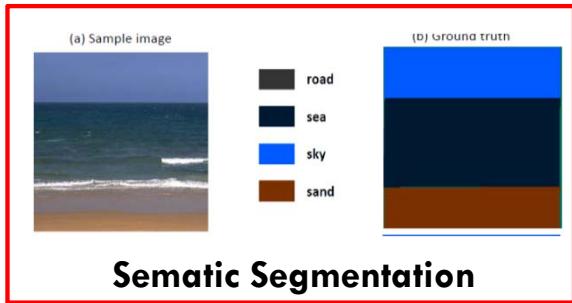
He ___ up the steps of the stand and away. (Runs)

Video Fill In The Blank



Reading The Mind

CONTENTS



Facial Attributes Detection



Target Detection in WAMI



Anomaly Detection



Human Action Localization



Video Fill In The Blank





**Center for Research
in Computer Vision**

UNIVERSITY OF CENTRAL FLORIDA

Semi Supervised Semantic Segmentation Using Generative Adversarial Network

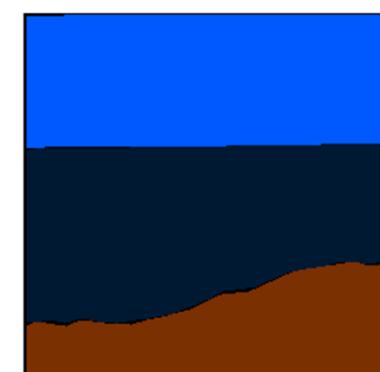
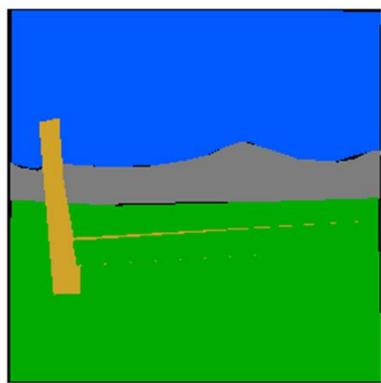
Nasim Souly, Concetto Spampinato and Mubarak Shah

ICCV 2017

http://crcv.ucf.edu/papers/iccv17/GAN_Semantic_cameraReady.pdf

SEMANTIC SEGMENTATION (SCENE LABELLING)

Assigning a semantic label to each pixel of an image.

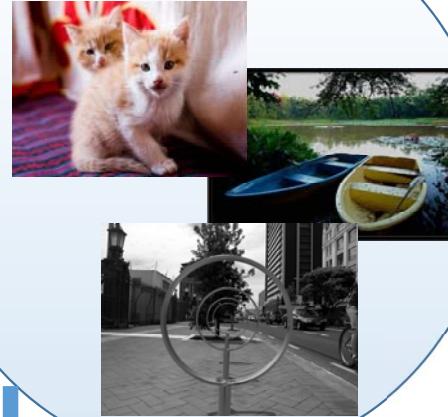


█	building
█	car
█	crosswalk
█	door
█	field
█	grass
█	mountain
█	person
█	plant
█	river
█	road
	rock
█	sand
█	sea
█	sidewalk
█	sky
█	tree
█	window

Motivation

- Lack of enough annotated data
- Plentiful unlabeled data
- Use generative model to improve classifiers

Labeled Data



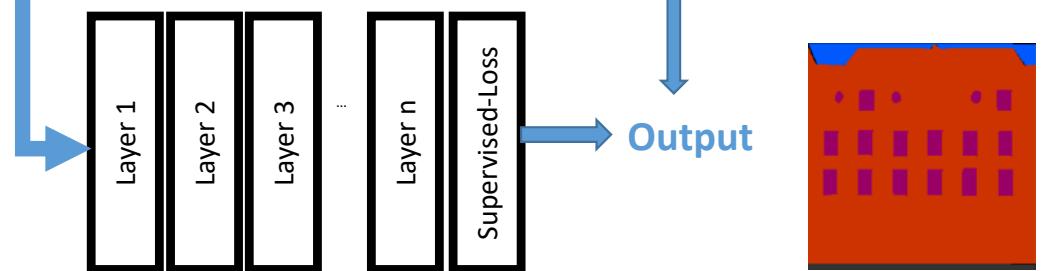
Unlabeled Data



Generated Data



Semi-supervised Loss



SEMI SUPERVISED LEARNING (SSL)

Halfway between supervised and unsupervised learning

Data points lying on the same feature manifold are more expected to be classified into the same class

Leverage the unlabeled data to find this structure.

Cost function for SSL

$$\textit{Loss} = \sum_{n=1}^{Nl} \textit{Loss}_l(y_n, x_n) + w \sum_{n=1}^{Nu} \textit{Loss}_u(x_n)$$

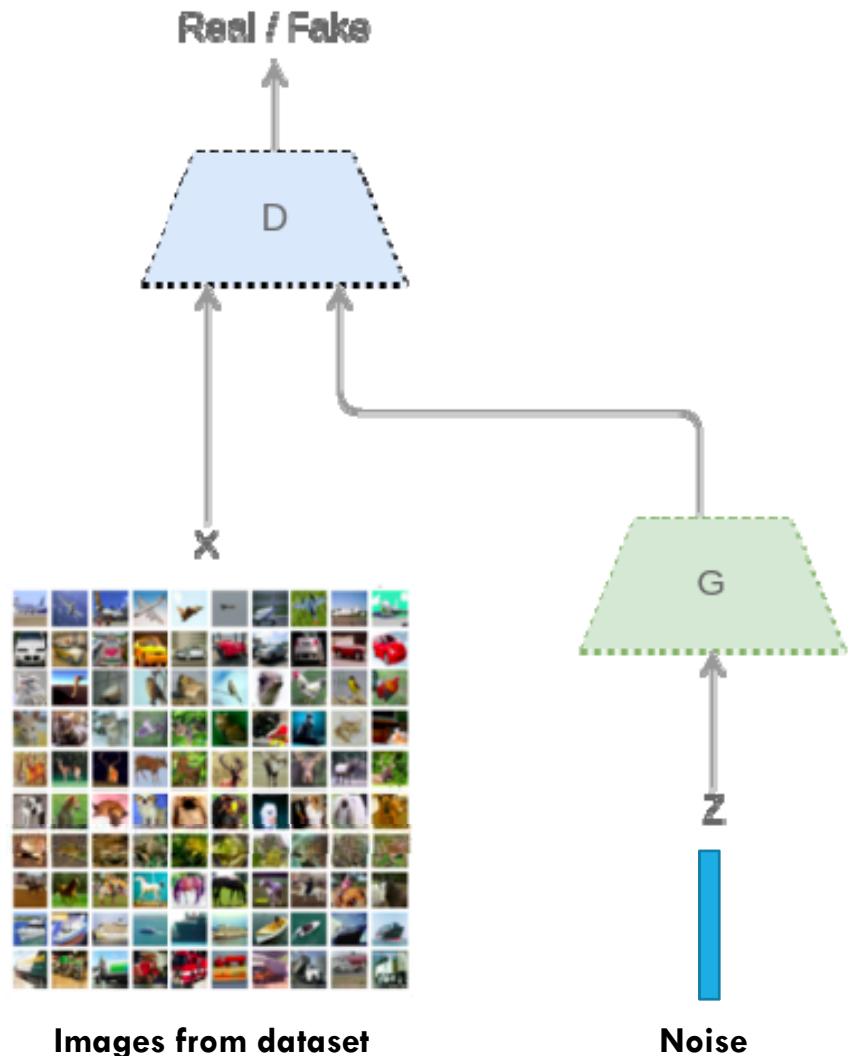
GENERATIVE ADVERSARIAL NETWORK

Enables models to tackle unsupervised learning

The intuitive idea:

- A painter who wants to do art forgery (G), (of Picasso)
- Someone is judging paintings (D)
- Then G produces paintings in an attempt to fool D
- D starts learning more about Picasso, G has a harder time fooling D
- D gets really good in telling apart what is Picasso and what is not?
- G gets really good at forging Picasso paintings

From Kdnuggets <http://www.kdnuggets.com>



GAN

Constant competition between two networks :

- a **generator** (G) and
- **discriminator** (D).

G starts from some noise, z , generate images $G(z)$.

D takes images from the distribution (real) and fake (from G) and classifies them: $D(x)$ and $D(G(z))$.

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log(D(x))] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$

SEMI SUPERVISED LEARNING USING GANS

Labels are not available for all training images,

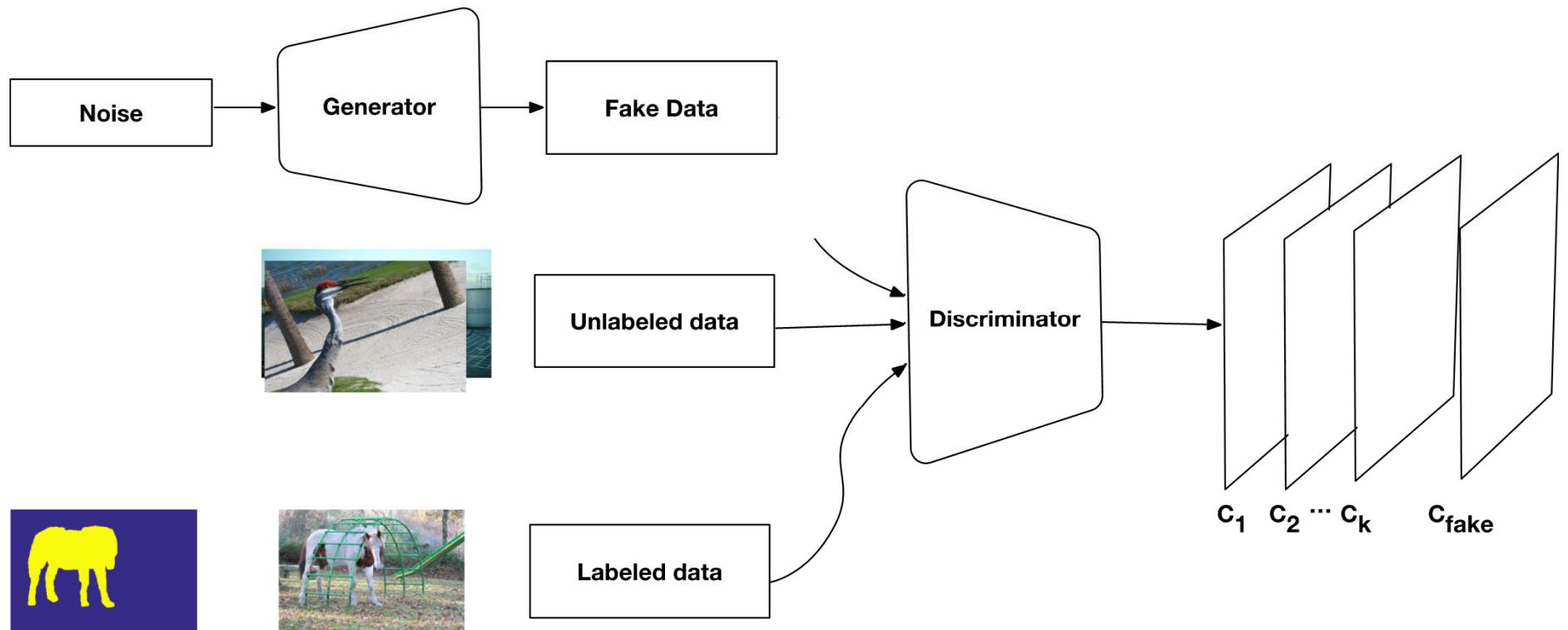
- leverage the unlabeled data by estimating a proper prior.

This prior is used by a classifier to improve.

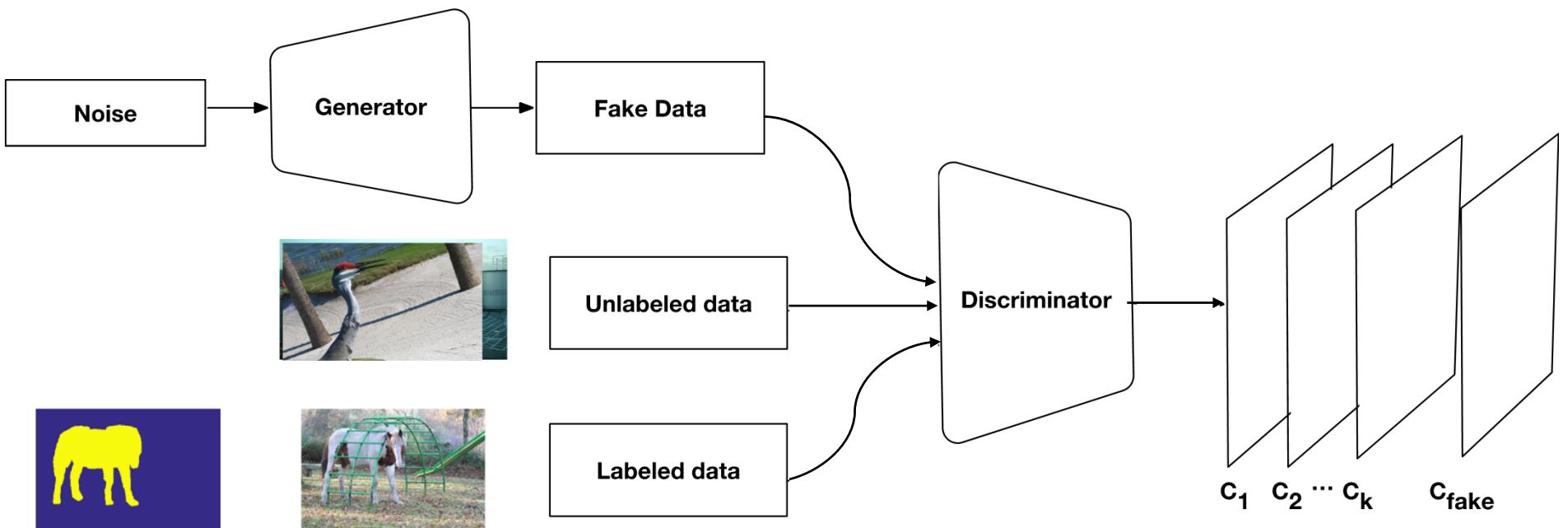
In GAN :

- Unlabeled data belongs to the same distribution of labeled data
- Generated (fake) data does not.

SEMI SUPERVISED LEARNING USING GANS

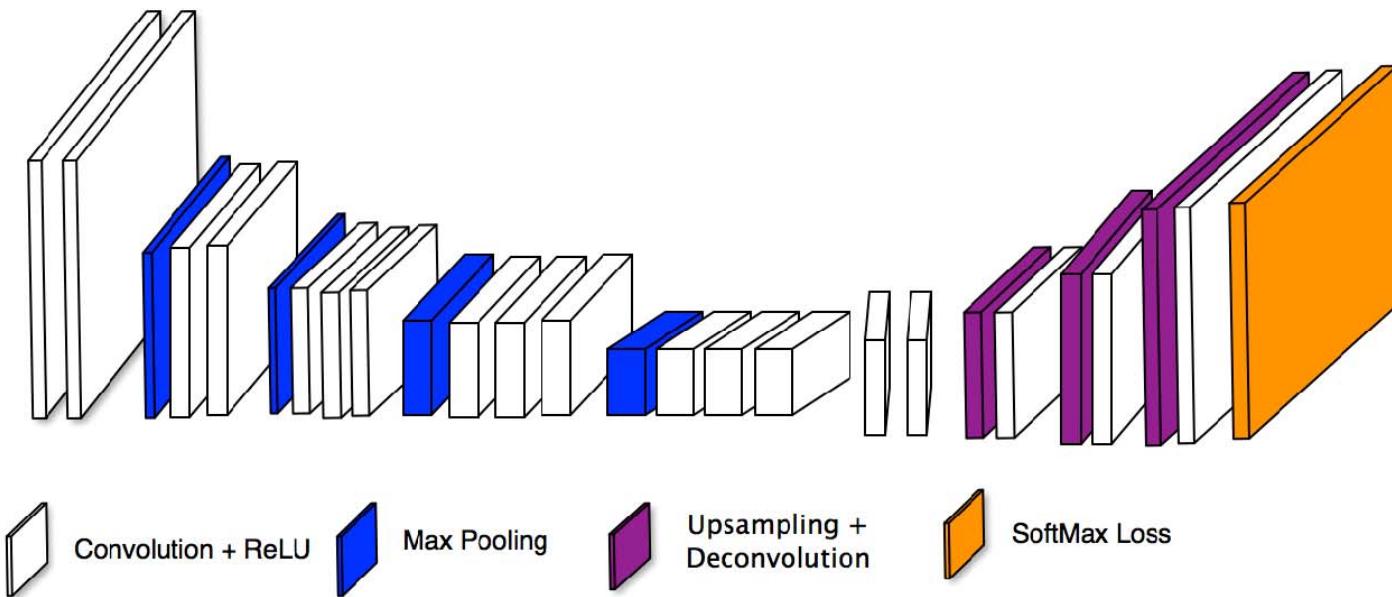


SEMI SUPERVISED LEARNING USING GANS



DISCRIMINATOR (CLASSIFIER) NETWORK

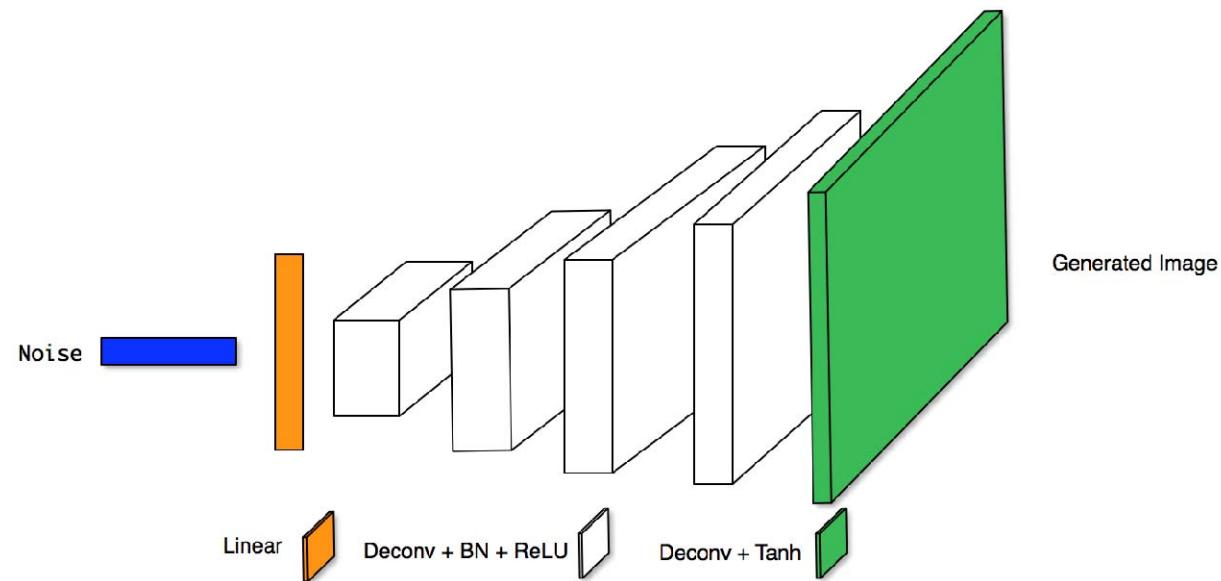
VGG 16 network with deconvolution layers.



GENERATOR NETWORK

Noise: 100D from uniform distribution.

Feature maps: 768, 384, 256, 192 and 3.



OPTIMIZATION: DISCRIMINATOR

Discriminator Loss:

- Unlabeled data
- Supervised Loss
- Fake images Loss

$$\mathcal{L}_D = -\mathbb{E}_{x \sim p_{data}(x)} \log(D(x)) + \gamma \mathbb{E}_{x, y \sim p(y, x)} [\text{CE}(y, P(y|x, D))] - \mathbb{E}_{z \sim p_z(z)} \log(1 - D(G(z)))$$

OPTIMIZATION: DISCRIMINATOR

Discriminator Loss :

- Unlabeled data
- Supervised Loss
- Fake (generated) images belonging to data distribution

$$\mathcal{L}_D = -\mathbb{E}_{x \sim p_{data}(x)} \log(D(x)) + \gamma \mathbb{E}_{x,y \sim p(y,x)} [\text{CE}(y, P(y|x, D))] - \mathbb{E}_{z \sim p_z(z)} \log(1 - D(G(z)))$$

$$D(x) = [1 - P(y = fake|x)]$$

$$P(y = k|x, D) = \frac{e^{D_k(x)}}{\sum^K_{i=1} e^{D_i(x)}}$$

OPTIMIZATION

Discriminator Loss : Supervised Loss + Unlabeled data Loss + Loss of fake generated belongs to data.

K classes - > K+1 classes (all classes plus fake)

Loss for Fake generated data - > $\min P(D(G(z)) \mid y \text{ in classes}) \Leftrightarrow \max P(D(G(z) \mid y = \text{fake class}^1)$

$$\begin{aligned}\mathcal{L}_D = & -\mathbb{E}_{x \sim p_{data}(x)} \log(D(x)) - \mathbb{E}_{z \sim p_z(z)} \log(1 - D(G(z))) \\ & + \gamma \mathbb{E}_{x,y \sim p(y,x)} [\text{CE}(y, P(y|x, D))],\end{aligned}$$

$$D(x) = [1 - P(y = \text{fake} | x)]$$

$$P(y = k | x, D) = \frac{e^{D_k(x)}}{\sum_{k=1}^K e^{D_k(x)}}$$

OPTIMIZATION: GENERATOR

Generator tries to generate samples close to real data

$$\min_G \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

OPTIMIZATION: GENERATOR

Generator tries to generate samples close to real data

Generator Loss = $\max P(D(G(z)) \mid y \text{ in classes})$

$$\min_G \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

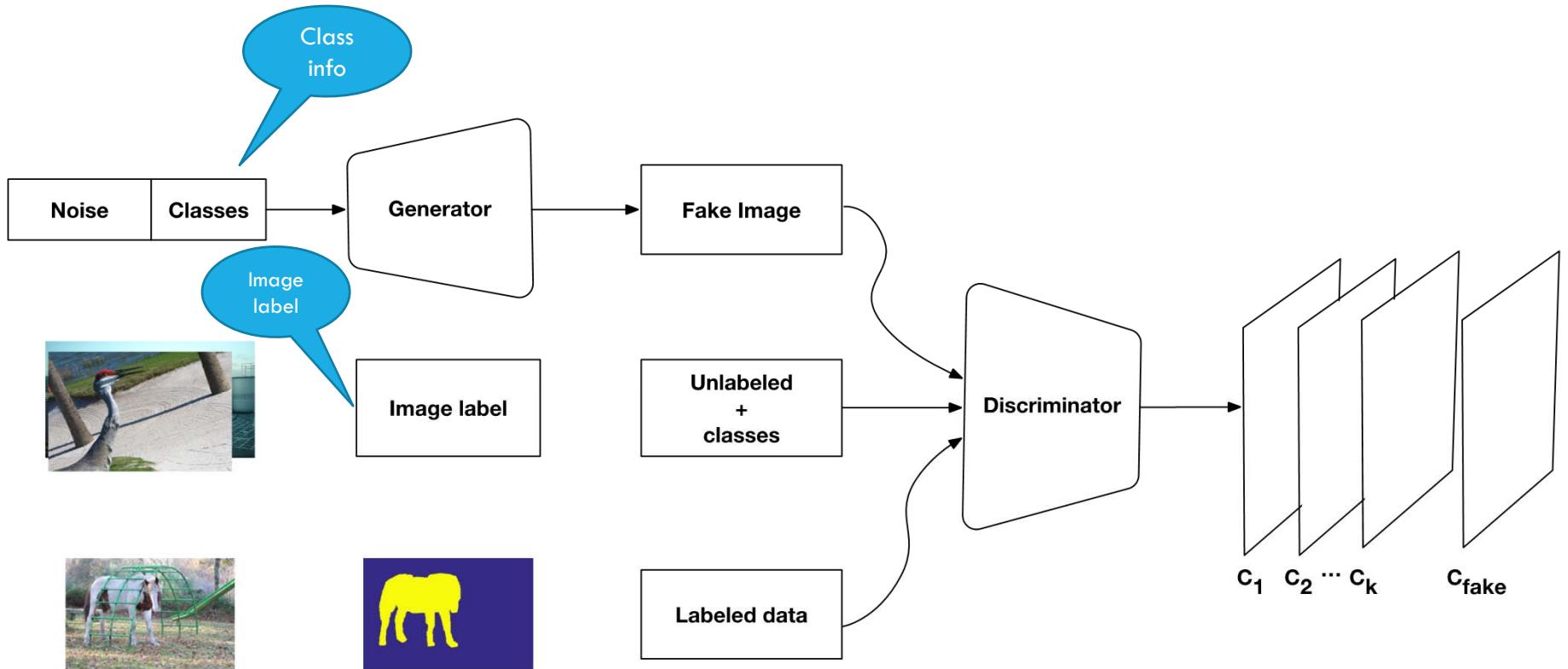
$$\max_G \mathbb{E}_{z \sim p_z(z)} [\log(D(G(z)))]$$

SEMI SUPERVISED LEARNING WITH WEAKLY LABELED DATA

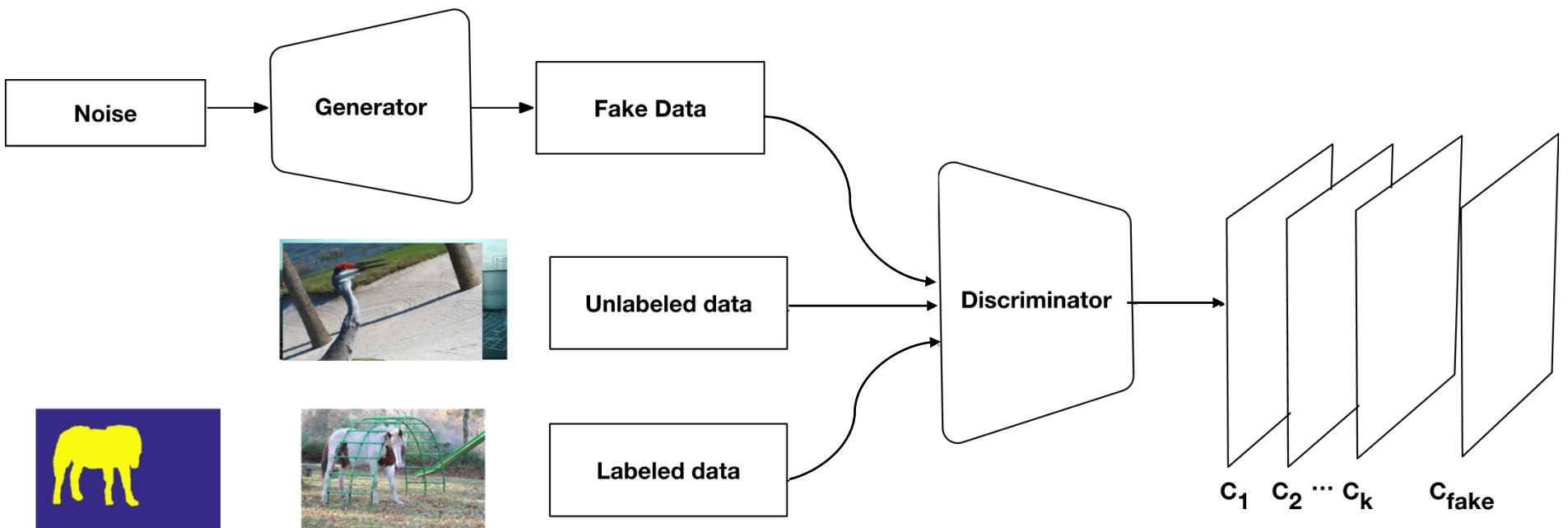
Discriminator learns class confidences and produces

- Confidence maps for each class
- A label for the fake data.

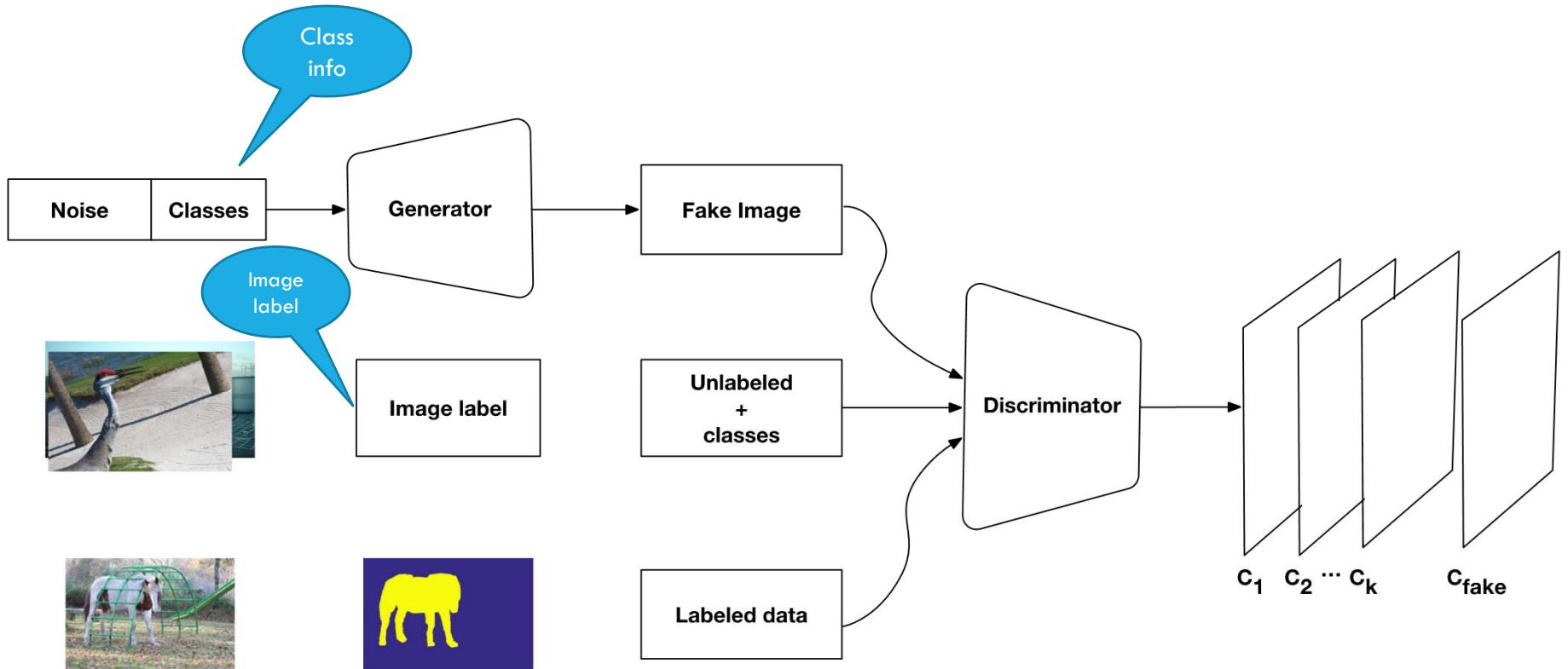
SEMI SUPERVISED LEARNING WITH ADDITIONAL WEAKLY LABELED DATA USING CONDITIONAL GANS



SEMI SUPERVISED LEARNING USING GANS



SEMI SUPERVISED LEARNING WITH ADDITIONAL WEAKLY LABELED DATA USING CONDITIONAL GANS



SEMI SUPERVISED LEARNING WITH **WEAKLY LABELED DATA**

The Generator uses

- Noise
- Class label information

The Discriminator uses

- Generated data
- Unlabeled data
- Image-level labels
- Pixel-level labeled data

OPTIMIZATION IN WEAKLY SUPERVISED

Using conditional GAN

In Discriminator the unlabeled part of loss is changed

$$\min_G \max_D V(D, G) = \mathbb{E}_{x,l \sim p_{data}(x,l)} [\log(D(x, l))] + \mathbb{E}_{z \sim p_z(z,l), l \sim p_l(l)} [\log(1 - D(G(z, l), l))]$$

$$\begin{aligned} \mathcal{L}_D = & -\mathbb{E}_{x,l \sim p_{z,l}(x,l)} \log[p(y = fake|x)] - \mathbb{E}_{x,l \sim p_{data(x,l)}} \log[p(y \in K_i \subset 1...K|x)] + \\ & \gamma \mathbb{E}_{x,y \sim p(y,x)} [\text{CE}(y, P(y|x, D))], \end{aligned}$$

OPTIMIZATION IN WEAKLY SUPERVISED

Using conditional GAN

In Discriminator the unlabeled part of loss is changed

Unlabeled Loss -> $\max P(D(x) \mid y \text{ in image level ground-truth classes})$

$$\min_G \max_D V(D, G) = \mathbb{E}_{x,l \sim p_{data}(x,l)} [\log(D(x, l))] + \mathbb{E}_{z \sim p_z(z,l), l \sim p_l(l)} [\log(1 - D(G(z, l), l))]$$

$$\begin{aligned} \mathcal{L}_D = & -\mathbb{E}_{x,l \sim p_{z,l}(x,l)} \log[p(y = fake|x)] - \mathbb{E}_{x,l \sim p_{data(x,l)}} \log[p(y \in K_i \subset 1...K|x)] + \\ & \gamma \mathbb{E}_{x,y \sim p(y,x)} [\text{CE}(y, P(y|x, D))], \end{aligned}$$

EXPERIMENTAL RESULTS:

We evaluated our method on

- PASCAL VOC 2012
- SiftFlow
- StanfordBG
- CamVid datasets.

Example: For Pascal dataset, we use all training data (1400 images) for which the pixel-level label are provided

10k additional images with image-level class labels

3 metrics : pixel accuracy , per-pixel classification accuracy and average of region intersection over union

EXPERIMENTAL RESULTS:

We evaluated our method on

- PASCAL VOC 2012
- SiftFlow
- StanfordBG
- CamVid datasets.

Example: For Pascal dataset, we use all training data (1400 images) for which the pixel-level label are provided

10k additional images with image-level class labels

3 metrics : pixel accuracy , per-pixel classification accuracy and average of region intersection over union

EXPERIMENTAL RESULTS

Datasets

- PASCAL VOC 2012
- SiftFlow
- StanfordBG
- CamVid datasets.

Evaluation metrics

- Pixel accuracy
- Per-pixel classification accuracy
- Average of region intersection over union

QUANTITATIVE RESULTS

StanfordBG

method	pixel accuracy	mean accuracy	mean IU
Standard [15]	73.3	66.5	51.3

• CamVid

method	pixel accuracy	mean accuracy	mean IU
SegNet(Basic) [1]	82.2	62.3	43.6

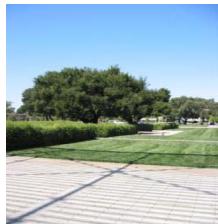
STANFORD BG

Image



Stanford BG

Image



Fully*Supervised

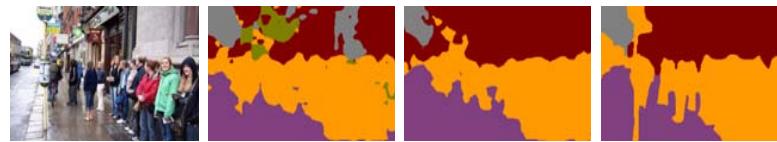


Semi*Supervised



GroundTruth5





QUANTITATIVE RESULTS: PASCAL VOC 2012

Using all fully labeled and unlabeled data in train set.

method	pixel accuracy	mean accuracy	mean IU
Fully supervised	90.3	75.9	62.2

Using 30% of fully labeled data and all unlabeled data in train set.

method	pixel accuracy	mean accuracy	mean IU
Fully supervised	83.15	53.1	38.9

QUANTITATIVE RESULTS: PASCAL VOC 2012

Using all fully labeled and unlabeled data in train set.

method	pixel accuracy	mean accuracy	mean IU
Fully supervised	90.3	75.9	62.2

Using 30% of fully labeled data and all unlabeled data in train set.

method	pixel accuracy	mean accuracy	mean IU
Fully supervised	83.15	53.1	38.9

QUANTITATIVE RESULTS: PASCAL VOC 2012

Using all fully labeled and unlabeled data in train set.

method	pixel accuracy	mean accuracy	mean IU
Fully supervised	90.3	75.9	62.2

Using 30% of fully labeled data and all unlabeled data in train set.

method	pixel accuracy	mean accuracy	mean IU
Fully supervised	83.15	53.1	38.9

QUALITATIVE RESULTS: VOC 2012

Image



- █ Human
- █ Bicycle
- █ MotorBike
- █ Bottle
- █ Cat
- █ Potted plant
- █ Aeroplane
- █ Bus

QUALITATIVE RESULTS: VOC 2012

Image



- █ Human
- █ Bicycle
- █ MotorBike
- █ Bottle
- █ Cat
- █ Potted plant
- █ Aeroplane
- █ Bus

QUANTITATIVE RESULTS: SIFTFLOW

Using fully labeled data and 2000 unlabeled images from SUN2012

method	pixel accuracy	mean accuracy	mean IU
Fully supervised	79.2	40.0	25.8

QUALITATIVE RESULTS: SIFTFLOW

Image

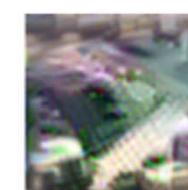
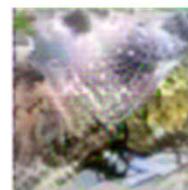
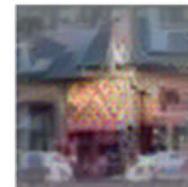
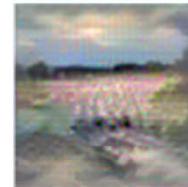
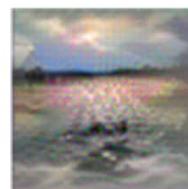
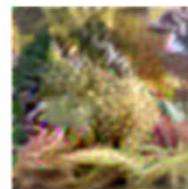
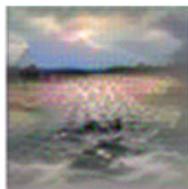


QUALITATIVE RESULTS: SIFTFLOW

Image



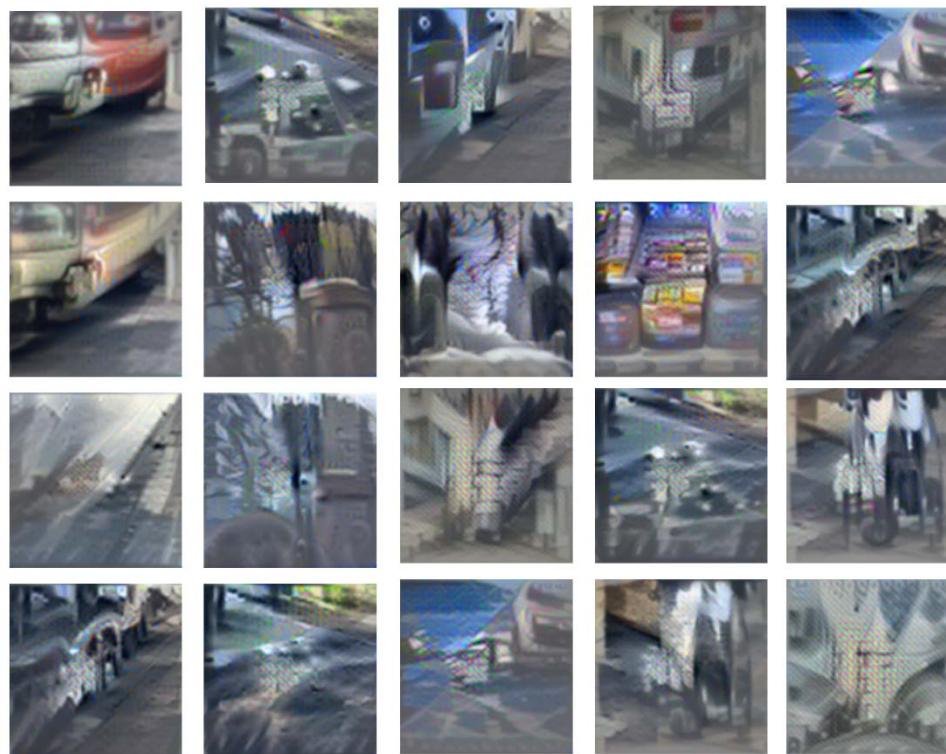
GENERATED IMAGES



GENERATED IMAGES SIFTFLOW

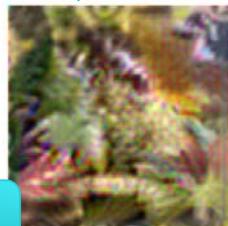
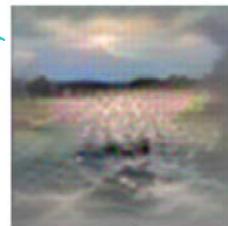


GENERATED IMAGES FROM CAMVID

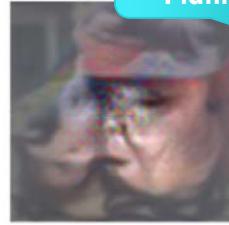
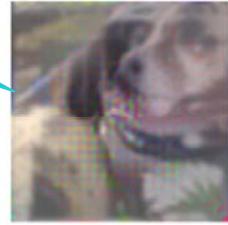


GENERATED IMAGES

Sky-Sea



Dog



Potted Plant



Forest



Car





**Center for Research
in Computer Vision**

UNIVERSITY OF CENTRAL FLORIDA

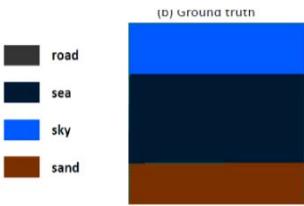
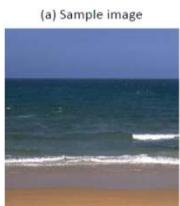
Semi Supervised Semantic Segmentation Using Generative Adversarial Network

Nasim Souly, Concetto Spampinato and Mubarak Shah

ICCV 2017

http://crcv.ucf.edu/papers/iccv17/GAN_Semantic_cameraReady.pdf

Contents



Sematic Segmentation



Human Re-Identification



Target Detection in WAMI



Anomaly Detection

Diving



Human Action Localization



Single Blank:

He ___ up the steps of the stand and away. (Runs)

Video Fill In The Blank



Reading The Mind

Improving Facial Attribute Prediction using Semantic Segmentation

Mahdi Kalayeh, Boqing Gong and Mubarak Shah

CVPR 2017

http://crcv.ucf.edu/papers/cvpr2017/Kalayeh_CVPR2017.pdf

Problem Definition

- Facial attribute prediction



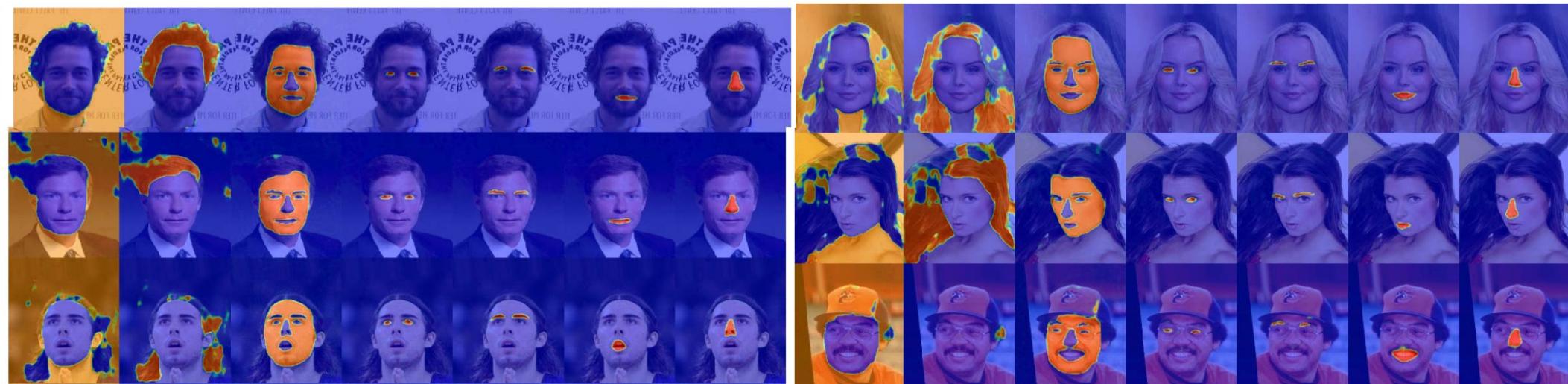
Improving Facial Attribute Prediction using Semantic
Segmentation (CVPR17)

40 Facial Attributes

5 o Clock Shadow	Mouth Slightly Open
Arched Eyebrows	Mustache
Attractive	Narrow Eyes
Bags Under Eyes	No Beard
Bald	Oval Face
Bangs	Pale Skin
Big Lips	Pointy Nose
Big Nose	Receding Hairline
Black Hair	Rosy Cheeks
Blond Hair	Sideburns
Blurry	Smiling
Brown Hair	Straight Hair
Bushy Eyebrows	Wavy Hair
Chubby	Wearing Earrings
Double Chin	Wearing Hat
Eyeglasses	Wearing Lipstick
Goatee	Wearing Necklace
Gray Hair	Wearing Necktie
Heavy Makeup	Young
High Cheekbones	
Male	

Proposed Idea

- Exploiting semantic face parsing



From left to right: background, hair, face, eyes, eyebrows, mouth and nose

Improving Facial Attribute Prediction using Semantic
Segmentation (CVPR17)

Overview

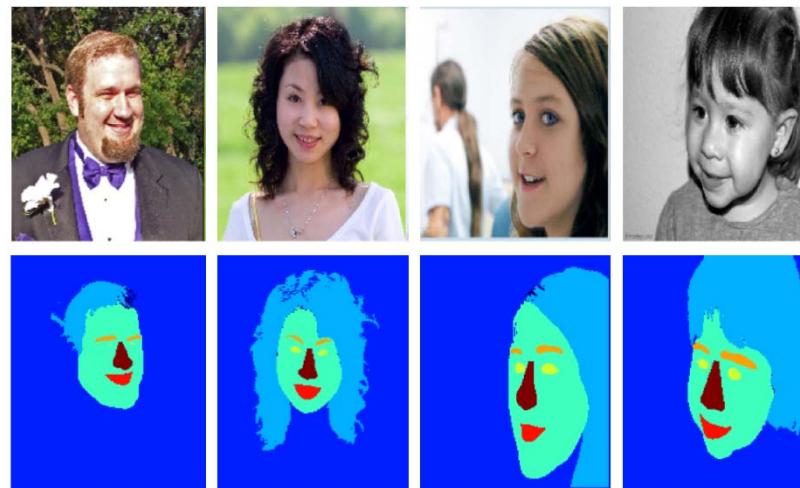
- Problem Definition: Attribute Prediction
- Current Approaches: Holistic v.s. Part-based
- Deep Learning:
 - Pass the entire image to CNN, predict all attributes jointly
 - Extract parts, extract CNN features from them, aggregate features, train multiple binary SVMs
- Part-based > Holistic

Proposed Idea

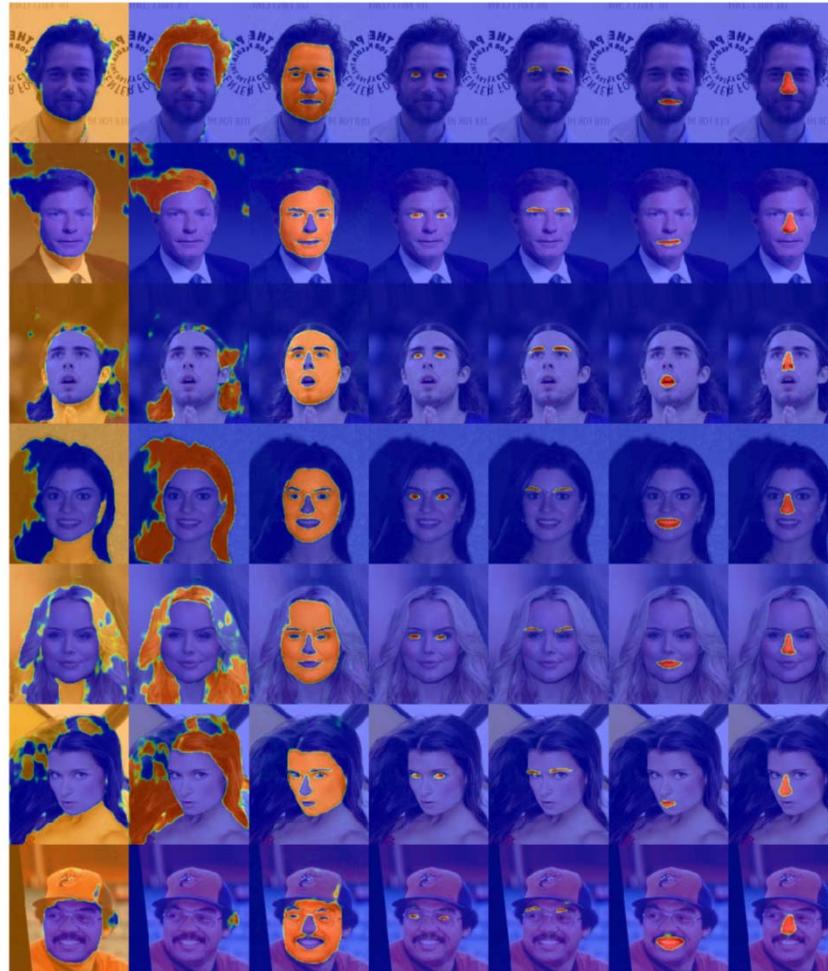
- Attributes are additive to the object
- Attributes do not appear in arbitrary regions
- Spatially decomposing objects into semantic regions
- Learning attributes in per-region fashion
- Output: attribute scores and where (spatially) they are inferred from

Semantic Segmentation Network

- An encoder-decoder de/convolutional network
- 16 layers deep architecture
- Using only 2K training data, with 7 semantic labels

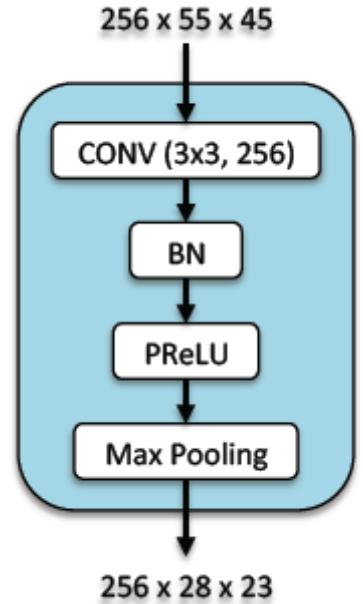


Semantic Segmentation Network: Results



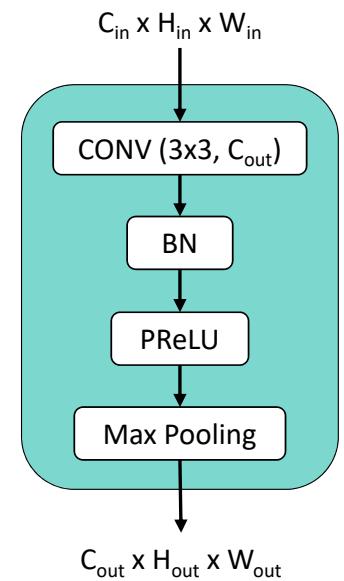
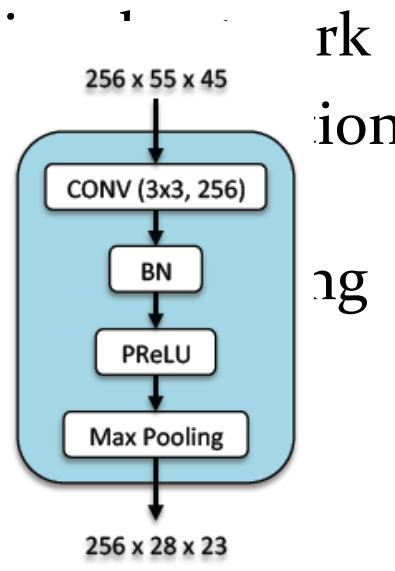
Basic Attribute Prediction Model

- 12 layers-deep fully convolutional network
- Blocks of convolution, batch normalization and non-linearity
- Reduce spatial resolution via max-pooling
- Global average pooling instead of FC



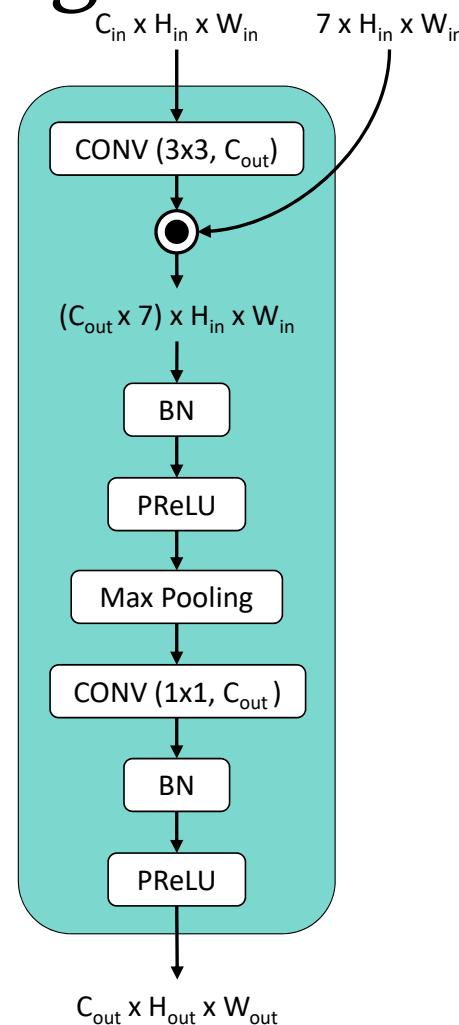
Basic Attribute Prediction Model

- 12 layers-deep fully convolutional network
- Blocks of convolution, batch normalization and non-linearity
- Reduce spatial resolution via max pooling
- Global average pooling instead of fully connected layer



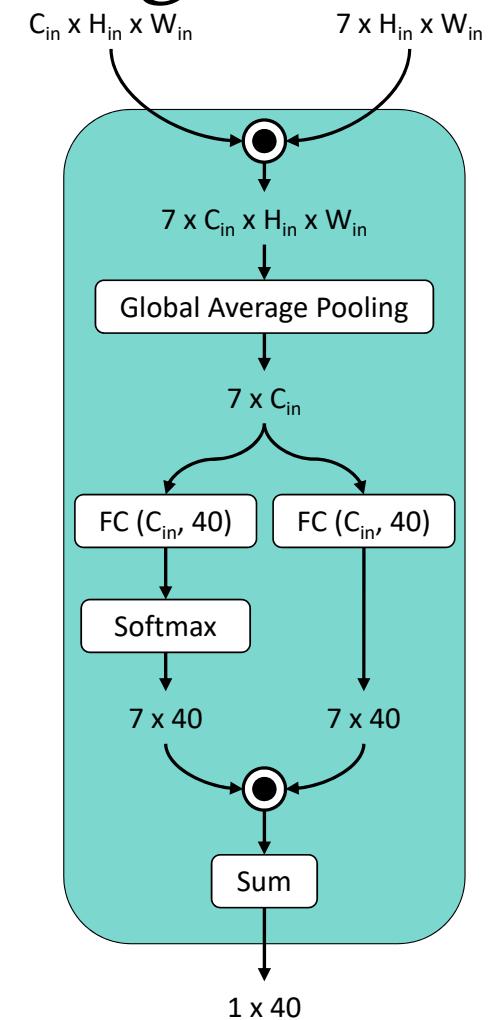
Semantic Segmentation-based Gating

- Max-pooling?
- Spatial aggregation? Single label vs. Multi-label
- Restrict pooling to within regions of the same semantic label



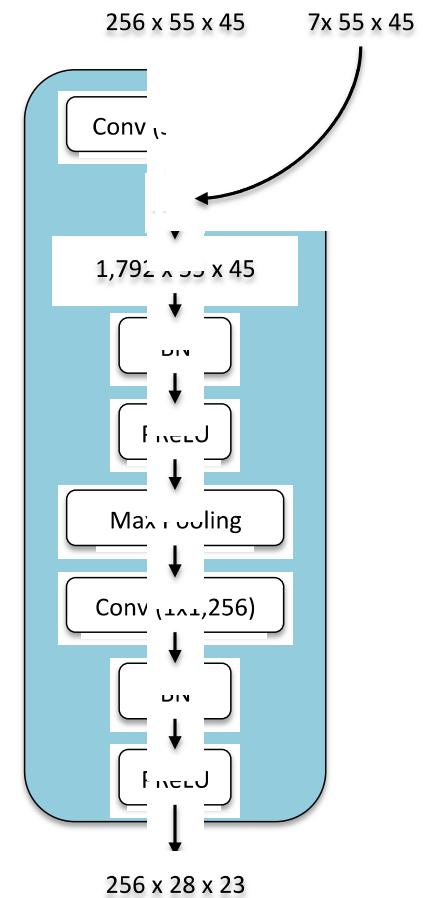
Semantic Segmentation-based Pooling

- Global Average Pooling: agnostic w.r.t spatial permutation
- Natural correspondence to the semantic regions
- Learn the correspondence
- Localization and Detection branches



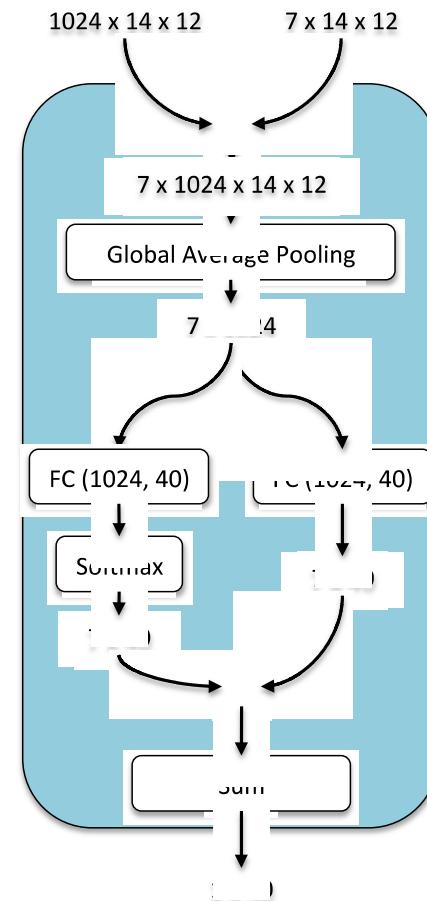
SSG: Semantic Segmentation-based Gating

- Problem with Max-pooling
- Spatial aggregation? Single label v.s. Multi-Label
- Restrict pooling to within regions of the same semantic label



SSP: Semantic Segmentation-based Pooling

- Global Average Pooling is spatially agnostic
- Attributes have natural correspondence to semantic regions
- We should learn the correspondence as well



Experiments

- Dataset: CelebA Facial Attribute Prediction
- 160K training, 20K validation, 20K testing
- 40 attributes



Experiments

Method	Classification Error%
FaceTracer [11]	18.88
PANDA [22]	15.00
Liu <i>et al.</i> [15]	12.70
Samangouei <i>et al.</i> [18]	10.50
Zhong <i>et al.</i> [23]	10.20
Avg. Pooling	9.36
SSP	8.98
SSG	9.13
SSP + SSG	8.84

Experiments

Method	Classification Error%
FaceTracer [11]	18.88
PANDA [22]	15.00
Liu <i>et al.</i> [15]	12.70
Samangouei <i>et al.</i> [18]	10.50
Zhong <i>et al.</i> [23]	10.20
Avg. Pooling	9.36
SSP	8.98
SSG	9.13
SSP + SSG	8.84

Table 2. Attribute prediction performance measured by the Classification Error% on CelebA [15] original image set. Note that FaceTracer and PANDA use groundtruth landmark points to attain face parts.

Method	Classification Error%
Rudd <i>et al.</i> [17]: Separate	9.78
Rudd <i>et al.</i> [17]: MOON	9.06
SPPNet*	9.49
Avg. Pooling	8.84
SSP	8.33
SSG	8.38
SSP + SSG	8.20

Table 3. Attribute prediction performance measured by the Classification Error% on CelebA [15] pre-cropped image set.

Method	Original (AP%)	Pre-cropped (AP%)
SPPNet*	N/A	77.69
Avg. Pooling	77.05	79.80
SSP	78.01	81.02
SSG	77.46	80.55
SSP + SSG	78.74	81.45

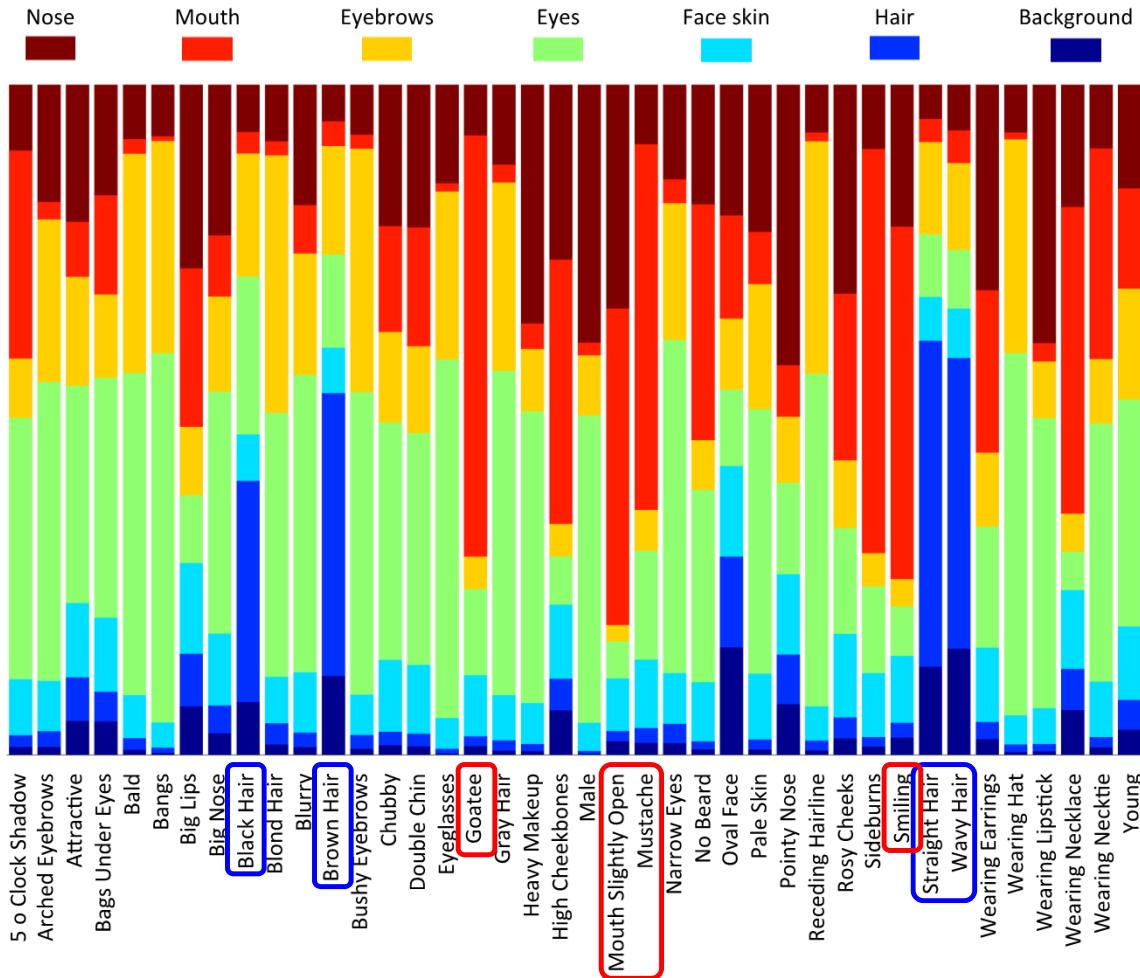
Table 4. Attribute prediction performance of our proposed variants measured by the Average Precision (AP)% on CelebA [15] original and pre-cropped image sets.

Experiments

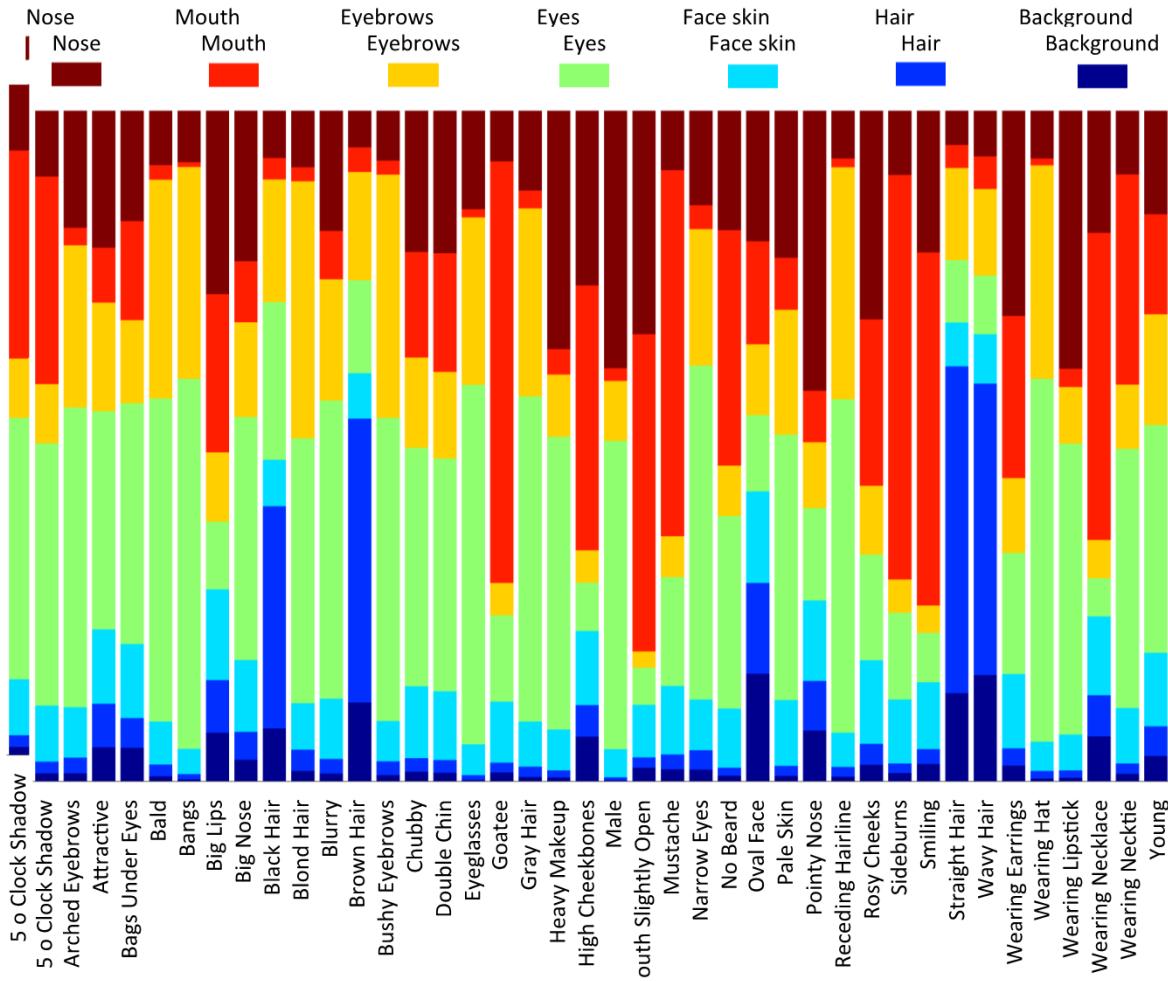
Method	Original (AP%)	Pre-cropped (AP%)
SPPNet*	N/A	77.69
Avg. Pooling	77.05	79.80
SSP	78.01	81.02
SSG	77.46	80.55
SSP + SSG	78.74	81.45

Table 4. Attribute prediction performance of our proposed variants measured by the Average Precision (AP)% on CelebA [15] original and pre-cropped image sets.

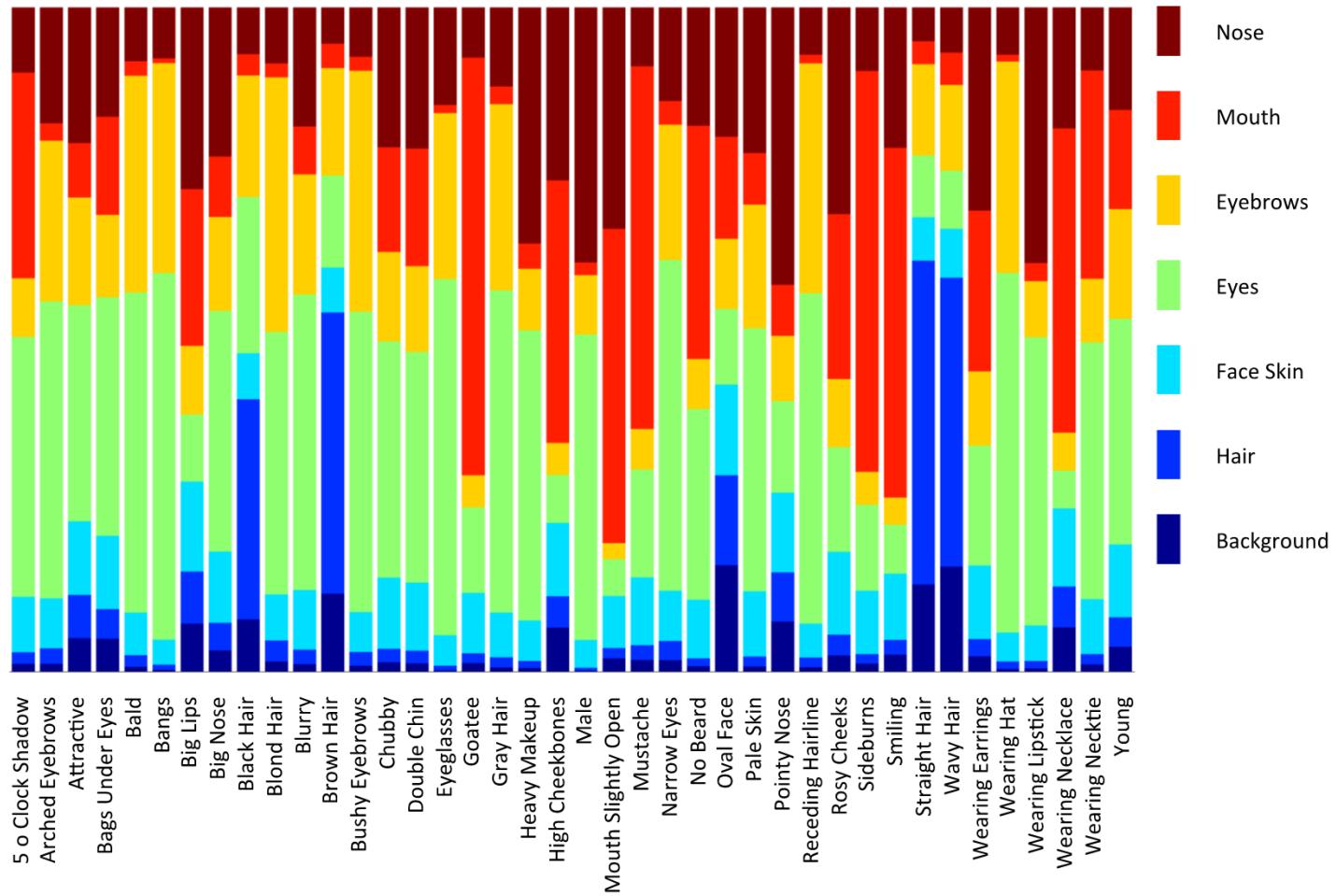
Learned Correspondences



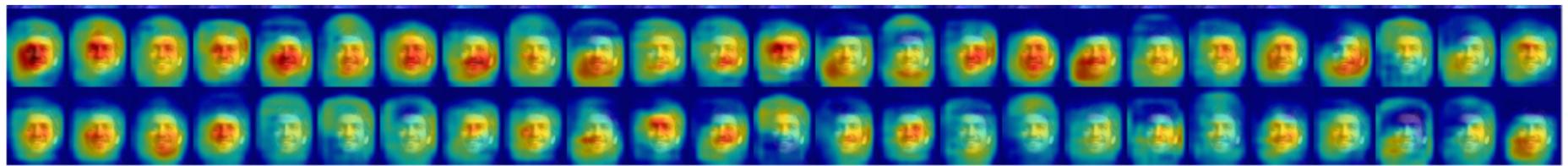
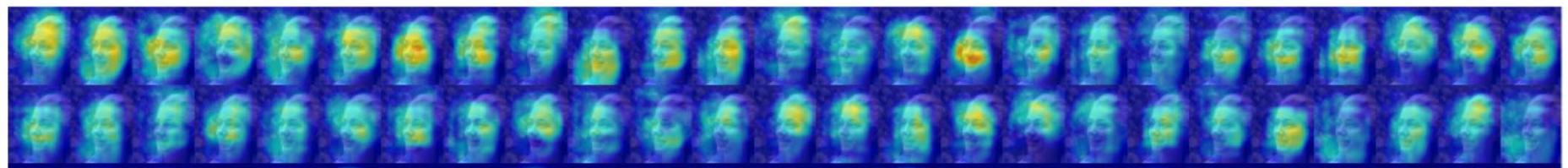
Learned Correspondences



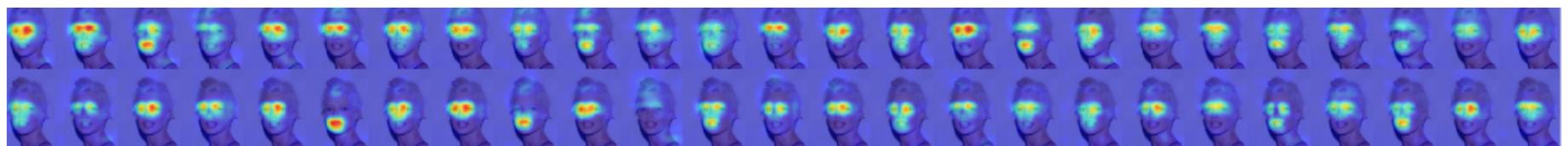
Learning the Correspondence



Visualizing the Activations: Global average pooling



Visualizing the Activation: SSP



Effect of Attributes on Semantic Face Parsing

- We used face parsing labels to improve facial attribute prediction
- What about the other way around?
 - Can attributes improve face parsing problem?

	Background	Hair	Face Skin	Eyes	Eyebrows	Mouth	Nose	Avg.
w/o attributes	89.25	47.56	78.65	46.83	31.22	62.03	77.40	61.84
w/ attributes	89.64	48.32	79.92	56.33	42.25	65.42	77.74	65.66

Table 2. Effect of attributes on facial parsing in IoU(%).

Qualitative Results



Attractive
Black hair
Brown hair
Male
Mouth slightly open
Smiling
Straight hair
Wearing lipstick
Young

Qualitative Results



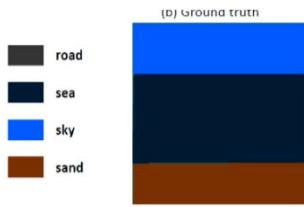
Attractive
Black hair
Brown hair
Male
Mouth slightly open
Smiling
Straight hair
Wearing lipstick
Young

Improving Facial Attribute Prediction using Semantic Segmentation

Mahdi Kalayeh, Boqing Gong and Mubarak Shah
CVPR 2017

http://crcv.ucf.edu/papers/cvpr2017/Kalayeh_CVPR2017.pdf

Contents



Semantic Segmentation



Facial Attributes Detection



Target Detection in WAMI



Anomaly Detection

Diving



Human Action Localization



Single Blank:

He ___ up the steps of the stand and away. (Runs)

Video Fill In The Blank



Reading The Mind

Human Semantic Parsing for Person Re-identification

**Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak
Mubarak Shah**

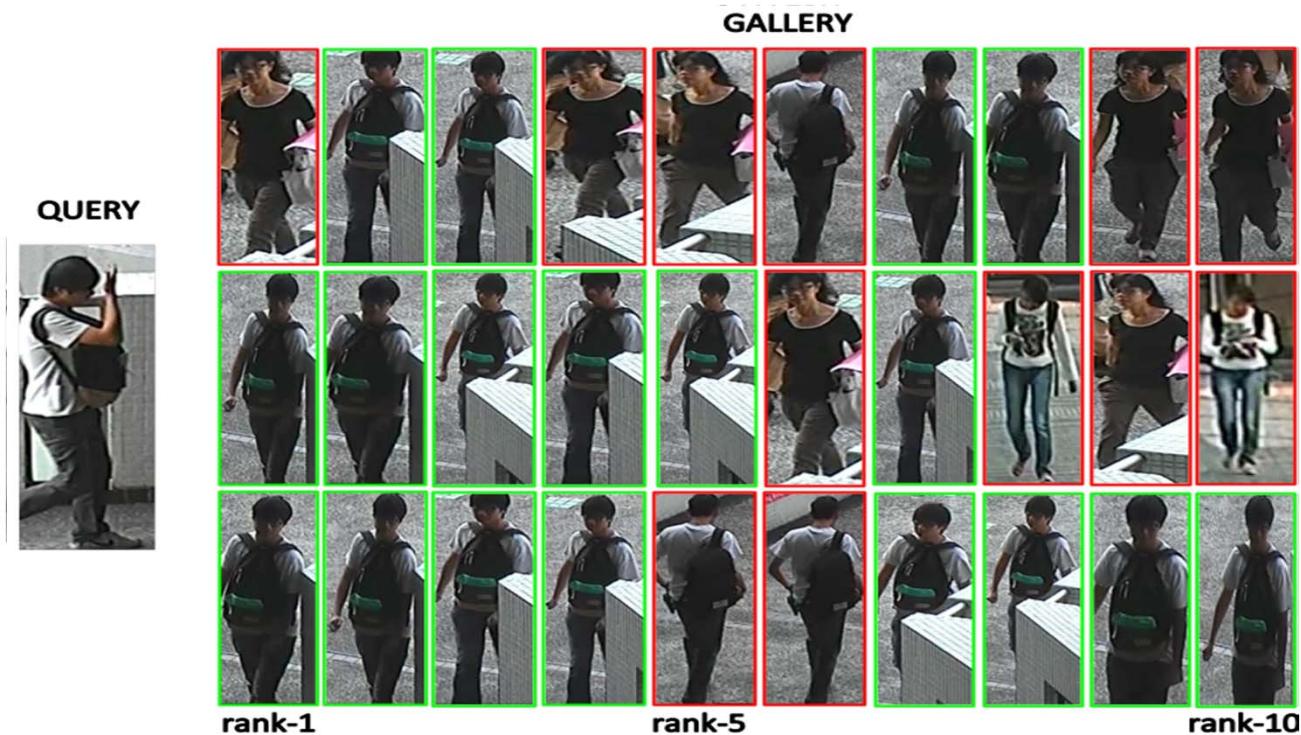
CVPR 2018

http://crcv.ucf.edu/people/phd_students/mahdi/papers/CVPR18.pdf

Person Re-Identification

- Retrieving the images of a person from a large gallery
- Query and gallery images are captured by different cameras
- A cross-camera data association problem

Person Re-Identification



Challenges

- Illumination conditions
- Observable human body parts
- Perceived posture of the person
- Background clutter
- Occlusion



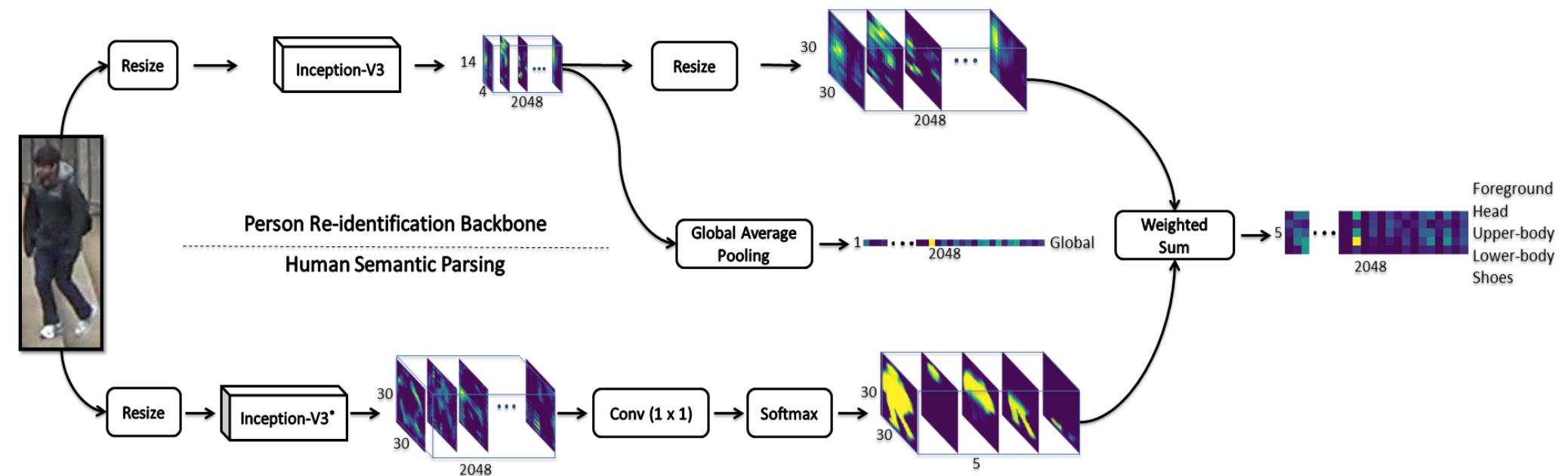
Extracting visual cues from human body parts

- Human pose estimation ✗
 - poor performance in low resolution images,
 - Unable to capture arbitrary contours of human body parts
- Dividing image into horizontal stripes ✗
 - Easy to implement but lacking any semantic alignment

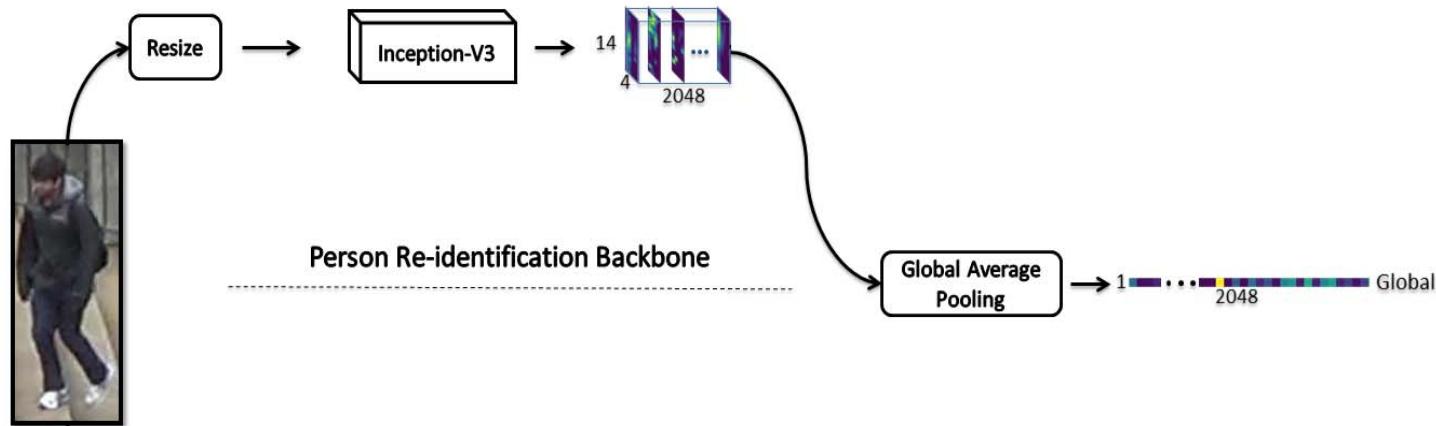
Proposed Approach

- A simple yet effective training strategy
 - Training on aggregation of multiple datasets in low resolution (492x164)
 - Finetuning on each individual dataset in high resolution (748x246)
 - Effective use of three architectures: Inception-V3, ResNet50 and ResNet152
- Human semantic parsing
 - Addressing aforementioned challenges

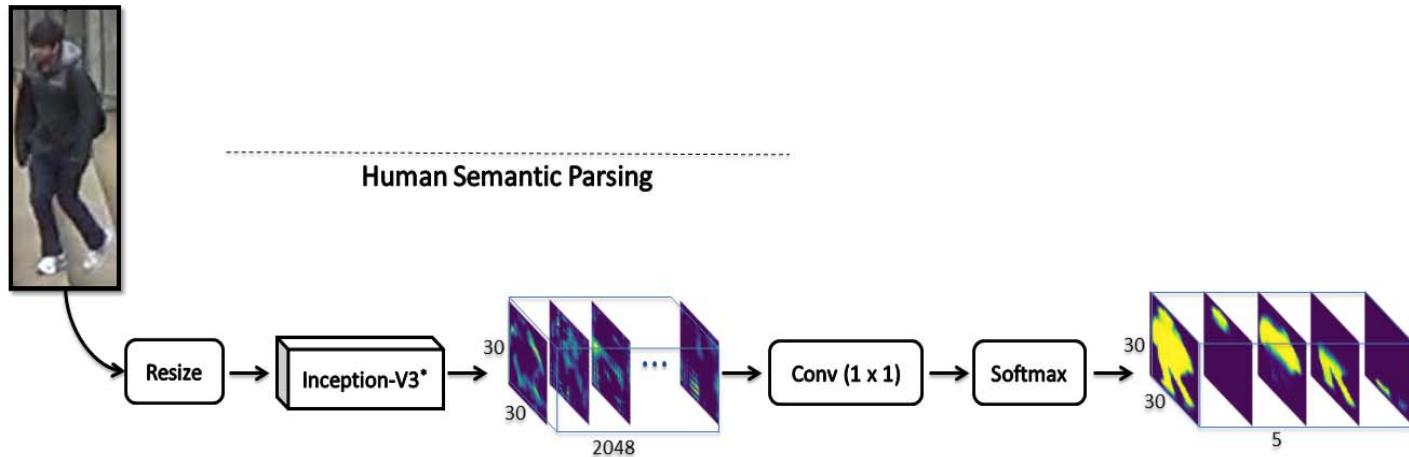
SPReID: Human Semantic Parsing for Person Re-identification



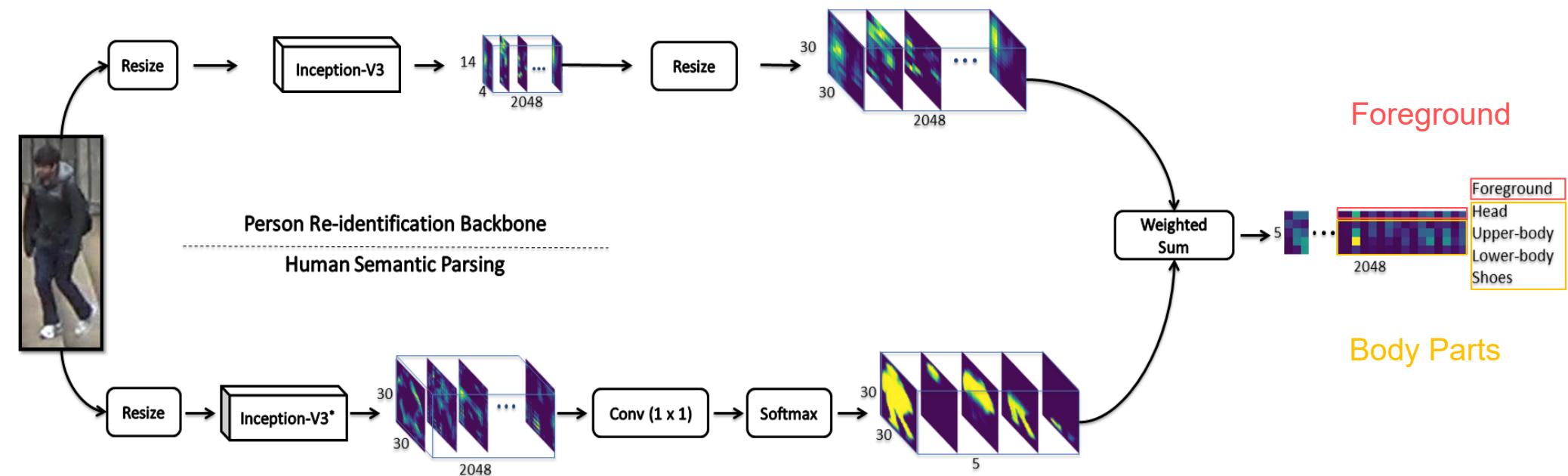
SPReID - Global Representation



SPReID - Human Semantic Parsing

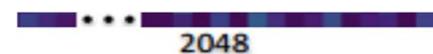
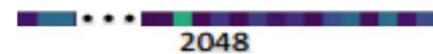


SPReID – Foreground & Region Representations



SPReID Representation

- Concatenation of
 - Global representation
 - Foreground representation
 - Body Part representation



Experimental Results

- **Datasets**
 - **Person Re-Identification**
 - CUHK03
 - Market-1501
 - DukeMTMC-reID
 - **Semantic parsing**
 - Look into Person (LIP)

CUHK03

- **1,360** identities captured by **6** cameras.
 - Each identity is viewed by **2** disjoint cameras.
- **13,164** person images
 - Each identity has on average **4.8** images in each viewpoint.
- The dataset partitions:
 - Training: **1160** persons
 - Validation: **100** persons
 - Test: **100** persons



Market-1501

- **32,668** labeled bounding boxes of **1501** subjects captured by **6** cameras
 - The bounding boxes are detected DPM
- The dataset partitions:
 - Training: **751** persons, **12936** images
 - Query: **750** persons, **3368** images
 - Gallery: **750** persons, **19734** images



DukeMTMC-reID

- Consists of the images extracted from the **DukeMTMC** tracking dataset.
 - Recorded by 8 cameras
 - Hand-annotated bounding boxes
- The dataset partitions:
 - Training: **702** persons, **16,522** images
 - Query: **702** identities, **2,228** images
 - Gallery: **702** identities, **16,522** images



Look into Person (LIP)

- 50,462 pixel-wise annotated images with **19** semantic human part labels and one background label.



■ Face ■ UpperClothes ■ Hair ■ RightArm ■ Pants ■ LeftArm ■ RightShoe ■ LeftShoe ■ Hat ■ Coat ■ RightLeg ■ LeftLeg ■ Gloves ■ Socks ■ Sunglasses ■ Dress ■ Skirt ■ Jumpsuits ■ Scarf

Training the Baselines

- **Aggregation of 10 datasets**
 - CUHK03, Market-1501, DukeMTMC-reID, 3DPeS, CUHK01, CUHK02, PRID, PSDB, Shinpuhkan, VIPeR
 - ~**111,000** images of ~**17,000** identities
- **Full images without semantic segmentation**
 - Training for **200K** iterations using input images of size **492×164**
- **Fine-tuning on evaluation datasets separately**
 - For **50K** iteration on higher input resolution of **748×246**

Training SPReID

- **Person Re-Id Backbone**
 - Training is done with the exact same setting as Baseline
- **Human semantic parsing**
 - Training on Look into Person (LIP) dataset
 - Different semantic regions are grouped to create 5 coarse labels
 - foreground, head, upper-body, lower-body and shoes

Human semantic parsing



CUHK03



InceptionV3-ft

SPReID-ft

InceptionV3-ft

SPReID-ft



InceptionV3-ft

SPReID-ft

InceptionV3-ft

SPReID-ft

Market-1501



InceptionV3-ft

SPReID-ft

InceptionV3-ft

SPReID-ft



InceptionV3-ft

SPReID-ft

InceptionV3-ft

SPReID-ft

DukeMTMC-reID



Experimental Results – LIP (semantic parsing)

State-of-the-art

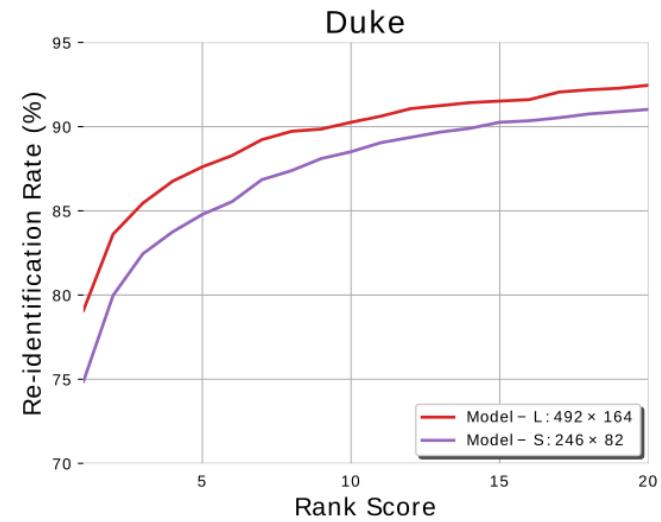
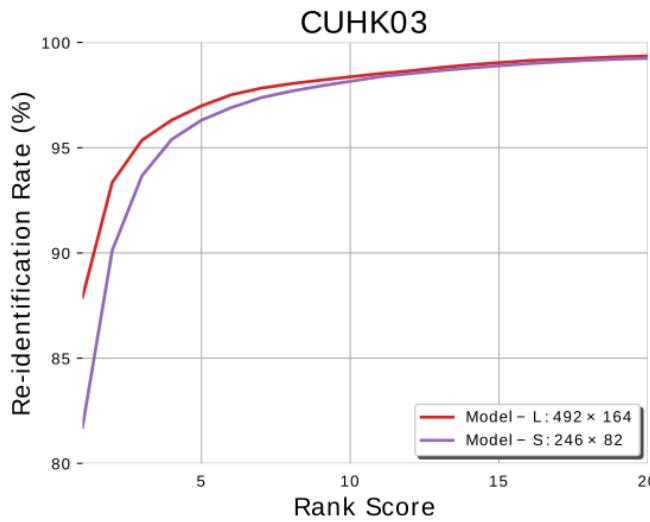
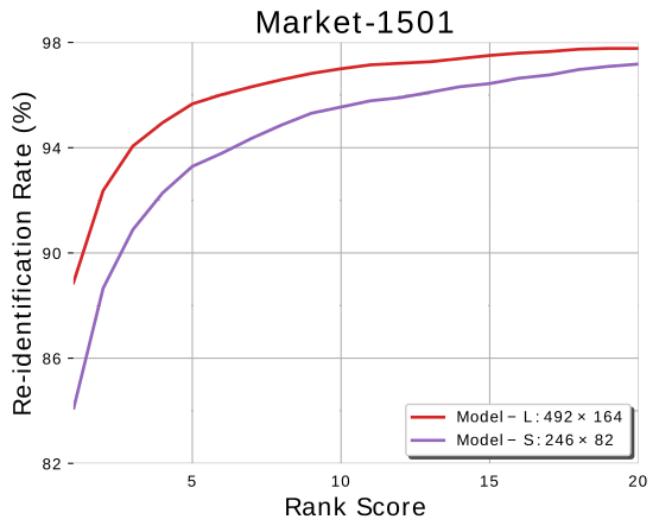
Ours

Method	Overall Accuracy	Mean Accuracy	Mean IoU
SegNet	69.04	24.0	18.17
FCN-8n	76.06	36.75	28.29
DeepLabV2	82.66	51.64	41.64
Attention	83.43	54.39	42.92
DeepLabV2+SSL	83.16	52.55	42.44
Attention+SSL	84.36	54.94	44.73
Ours	85.07	60.54	48.16

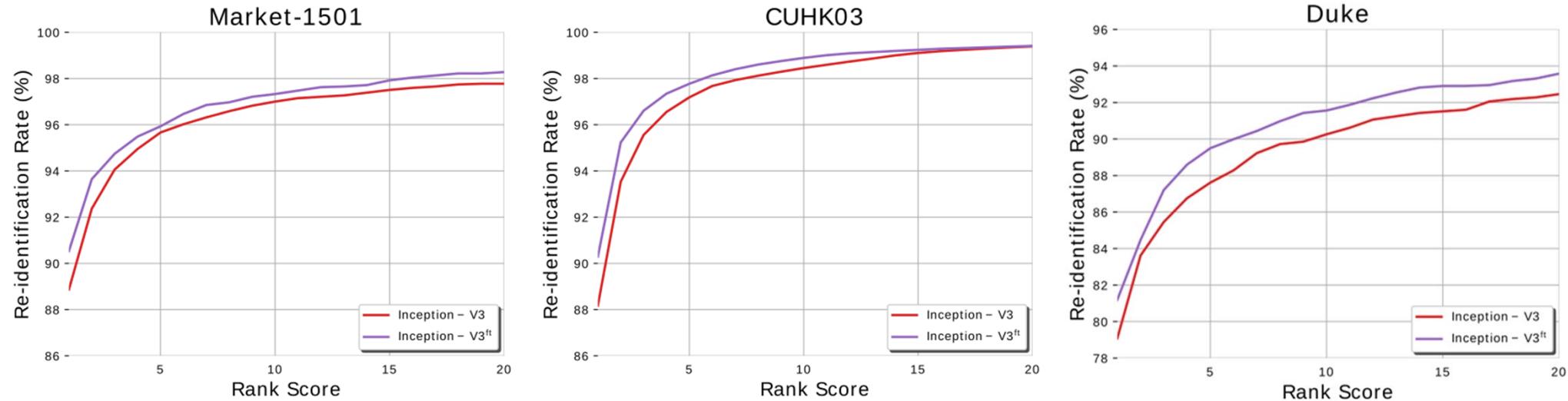
Experimental Results: Person Re-ID



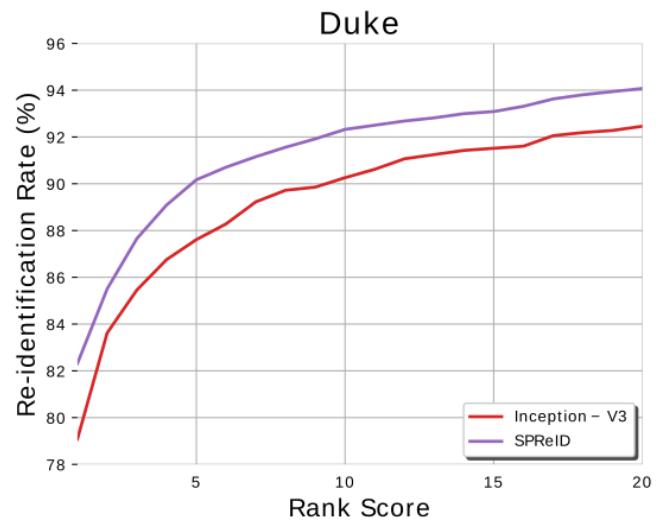
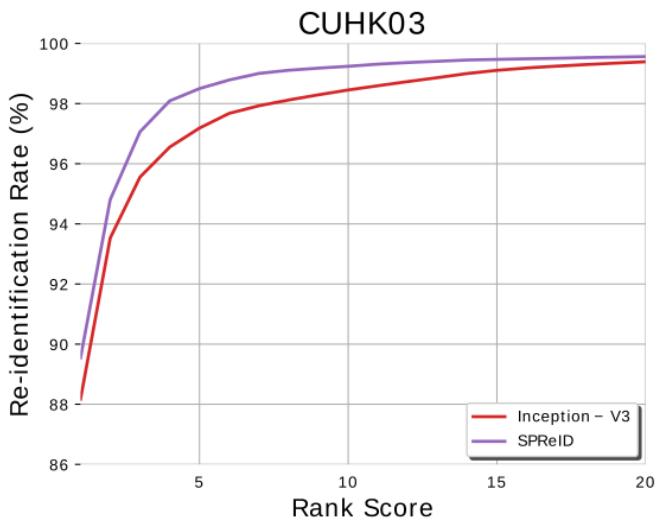
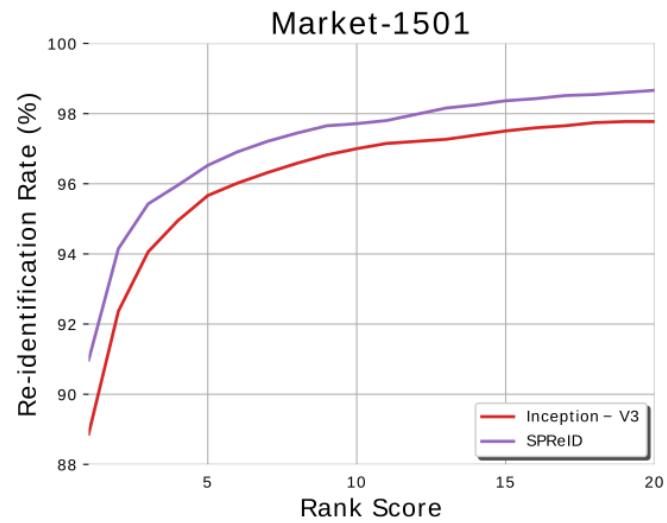
Effect of Input Image Size



- **Effect of Fine Tuning for CNN Backbone**



Effect of Semantic Parsing



Results – Market-1501

State-of-the-art

Ours - without
semantic parsing

Ours - with
semantic parsing

Method	mAP (%)	rank-1	rank-5	rank-10
DPAR	63.4	81.0	92.0	94.7
JMLL	65.5	85.1	-	-
Basel.+LSRO	66.1	84.0	-	-
SSM	68.8	82.2	-	-
DaF	72.4	82.3	-	-
Chen et. Al.	73.1	88.9	-	-
Inception-V3 ^{ft}	76.56	90.8	96.35	97.71
ResNet-152 ^{ft}	77.96	90.71	96.26	97.65
combined	82.87	93.14	97.27	98.22
SPReID ^{ft}	81.34	92.54	97.15	98.1
+ re-ranking	89.99	94.3	96.35	97.39
ResNet-152 ^{ft} + SPReID ^{ft}	83.36	93.68	97.57	98.4
+ re-ranking	90.96	94.63	96.82	97.65

Results – CUHK03

State-of-the-art

Ours - without
semantic parsing

Ours - with
semantic parsing

Method	mAP (%)	rank-1	rank-5	rank-10
SSM	-	76.6	94.6	98.0
Spindle	-	88.5	97.8	98.6
DPAR	-	85.4	97.6	99.4
Chen et. Al.	82.8	86.7	-	-
HydraPlus	-	91.8	98.4	99.1
Inception-V3 ^{ft}	-	88.73	97.82	98.94
ResNet-152 ^{ft}	-	90.38	98.71	99.46
combined	-	92.81	98.9	99.35
SPReID ^{ft}	-	93.89	98.76	99.51
+ re-ranking	-	96.31	99.25	99.71
ResNet-152 ^{ft} + SPReID ^{ft}	-	94.28	99.04	99.56
+ re-ranking	-	96.22	99.34	99.7

Results – Duke

State-of-the-art

Ours - without
semantic parsing

Ours - with
semantic parsing

Method	mAP (%)	rank-1	rank-5	rank-10
Basel. + LSRO	47.1	67.7	-	-
Basel. + OIM	-	68.1	-	-
Zheng et. Al.	49.3	68.9	-	-
ACRN	52.0	72.6	84.8	88.9
SVDNet	56.8	76.7	86.4	89.9
Chen et. Al.	60.6	79.2	-	-
Inception-V3 ^{ft}	63.27	80.48	88.78	91.65
ResNet-152 ^{ft}	67.02	83.26	90.93	92.95
combined	72.0	85.37	92.15	94.21
SPReID ^{ft}	70.97	84.43	91.88	93.72
+ re-ranking	83.16	87.21	92.37	93.9
ResNet-152 ^{ft} + SPReID ^{ft}	73.34	85.95	92.95	94.52
+ re-ranking	84.99	88.96	93.27	94.75

Conclusion

- A simple DCNN can outperform the current state-of-the-art
 - when trained properly on large number of images
- By exploiting human semantic parsing, the performance of a baseline model can be further improved



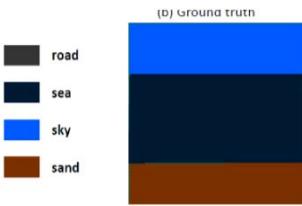
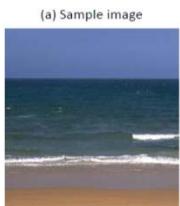
Human Semantic Parsing for Person Re-identification

**Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E. Kamasak
Mubarak Shah**

CVPR 2018

http://crcv.ucf.edu/people/phd_students/mahdi/papers/CVPR18.pdf

Contents



Sematic Segmentation



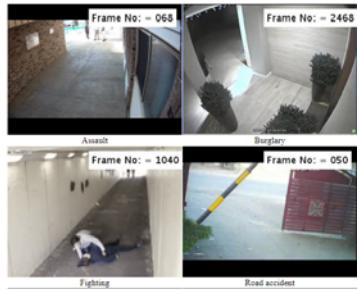
Facial Attributes Detection



Human Re-Identification



Target Detection in WAMI



Diving



Human Action Localization



Single Blank:

He ___ up the steps of the stand and away. (Runs)

Anomaly Detection

Video Fill In The Blank



Reading The Mind



Fully Convolutional Deep Neural Networks for Persistent Multi-Frame Multi-Object Detection in Wide Area Aerial Videos

Rodney LaLonde, Dong Zhang and Mubarak Shah
CVPR-2018

<http://crcv.ucf.edu/papers/cvpr2018/3460.pdf>

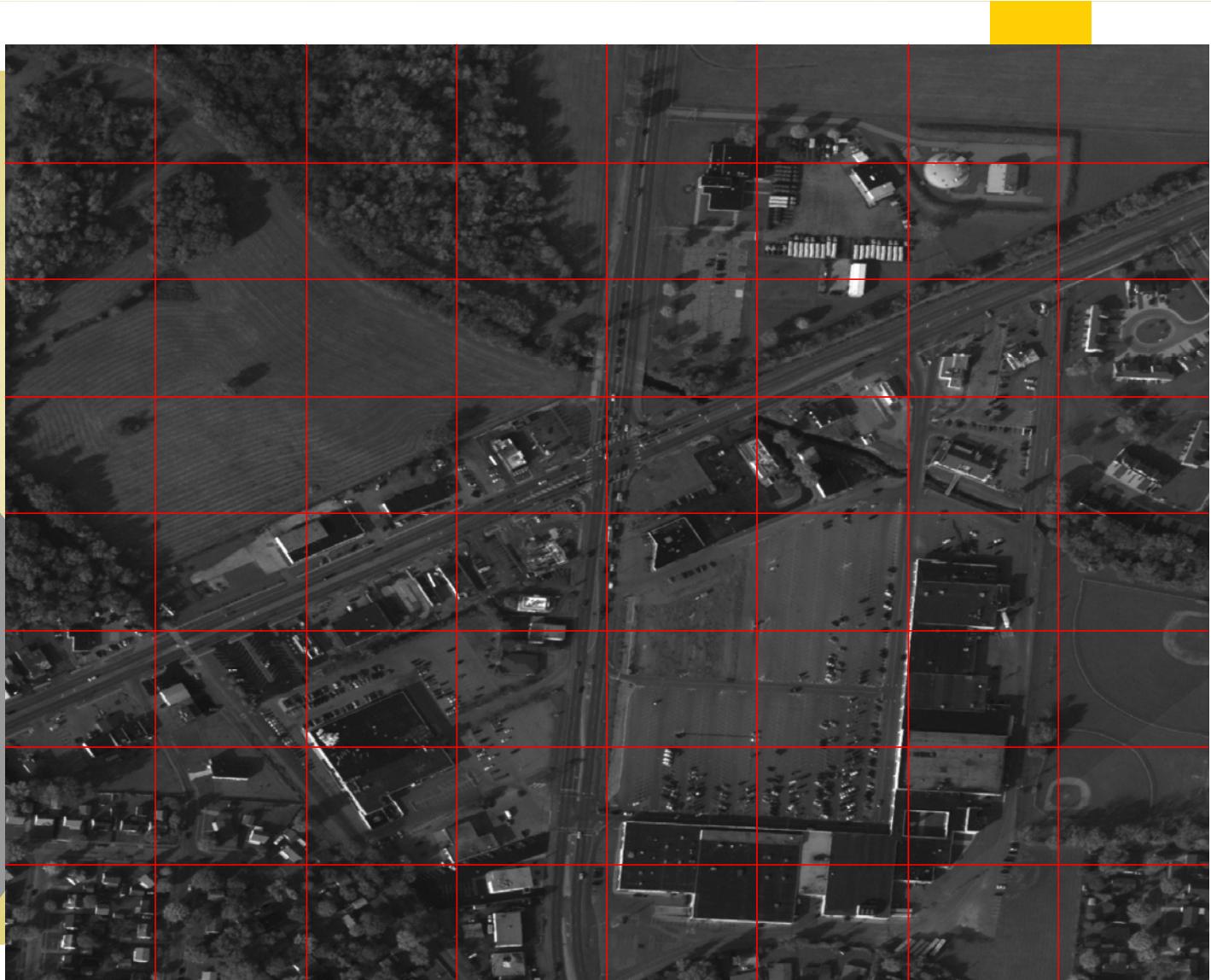
The Goal

Multiple Vehicle
Detection

Red dots are the ground
truth (x,y) annotations.
(WPAFB 2009)



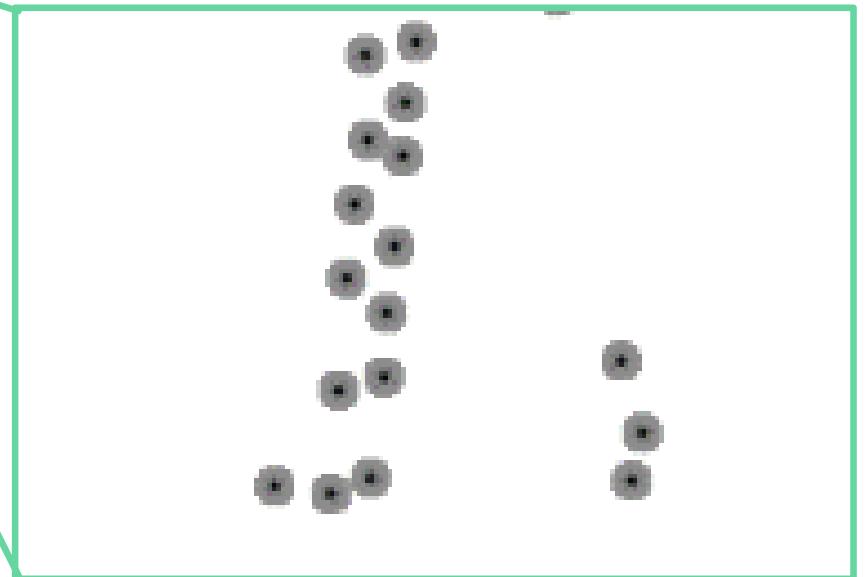
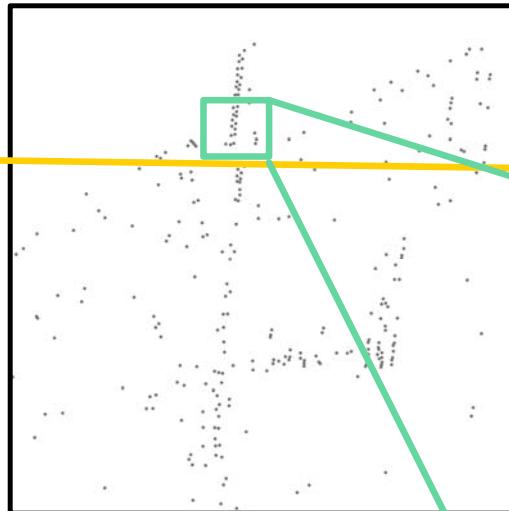
Video Frame Patch Creation



Ground Truth



Heat-maps

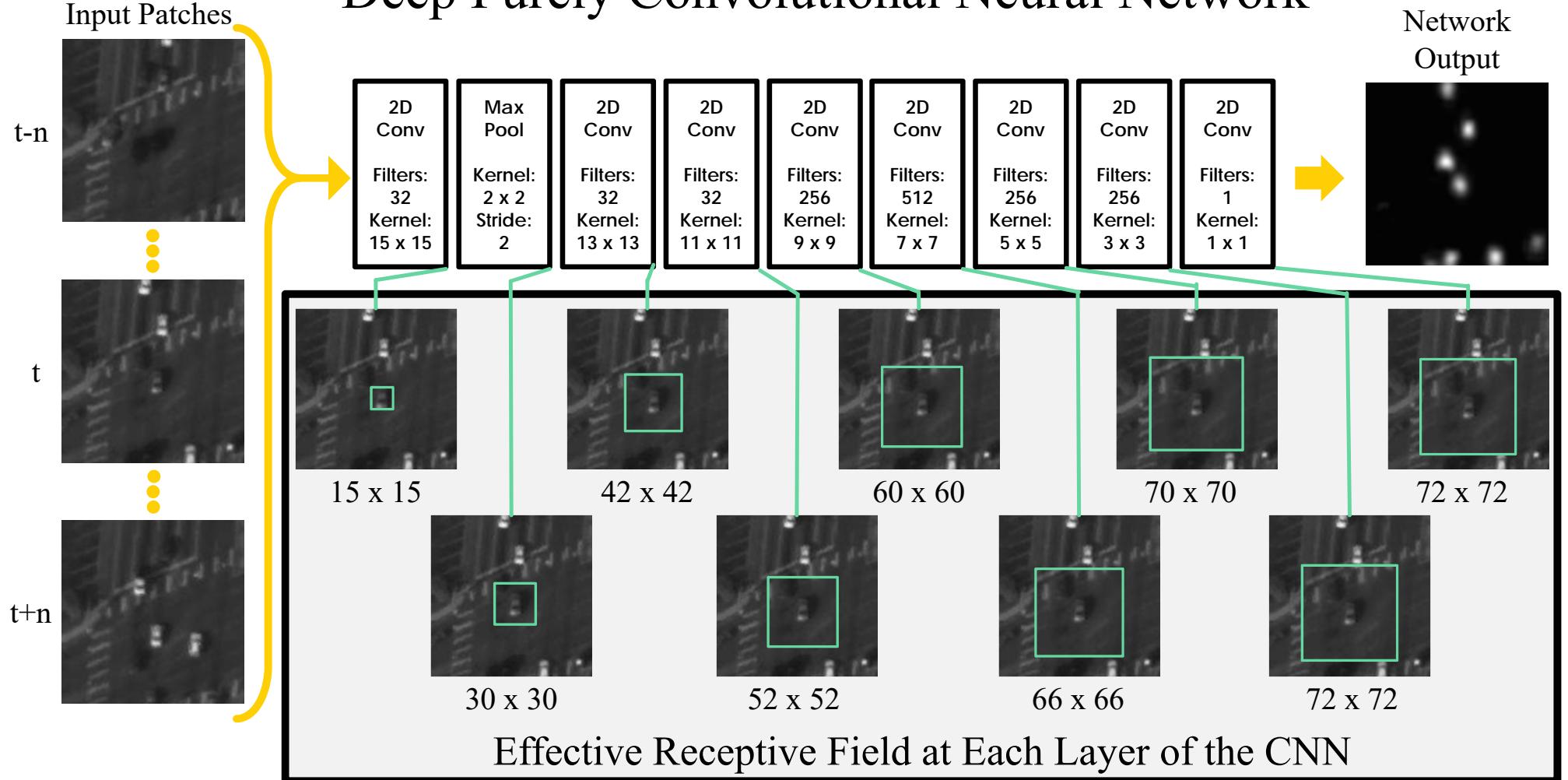


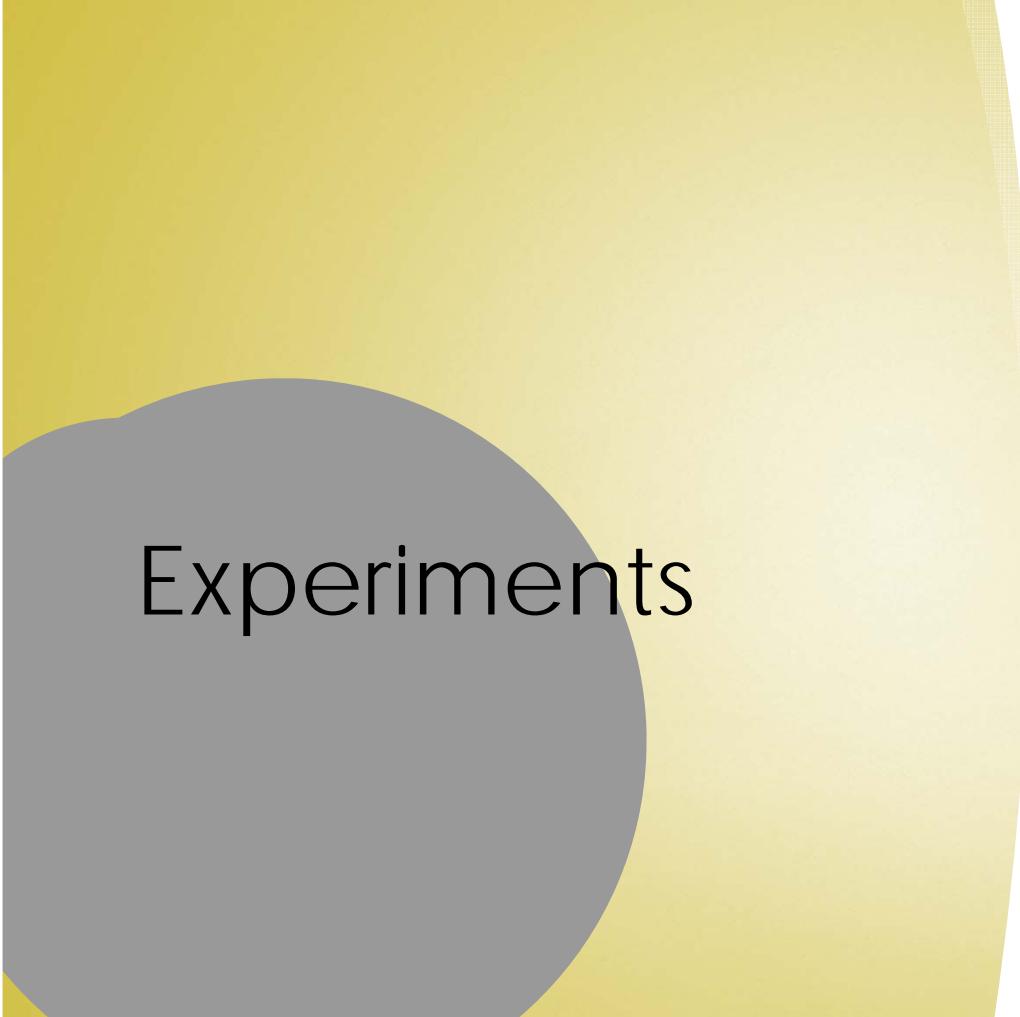
- ▶ Gaussian heat-maps have their colors inverted and σ slightly increased for visualization purposes.
- ▶ Binary segmentation heat-maps have spots with value 1 instead of Gaussians (and 0 background).

Supervised Learning

- ▶ 2D Convolutional Layer
- ▶ ReLU Layer
- ▶ Max Pooling Layer
- ▶ Dropout Layer (50%)
- ▶ Euclidean Loss Layer
- ▶ Solver Type
 - ▶ Adam
- ▶ Deep Learning Framework
 - ▶ Caffe
 - ▶ MATLAB interface
- ▶ GPU
 - ▶ 4 NVIDIA Titan X GPUs
- ▶ Training
 - ▶ Full training on a single GPU: 2 days

Deep Purely Convolutional Neural Network



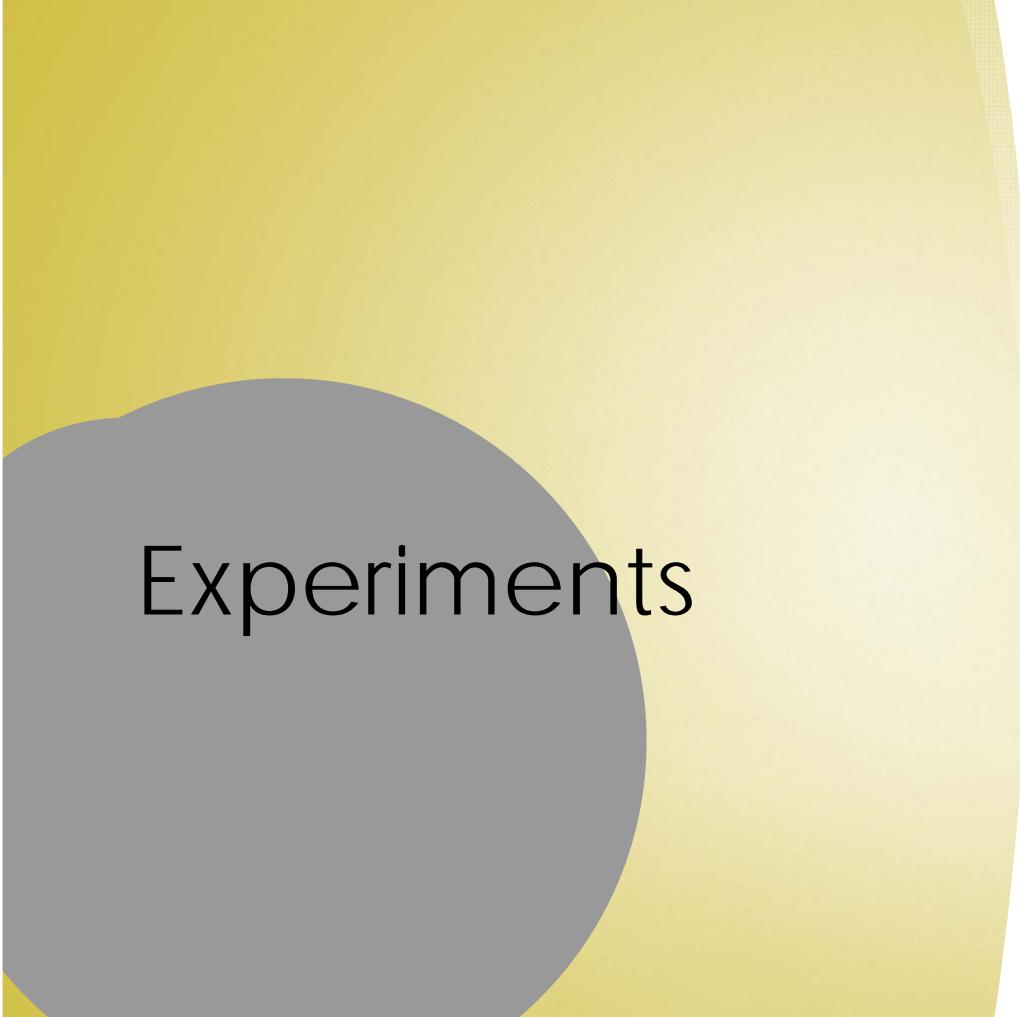


Experiments

1. Data Creation

WPAF 2009 Dataset

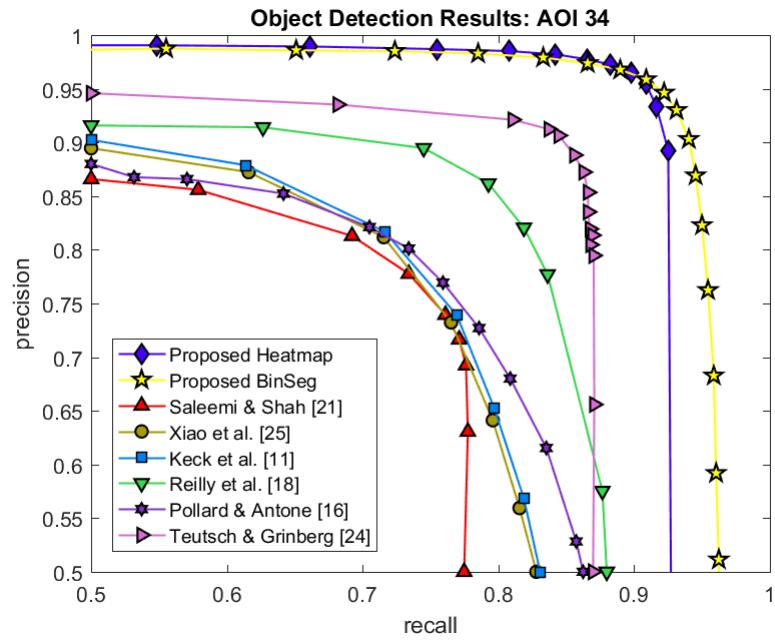
- ▶ Covering an area of over **19. 5 sq. km.**,
- ▶ Frame rate of roughly 1.25 Hz.
- ▶ Over 315 million pixels per frame, with each pixel corresponding to roughly **1/4 meter**.
- ▶ Vehicle makes up only approximately 9 X18 pixels
 - ▶ **2.4 million** vehicles in 1,025 frames of video,
 - ▶ Over **2,000** in every frame.
- ▶ **Eight AOIs**
 - ▶ AOI-1 to AOI-4: 2,278 X2,278 size
 - ▶ AOI 34 is 4,260 X2,604 . AOI 40 is 3,265X2,542 . AOI 41 is 3,207X 2,892
- ▶



Experiments

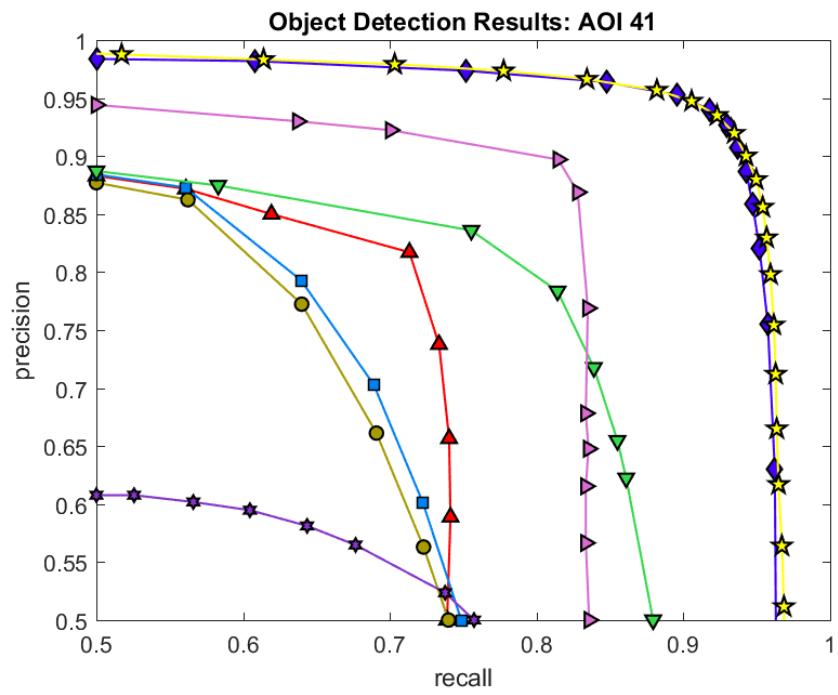
4. Multi-Frame Experiments:
Moving Object Detection

Results AOI 34 Gaussian Heat-map



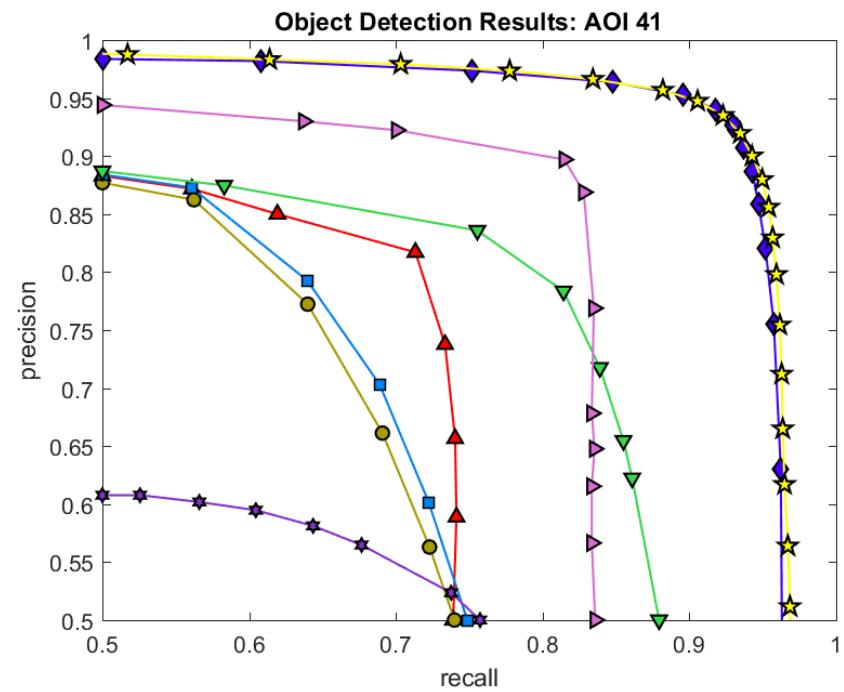
Results AOI 41

Gaussian Heat-map

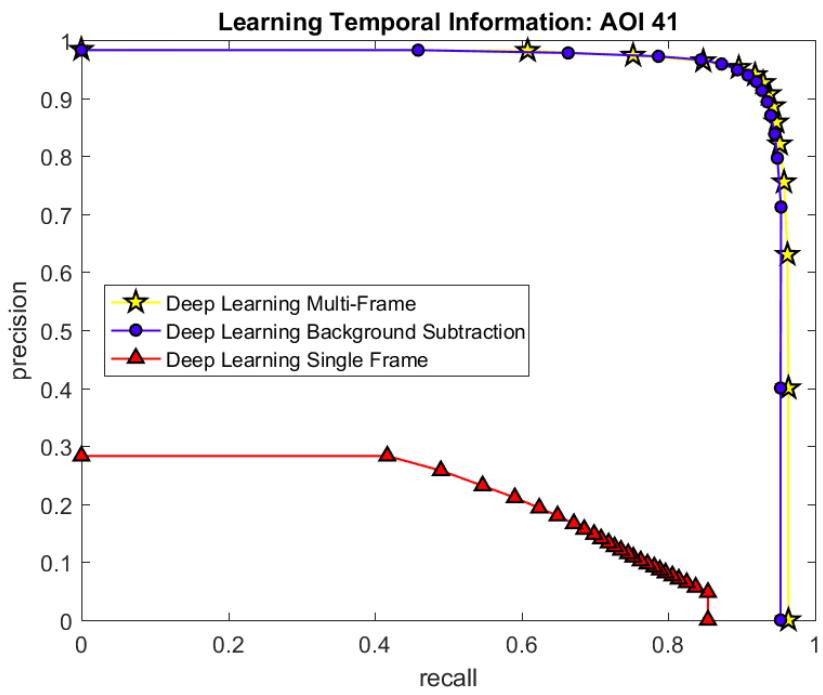


Results AOI 41

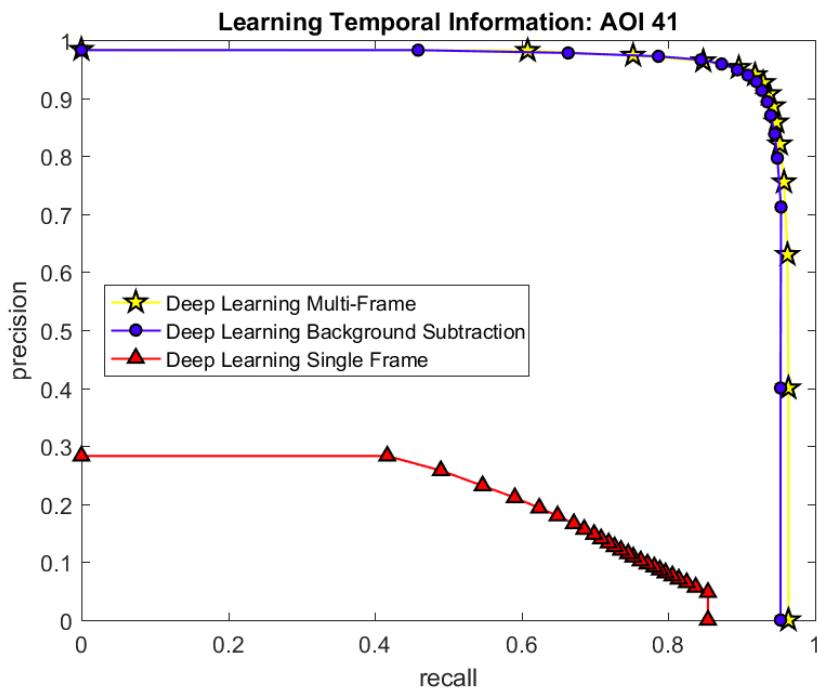
BinSeg Heat-map



Results AOI 41 Background Sub.

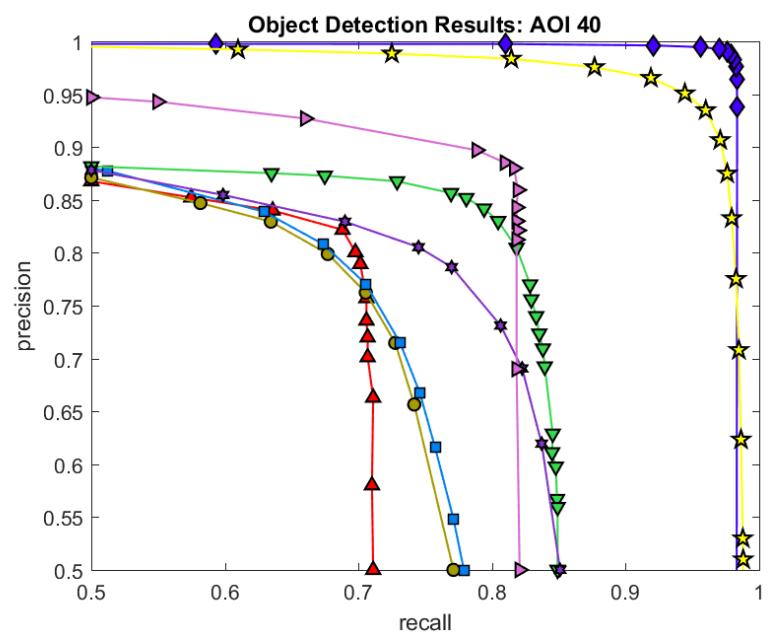


Results AOI 41 Single Frame



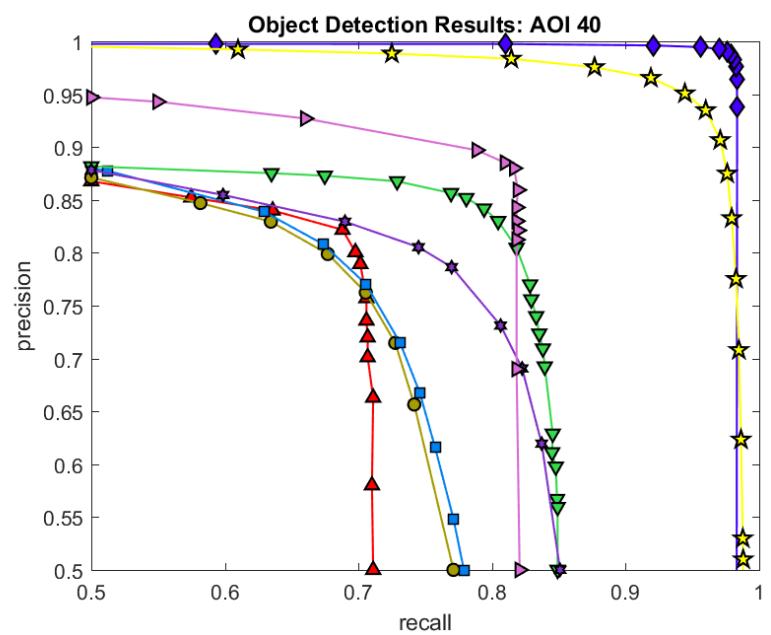
Results AOI 40

Gaussian Heat-map

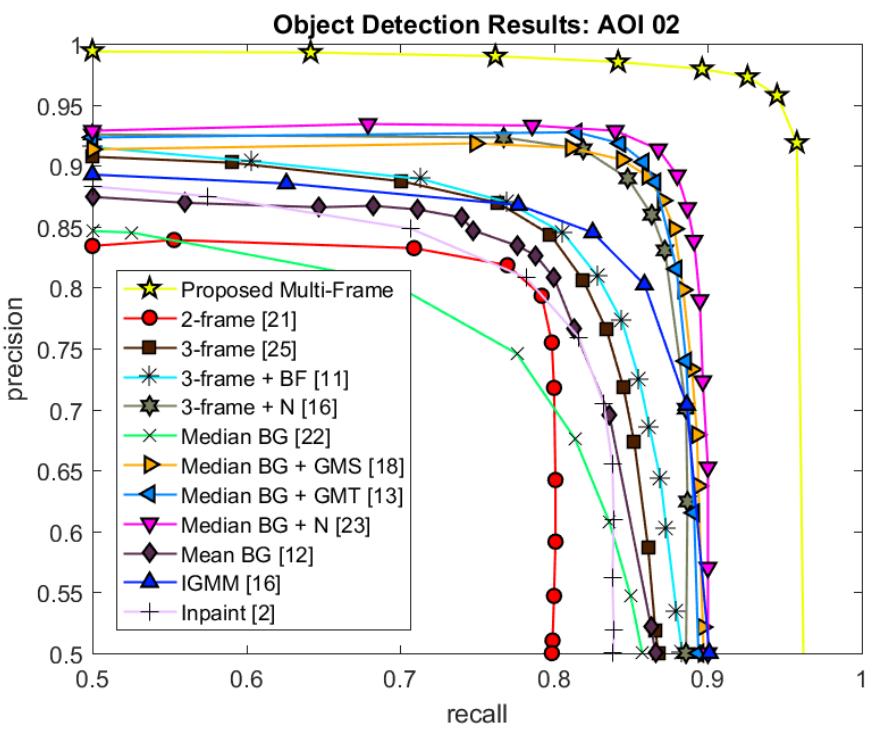


Results AOI 40

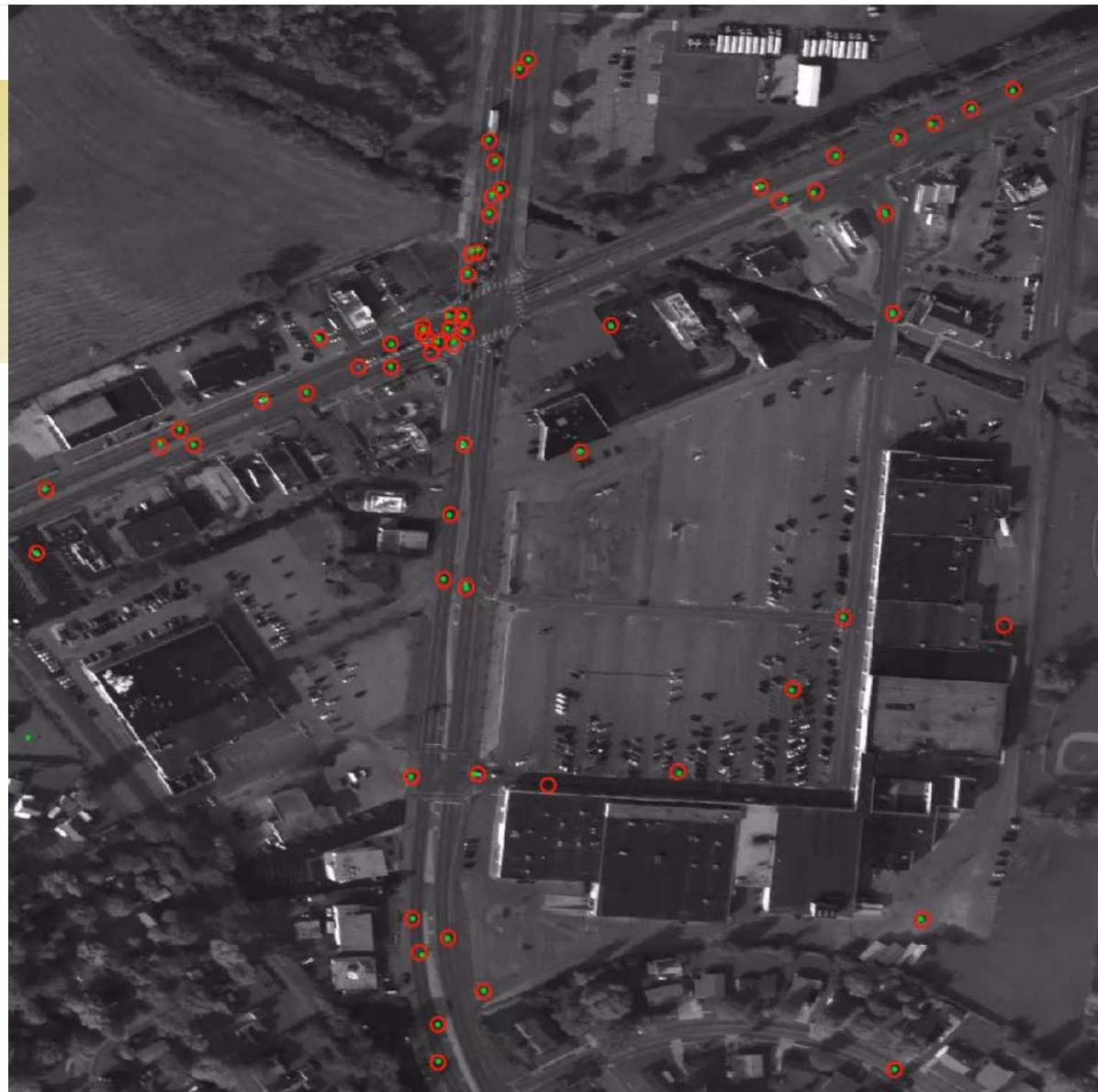
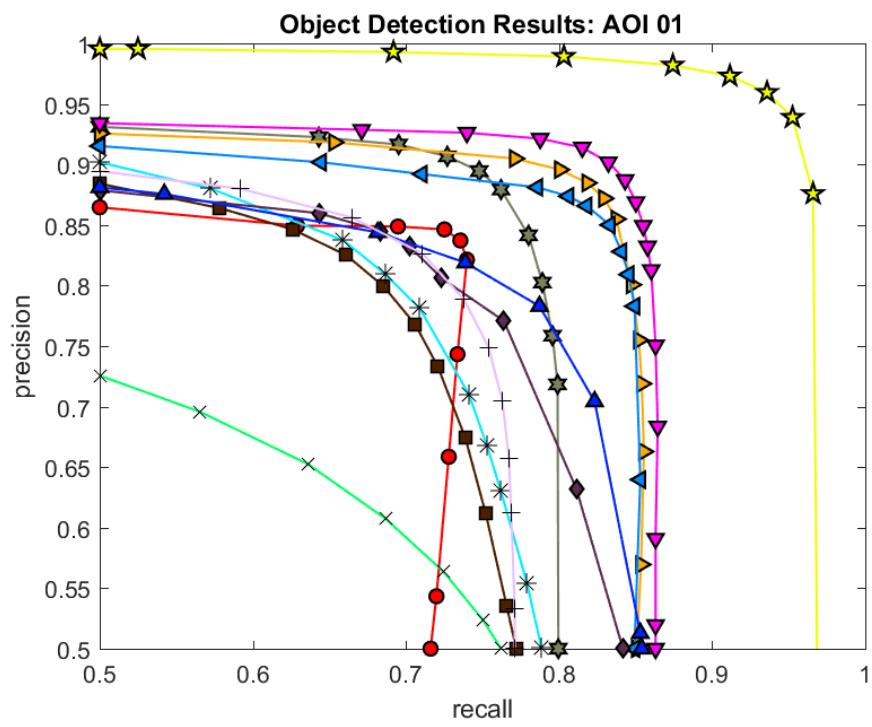
BinSeg Heat-map



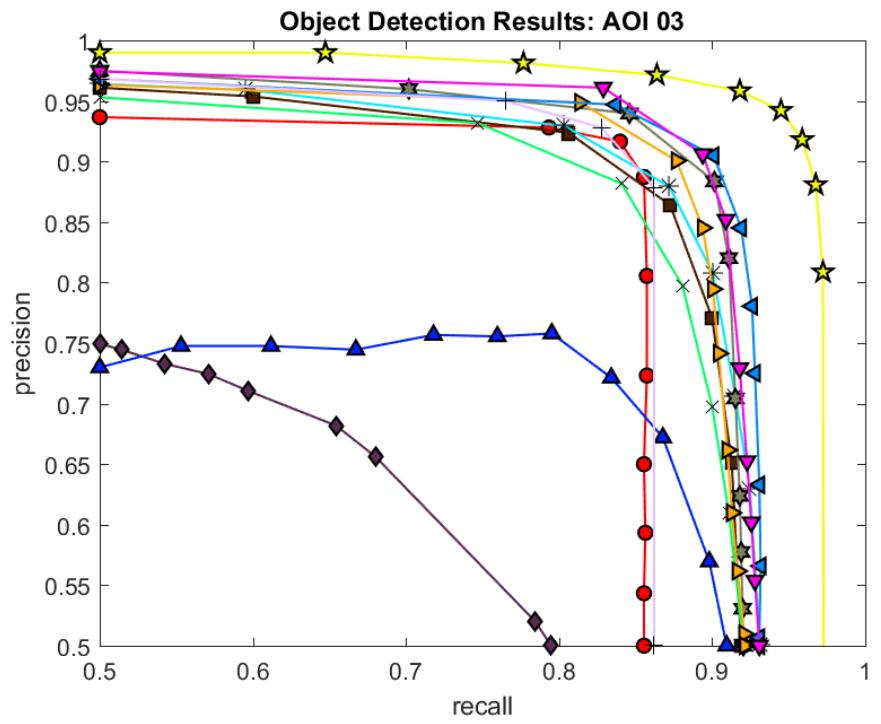
Results AOI 02



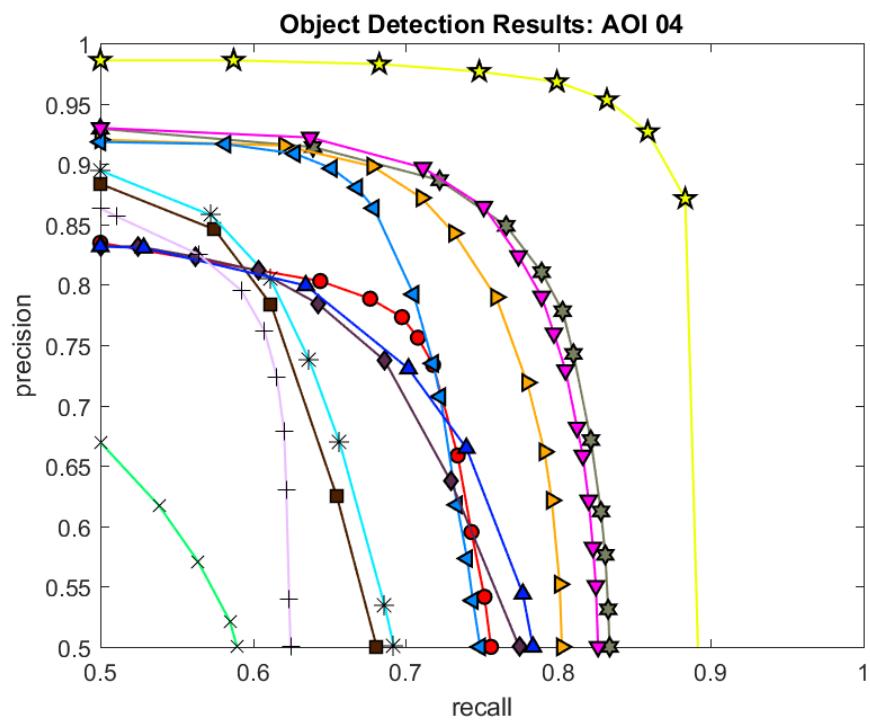
Results AOI 01



Results AOI 03



Results AOI 04



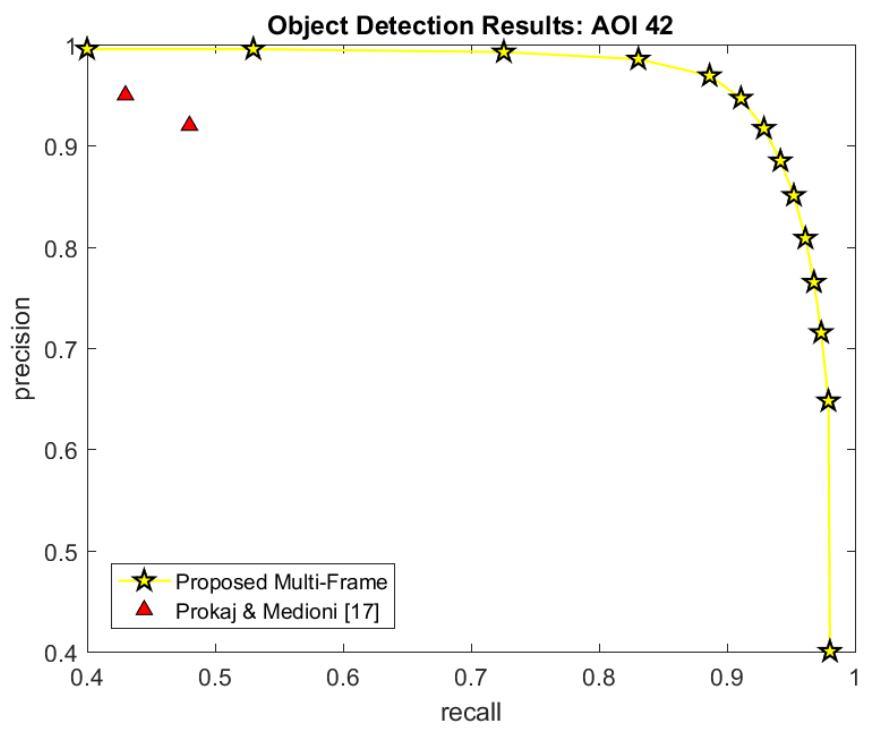
Experiments

4. Multi-Frame Experiments:
Slowing and
Stopped Vehicle
Detections

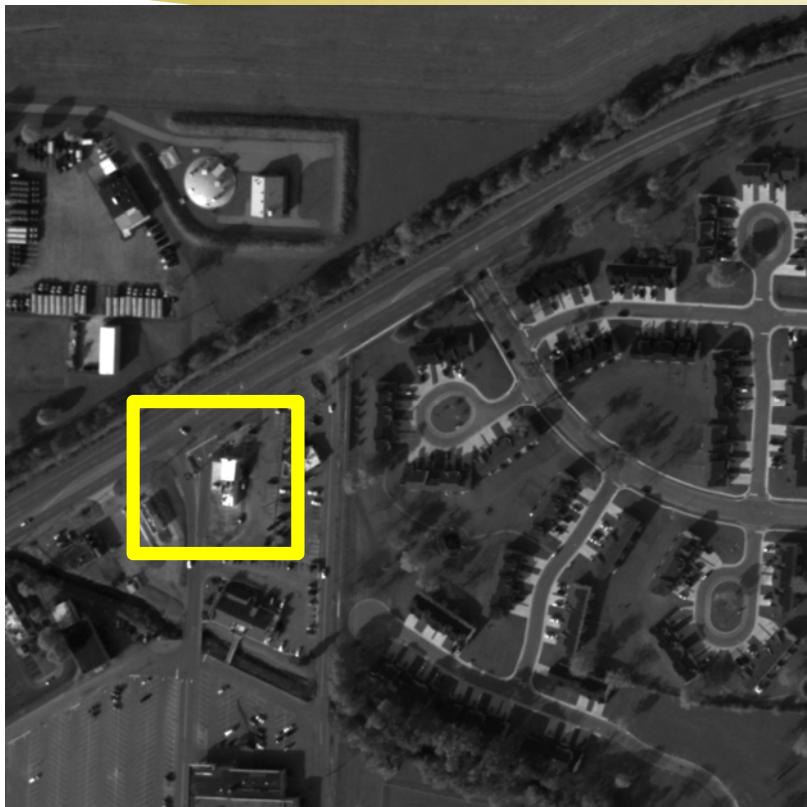
Stopped Vehicles

- ▶ Of the **11** state-of-the-art methods **none can handle** stopped vehicles.
- ▶ For evaluation all stopped vehicles were removed
 - ▶ The number of GT objects is reduced from originally **460,612** to **163,158**
- ▶ Our approach can detect stopped vehicles.
 - ▶ Experiments were ran on AOI 42 with no ground-truth coordinates removed.

Results AOI 42



Stopped Vehicle Detection Example



Comparison of F_1 Scores on Eight Crop and Aligned Sections of the WPAFB 2009 Dataset

Method
Sommer <i>et al.</i> [23]
Shi [22]
Liang <i>et al.</i> [13]
Kent <i>et al.</i> [12]
Aeschliman <i>et al.</i> [2]
Pollard & Antone (3-frame + N) [16]
Saleemi & Shah [21]
Xiao <i>et al.</i> [25]
Keck <i>et al.</i> [11]
Reilly <i>et al.</i> [18]
Pollard & Antone (IGMM) [16]
Teutsch & Grinberg [24]
Prokaj & Medioni [17]
Proposed Multi-Frame

Conclusions

- ▶ We have proposed a novel fully convolutional neural network based method for persistent multi-frame multi-object detection in aerial videos.
- ▶ In our method, we successfully taking advantage of both appearance and motion cues and integrate them into a single detection network, trained end to end.
- ▶ We have shown comparisons with many state-of-the-art methods, and the performance improvements are relatively 5 to 16% on moving objects for multiple videos in the WPAFB 2009 dataset as measured by F1 score and nearly 50% relative improvement on persistent detections compared to [17].
- ▶ Additionally, while detections are considered true positives if they fall within 20 pixels of the ground-truth, the proposed method's mean distance from ground truth annotations, averaged over all true positive detections, was roughly 2 pixels (0.5 m), compared to 5.5 pixels reported in [24].
- ▶ We further demonstrated that the proposed method can handle stopped vehicles well, which is often a failure case in other methods.

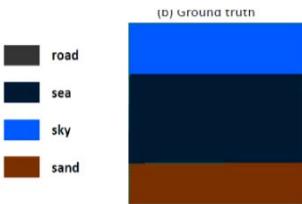


Fully Convolutional Deep Neural Networks for Persistent Multi-Frame Multi-Object Detection in Wide Area Aerial Videos

Rodney LaLonde, Dong Zhang and Mubarak Shah
CVPR-2018

<http://crcv.ucf.edu/papers/cvpr2018/3460.pdf>

Contents



Sematic Segmentation



Facial Attributes Detection



Human Re-Identification



Target Detection in WAMI



Anomaly Detection

Diving



Human Action Localization



Single Blank:

He ___ up the steps of the stand and away. (Runs)

Video Fill In The Blank



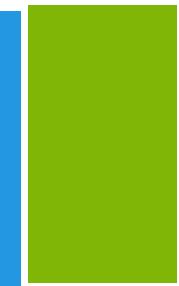
Reading The Mind



Real-World Anomaly Detection in Surveillance videos

Waqas Sultani, Chen Chen, Mubarak Shah

Computer Vision and Pattern Recognition (CVPR), 2018
<https://arxiv.org/pdf/1801.04264.pdf>





Motivation

- Over 30 Millions cameras in US
- Over 4 Billions hours of videos per week
- **Manual supervision is impossible**
- Automatic Analysis is highly needed





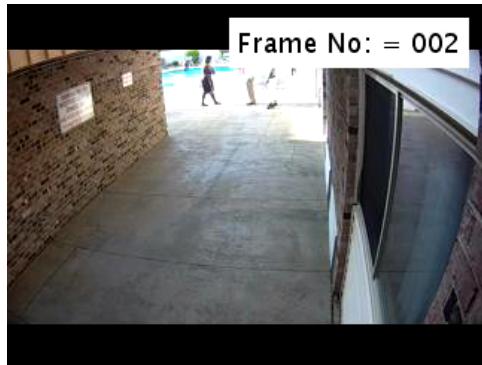
Abuse



Arrest



Arson



Assault



Burglary



Stealing

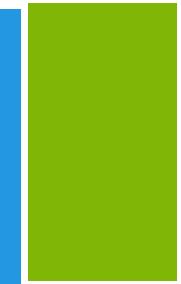
Note: we fast play or trim some videos due to their long durations.





Anomaly Detection

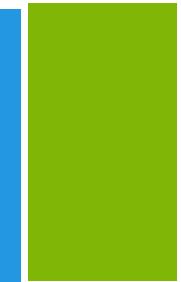
- The goal is to timely signal an activity that deviates from normal patterns.
- Anomalous events:
 - traffic accidents, crimes or illegal activities, etc
- Anomalous events rarely occur as compared to normal activities.





Our Approach

- Learn anomalies by exploiting both normal and anomalous videos.
- Avoid annotating the anomalous clips in training videos.
- Learn anomaly through the deep multiple instance ranking framework by leveraging weakly labeled training videos:
 - A video is normal or contains anomaly somewhere, but we do not know where.



Anomaly Detection

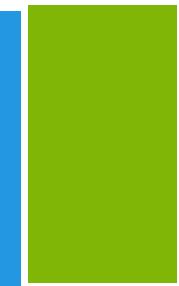
- The goal is to timely signal an activity that deviates normal patterns.
- Anomalous events:
 - traffic accidents, crimes or illegal activities, etc
- Anomalous events rarely occur as compared to normal activities.

Our Approach

- Learn anomalies by exploiting both normal and anomalous videos.
- Avoid annotating the anomalous clips in training videos.
- Learn anomaly through the deep multiple instance ranking framework by leveraging weakly labeled training videos:
 - A video is normal or contains anomaly somewhere, but we do not know where.

+

Example Anomalous Videos in the Dataset





Explosion



Robbery



Arrest



Fighting



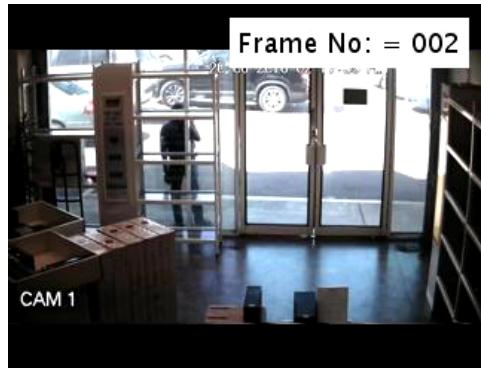
Road Accident



Shooting



+



Shoplifting



Vandalism



Weakly labeled Crime Detection Framework

Ranking

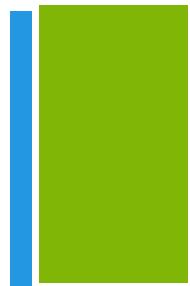
$$f(\mathcal{V}_a) > f(\mathcal{V}_n),$$

MIL Ranking

$$\max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) > \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i),$$

Loss Function

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i))$$



Weakly labeled Crime Detection Framework

Ranking

$$f(\mathcal{V}_a) > f(\mathcal{V}_n),$$

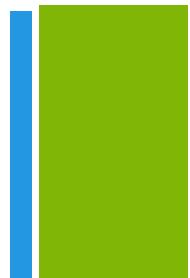
MIL Ranking

$$\max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) > \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i),$$

Loss Function

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i \in \mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i \in \mathcal{B}_n} f(\mathcal{V}_n^i))$$

$$+ \lambda_1 \underbrace{\sum_{i=1}^{n-1} (f(\mathcal{V}_a^i) - f(\mathcal{V}_a^{i+1}))^2}_{\text{Smoothness term}} + \lambda_2 \underbrace{\sum_{i=1}^n f(\mathcal{V}_a^i)}_{\text{Sparsity term}},$$



Weakly labeled Crime Detection Framework

Dataset & Experimental Results





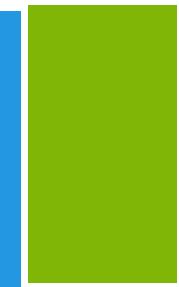
Anomaly	Number of videos
Burglary	100
Fighting	50
Road Accidents	150
Robbery	150
Shooting	50
Shoplifting	50
Stealing	100
Abuse	50
Arrest	50
Arson	50
Assault	50
Explosion	50
Vandalism	50
Normal	950

Number of videos of each category

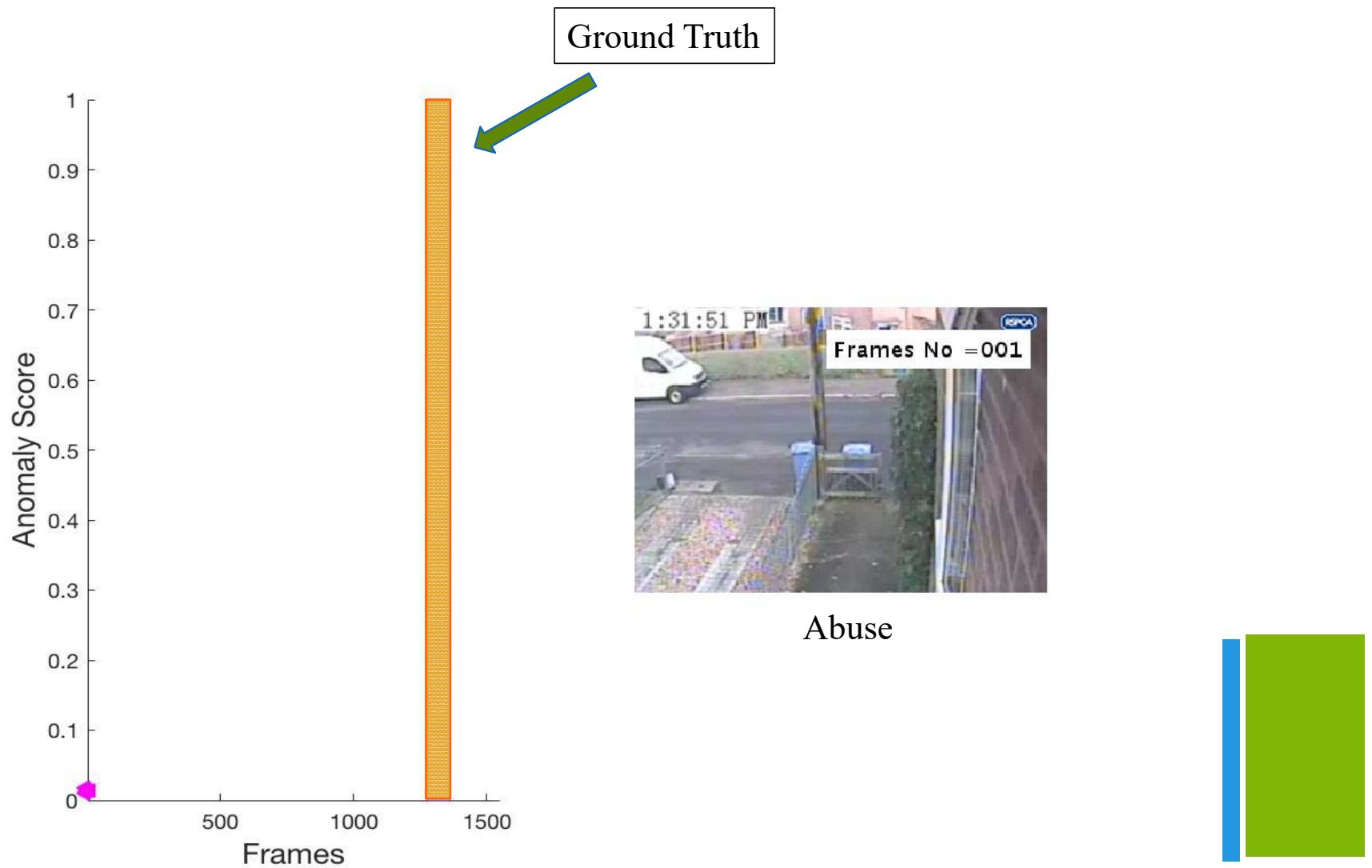




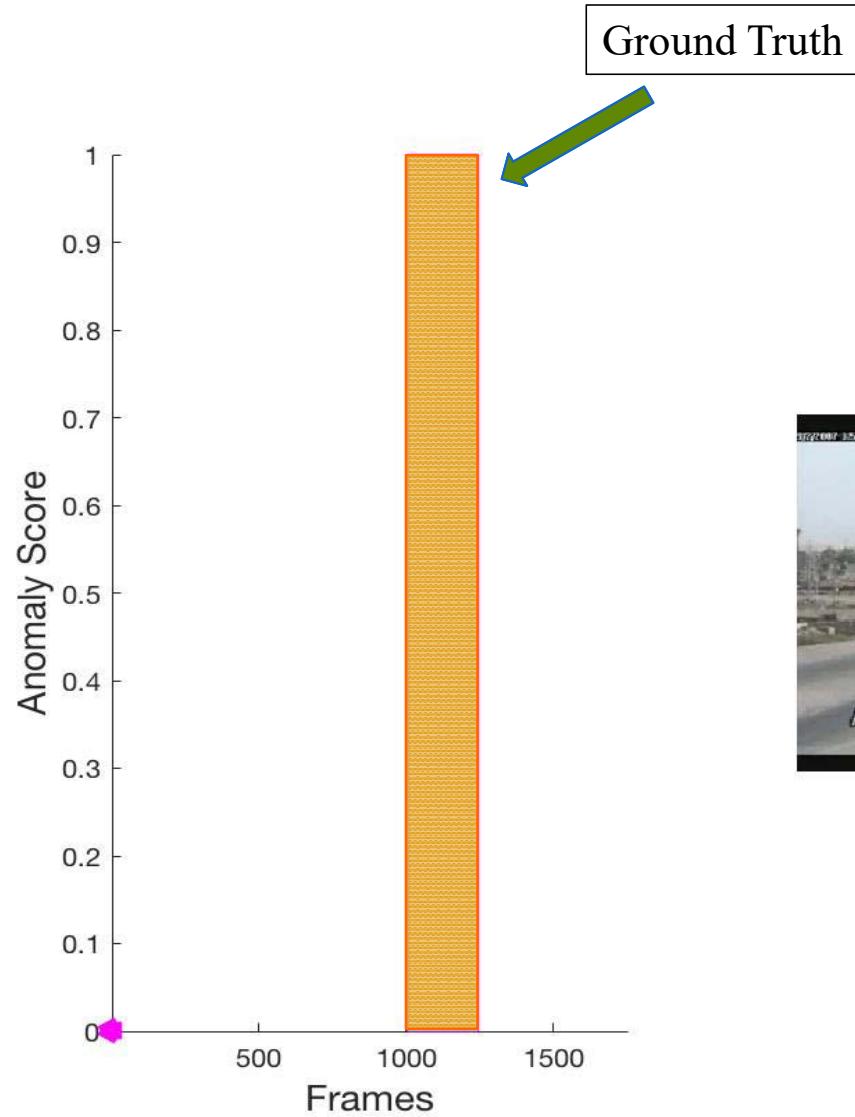
Qualitative Results



+



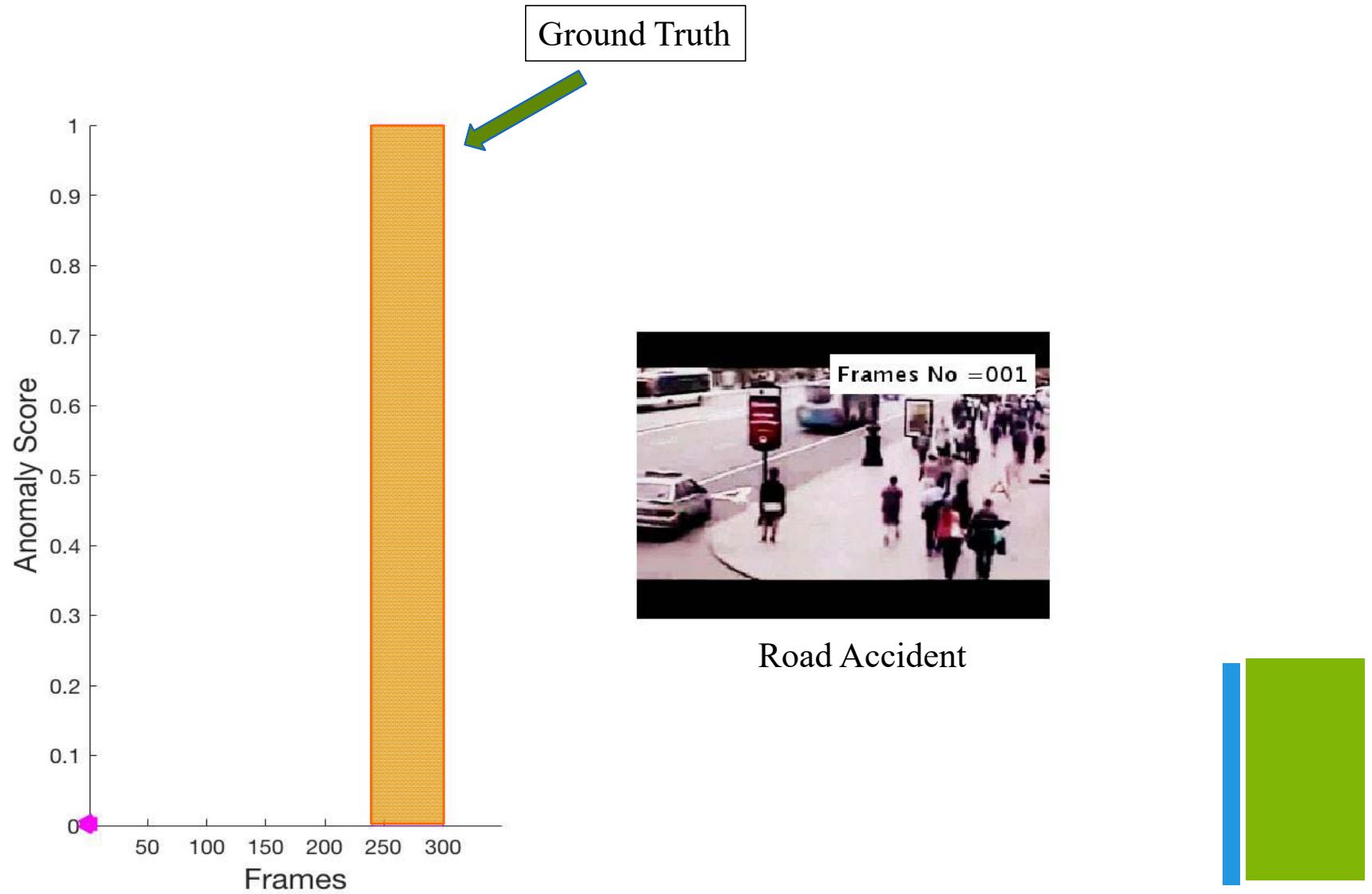
+



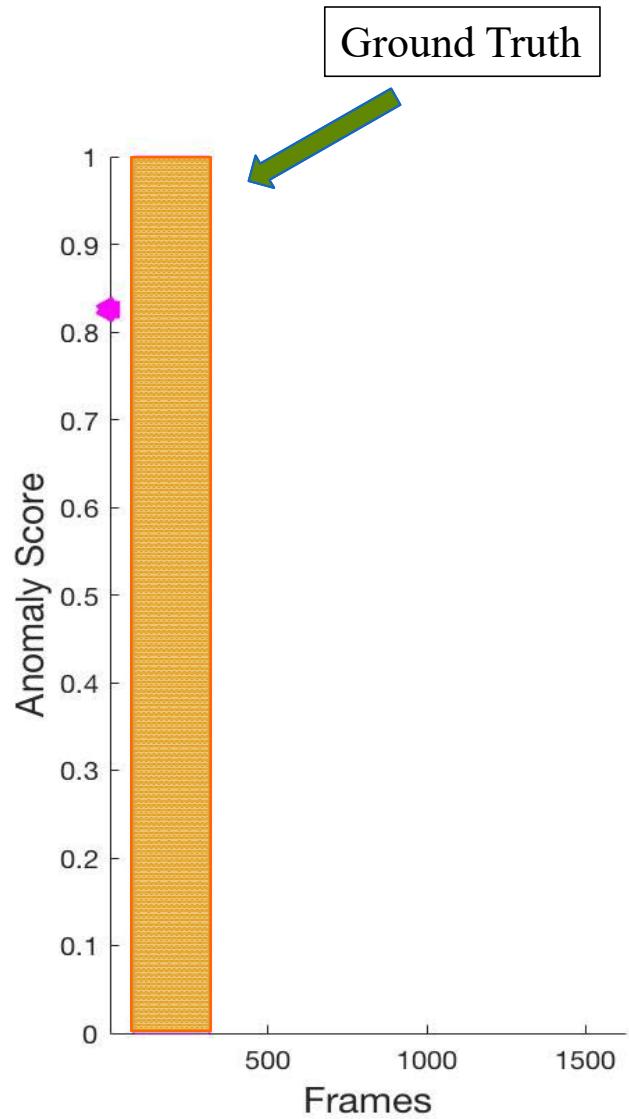
Explosion



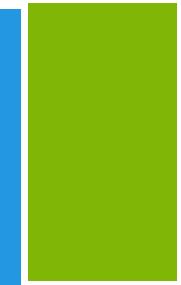
+



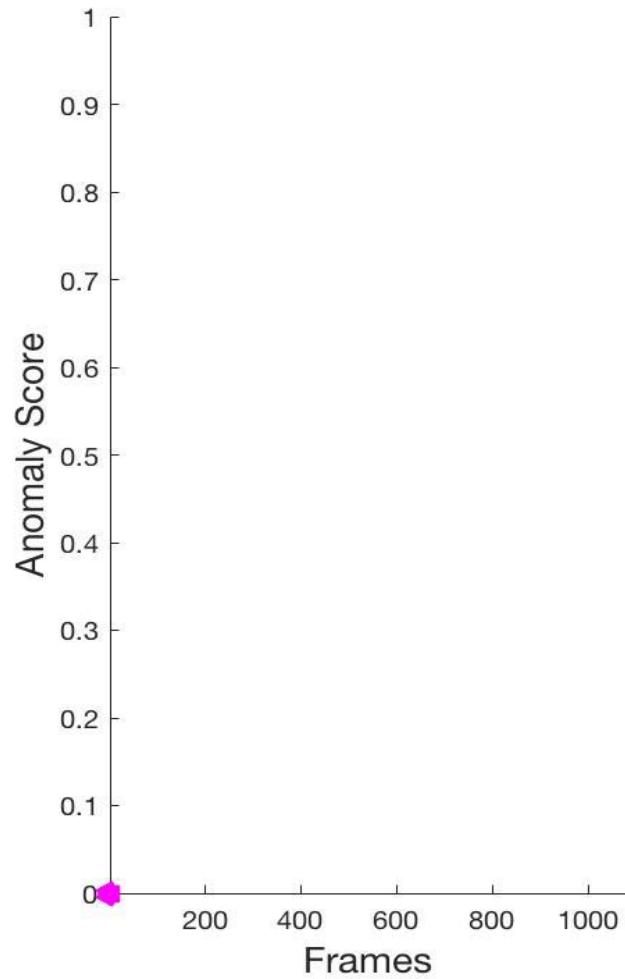
+



Shooting



+

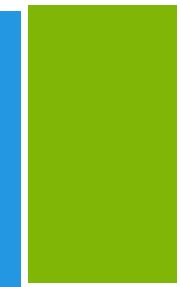


Normal video



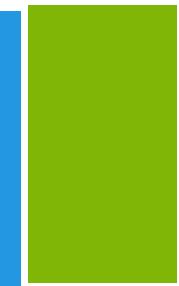
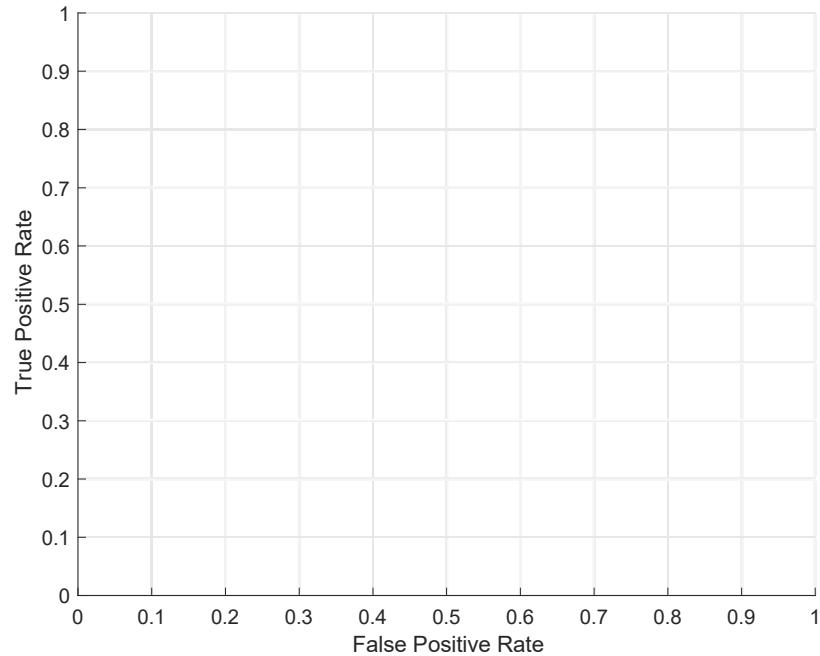


Quantitative Results



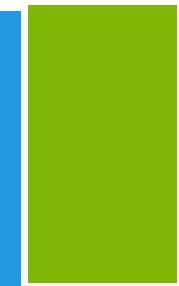
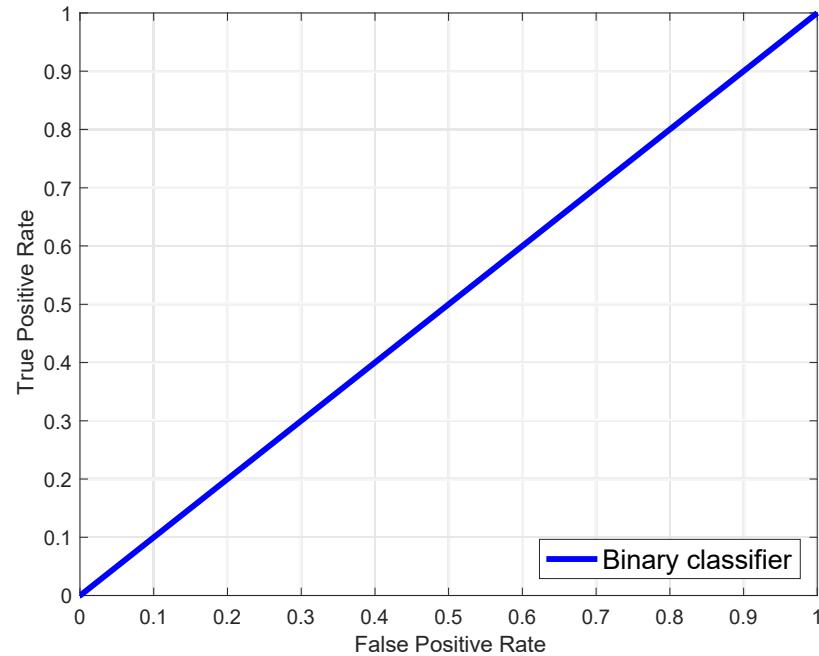


Comparison with stat-of-art anomaly detection methods

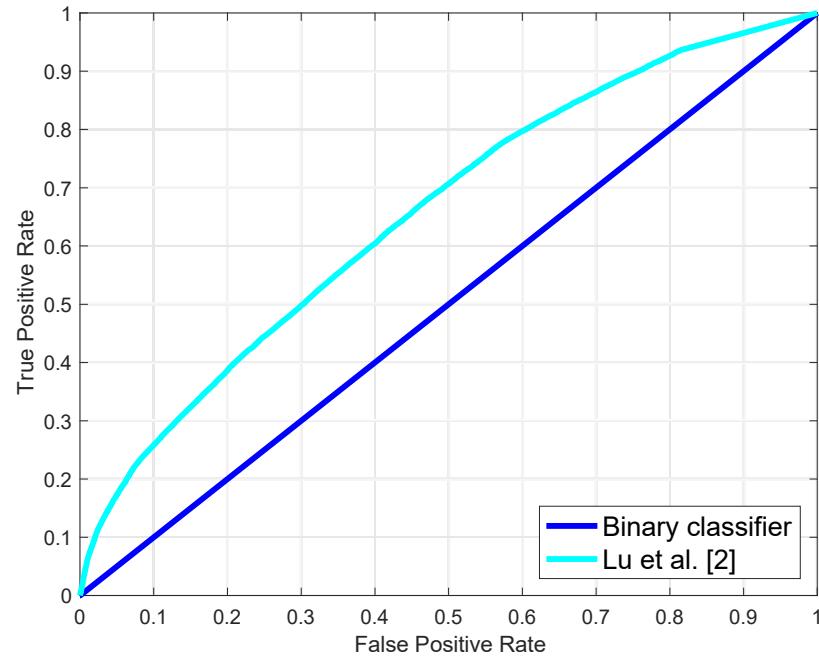




Comparison with stat-of-art anomaly detection methods

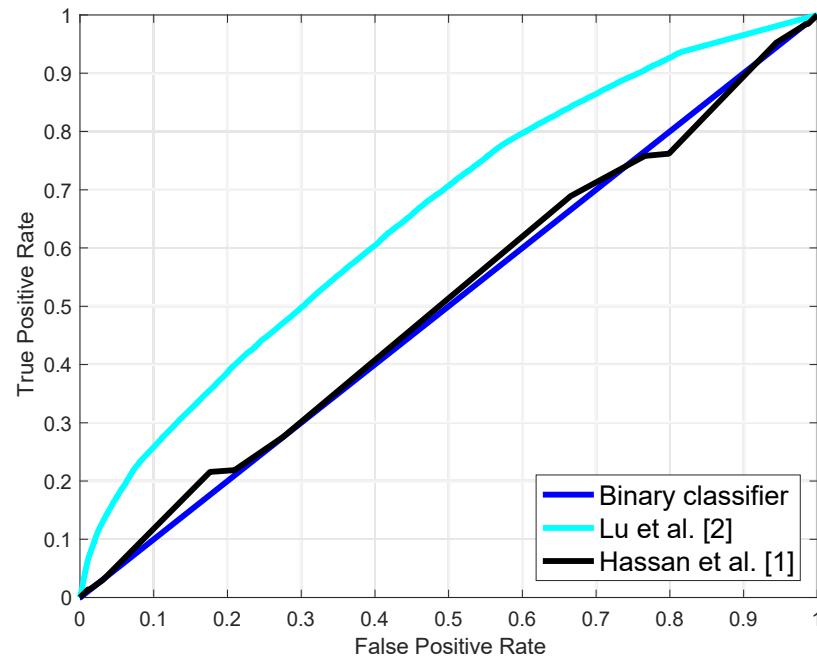


Comparison with stat-of-art anomaly detection methods



[2] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In ICCV, 2013.

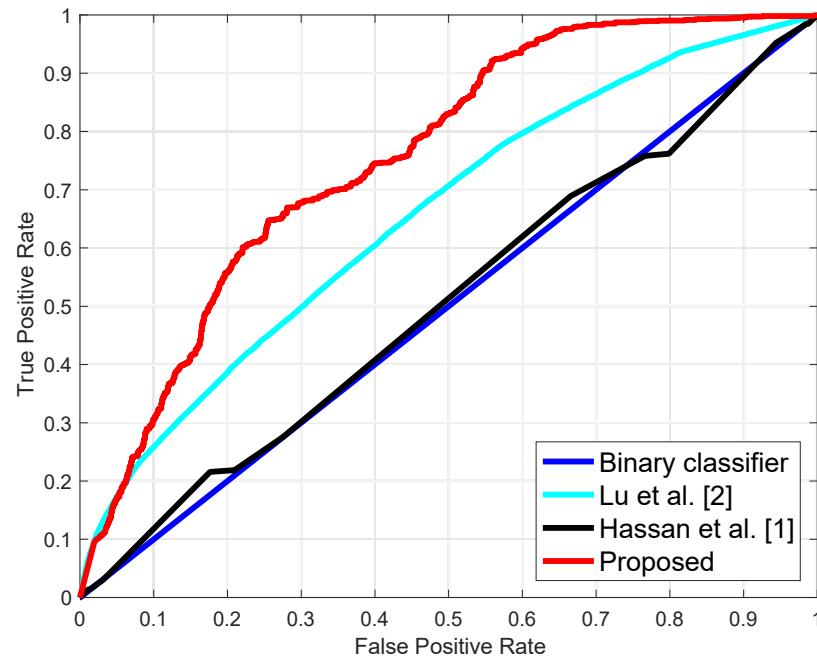
Comparison with stat-of-art anomaly detection methods



[1] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In CVPR, 2016.

[2] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In ICCV, 2013.

Comparison with stat-of-art anomaly detection methods



[1] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In CVPR, 2016.

[2] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In ICCV, 2013.



Comparison with stat-of-art anomaly detection methods

Method	AUC
Binary Classifier	50.0
Hasan et al. [1]	50.6
Lu et al. [2]	65.51
Proposed w/o constraints	74.44
Proposed w constraints	75.41

[1] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In CVPR, 2016.

[2] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In ICCV, 2013.



Comparison with stat-of-art anomaly detection methods

False Alarm Rate

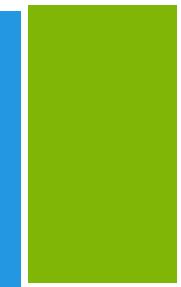
Method	False Alarm Rate
Hasan et al. [1]	27.2
Lu et al. [2]	3.1
Proposed	1.9

- Lower is better

[1] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis. Learning temporal regularity in video sequences. In CVPR, 2016.

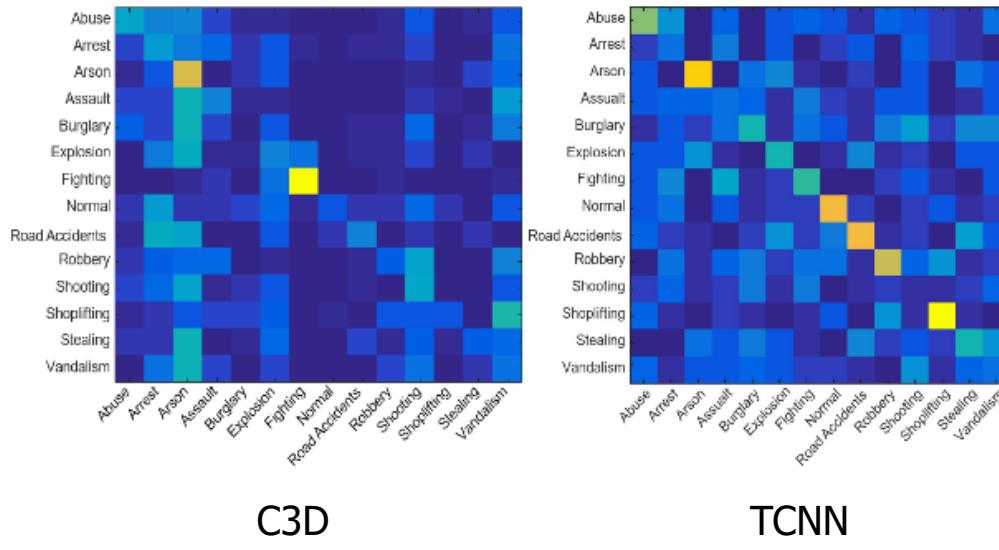
[2] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In ICCV, 2013.

Action Recognition Results





Method	Accuracy
C3D [1]	23.0
TCNN [2]	28.4



[1] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In ICCV, 2015

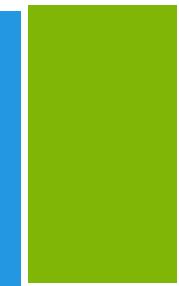
[2] R. Hou, C. Chen, and M. Shah. Tube convolutional neural network for action detection in videos. In ICCV, 2017.



Real-World Anomaly Detection in Surveillance videos

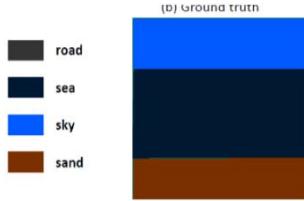
Waqas Sultani, Chen Chen, Mubarak Shah

Computer Vision and Pattern Recognition (CVPR), 2018
<https://arxiv.org/pdf/1801.04264.pdf>



Thank You!

Contents



Sematic Segmentation



Facial Attributes Detection

Human Re-Identification



Target Detection in WAMI



Anomaly Detection

Human Action Localization



Single Blank:

He ___ up the steps of the stand and away. (Runs)

Video Fill In The Blank



Reading The Mind

T-CNN for Action Detection in Videos

Rui Hou, Chen Chen, Mubarak Shah

ICCV-2017

<http://crcv.ucf.edu/papers/iccv17/T-CNN-camera-ready.pdf>

Action Recognition



Biking



SalsaSpin



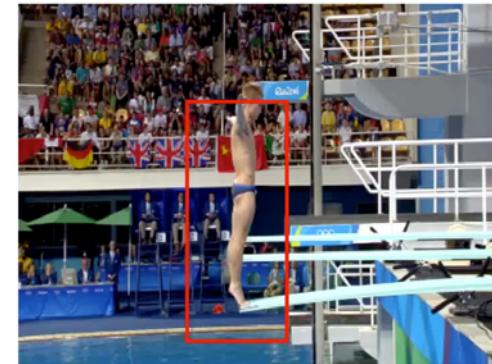
LongJump



TennisSwing

Action Detection

- Trimmed video
 - Spatio-temporal localization
- Untrimmed video
 - Temporal localization
 - Spatio-temporal localization



Diving



Tennis

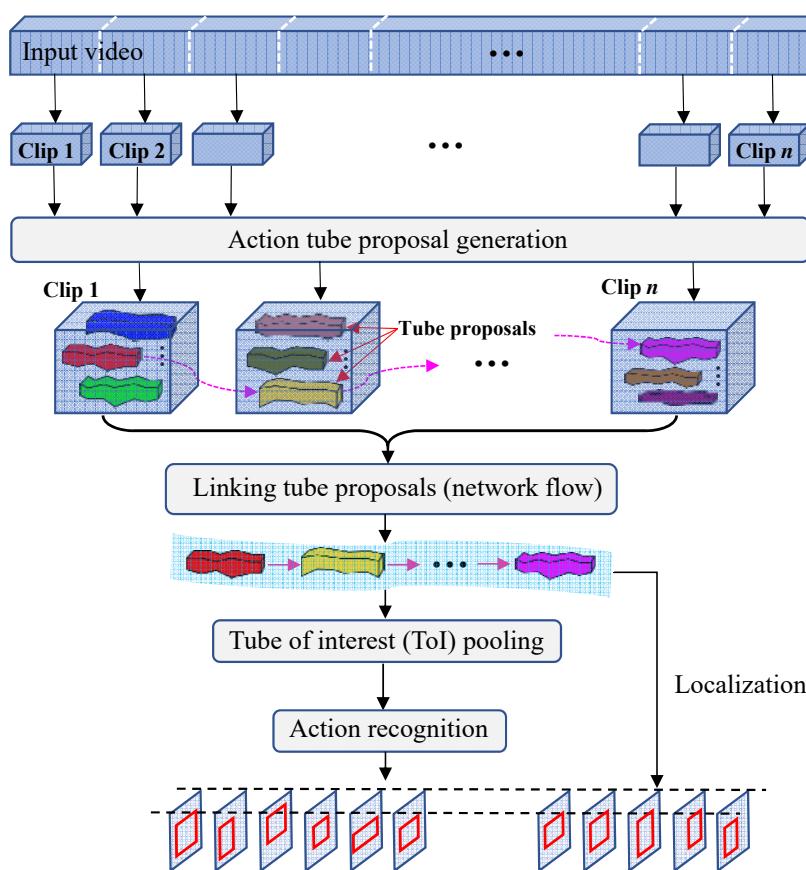
Action Segmentation

- Pixel-wise Spatio-Temporal localization



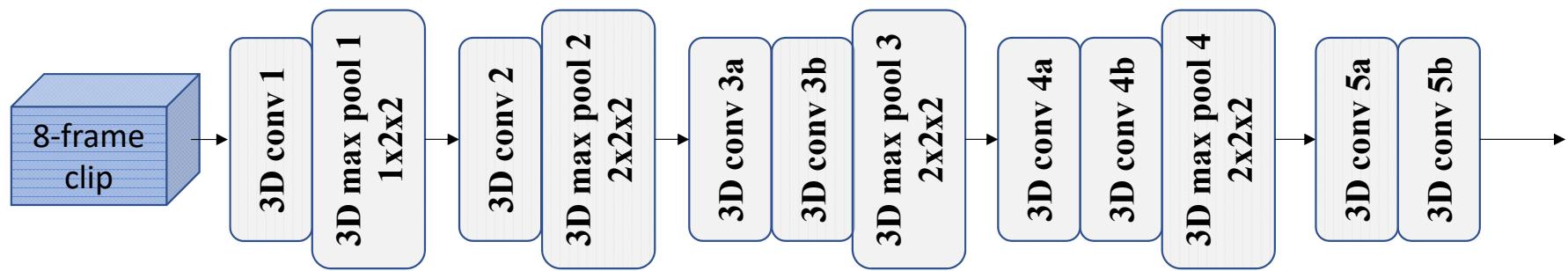
Golf

Overview of Tube-CNN

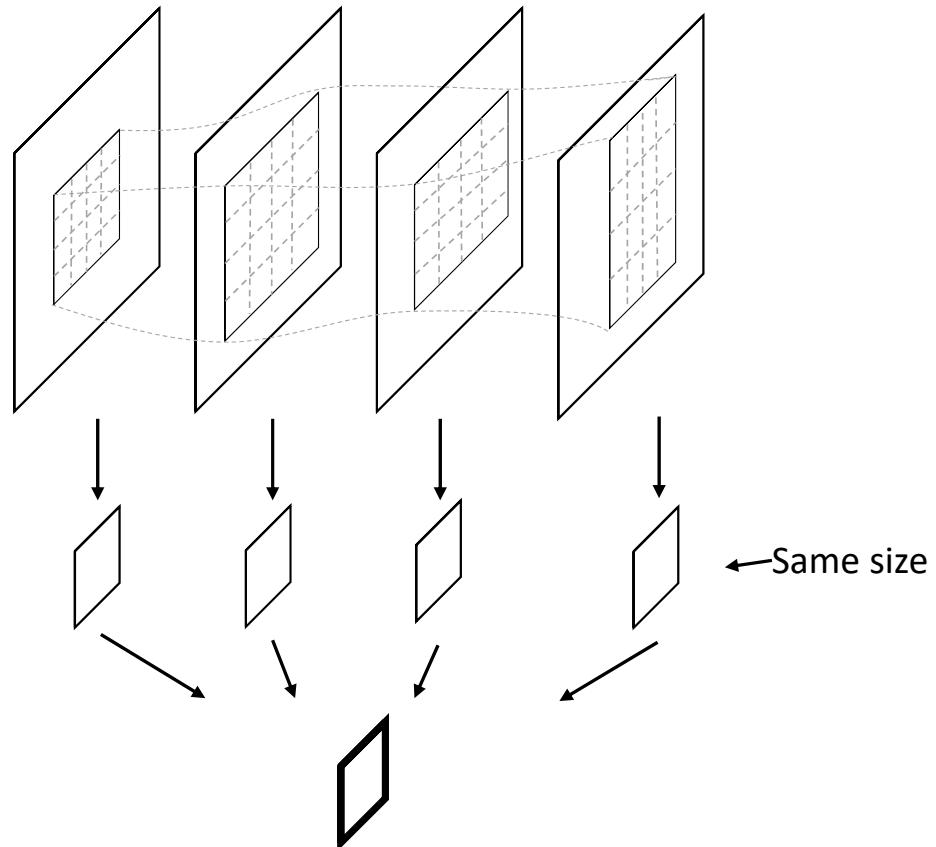


- Divide input video into equal length clips
- Generate tube action proposals
- Link tube proposals
- Tube of Interest Max-pooling
- Recognize and localize actions

3D ConvNet

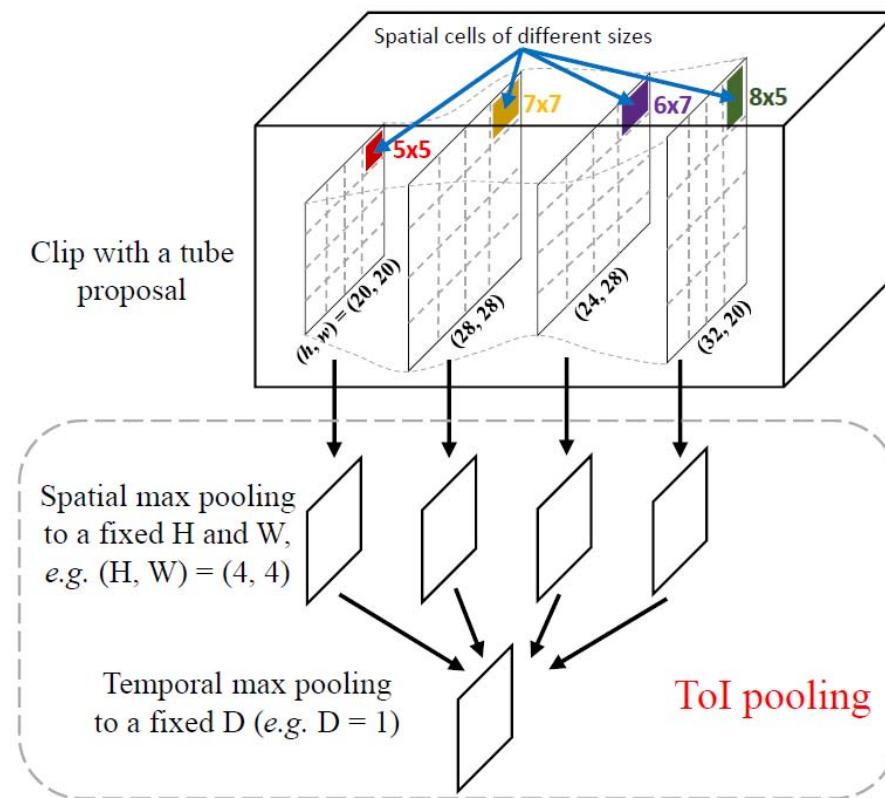


Tube of Interest Max Pooling (TOI)

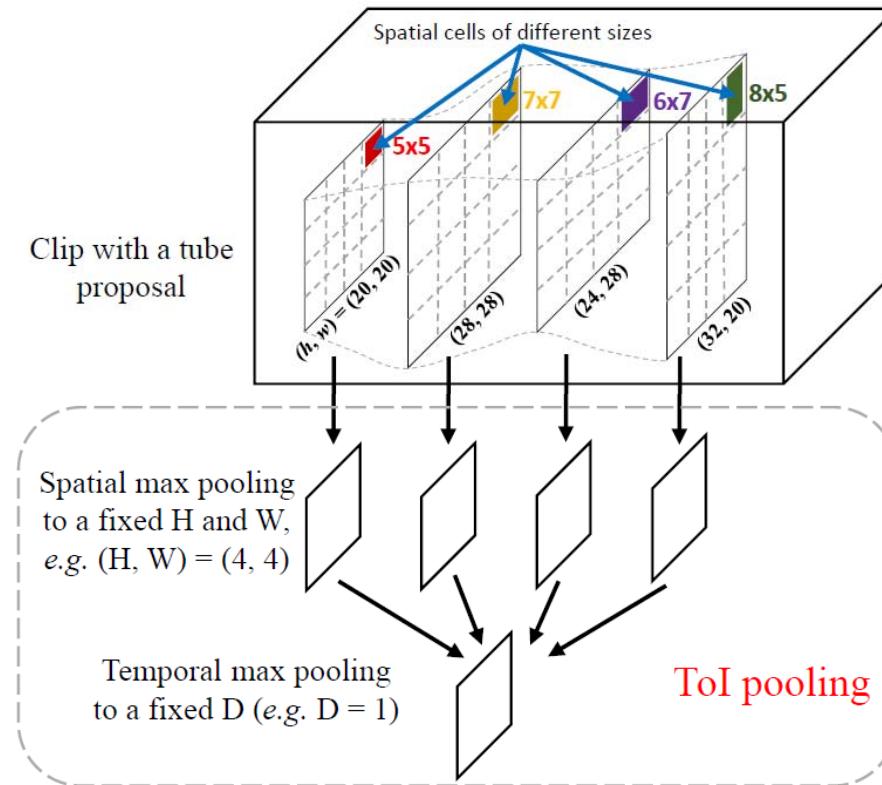


- Given a video clip
- And generated Tube of Interest
- Spatially max pool to a fixed shape
- Temporally max pool to a fixed duration

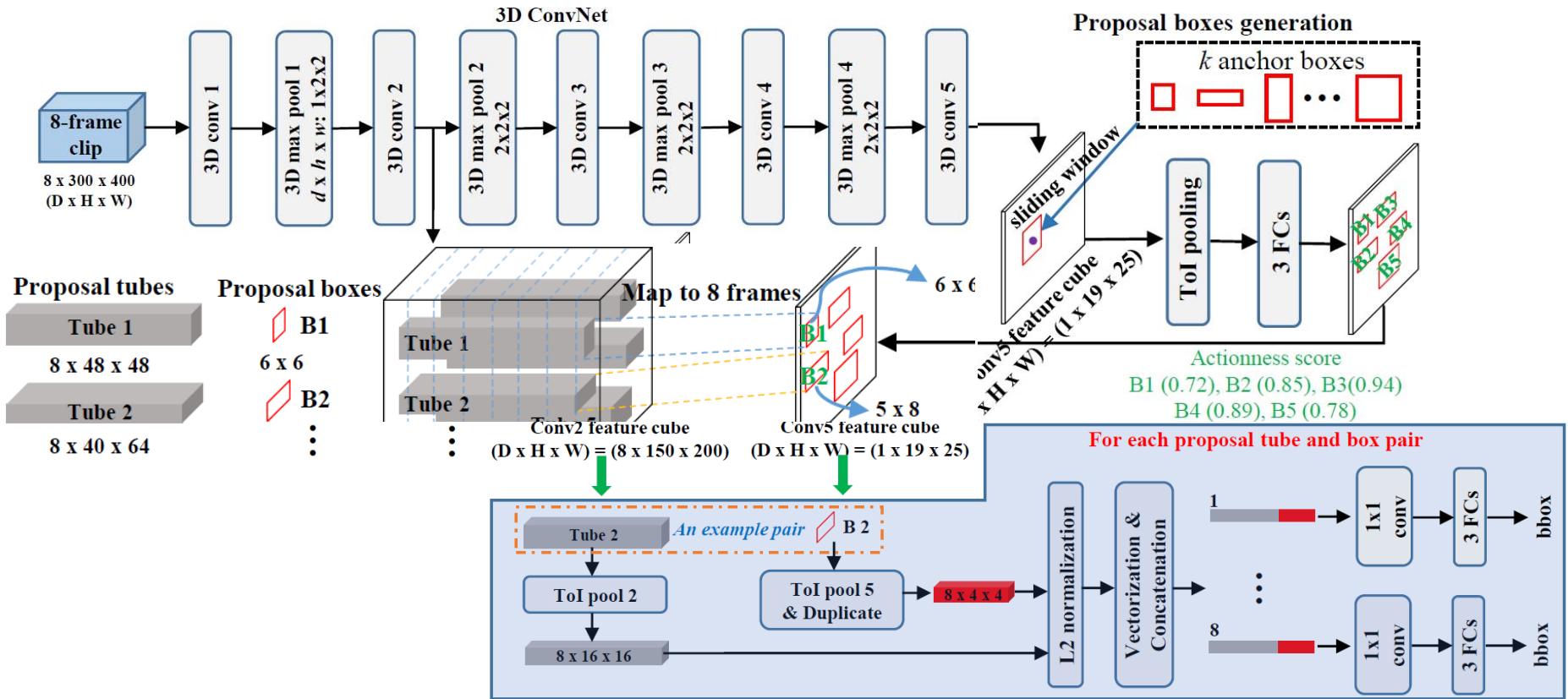
Tube of Interest Max Pooling(TOI)



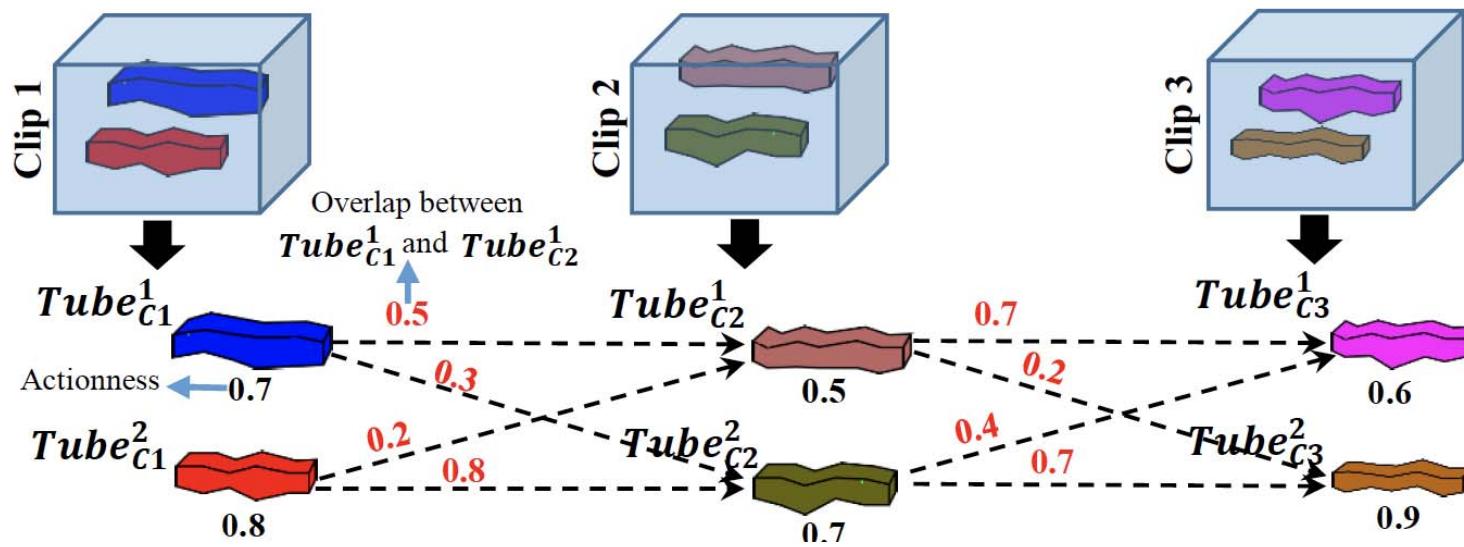
Tube of Interest Max Pooling(TOI)



Tube Proposal Network

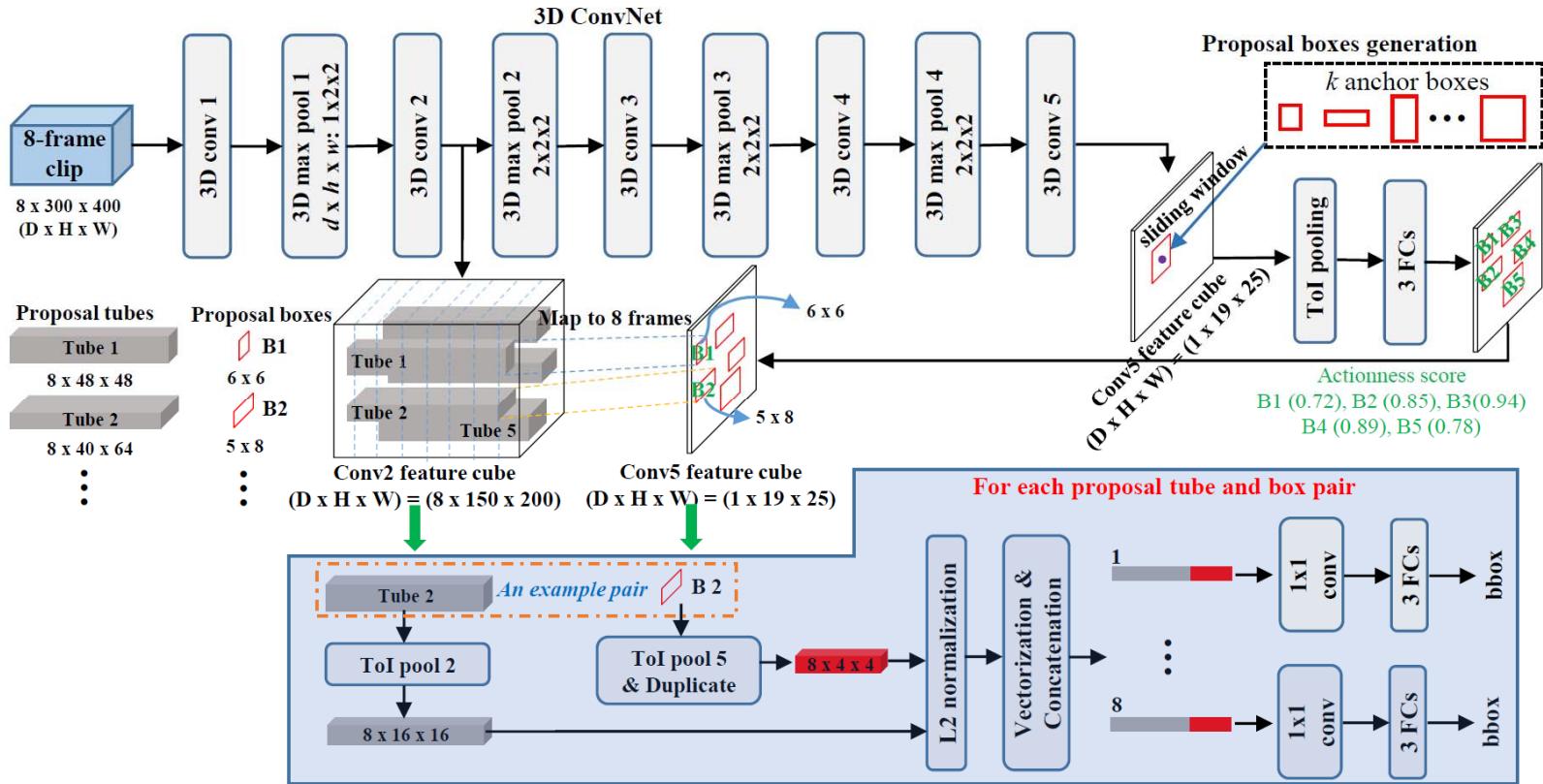


Proposal Linking



Network Details

name	kernel dims ($d \times h \times w$)	output dims ($C \times D \times H \times W$)
conv1	$3 \times 3 \times 3$	$64 \times 8 \times 300 \times 400$
max-pool1	$1 \times 2 \times 2$	$64 \times 8 \times 150 \times 200$
conv2	$3 \times 3 \times 3$	$128 \times 8 \times 150 \times 200$
max-pool2	$2 \times 2 \times 2$	$128 \times 4 \times 75 \times 100$
conv3a	$3 \times 3 \times 3$	$256 \times 4 \times 75 \times 100$
conv3b	$3 \times 3 \times 3$	$256 \times 4 \times 75 \times 100$
max-pool3	$2 \times 2 \times 2$	$256 \times 2 \times 38 \times 50$
conv4a	$3 \times 3 \times 3$	$512 \times 2 \times 38 \times 50$
conv4b	$3 \times 3 \times 3$	$512 \times 2 \times 38 \times 50$
max-pool4	$2 \times 2 \times 2$	$512 \times 1 \times 19 \times 25$
conv5a	$3 \times 3 \times 3$	$512 \times 1 \times 19 \times 25$
conv5b	$3 \times 3 \times 3$	$512 \times 1 \times 19 \times 25$
toi-pool2*	—	$128 \times 8 \times 8 \times 8$
toi-pool5	—	$512 \times 1 \times 4 \times 4$
1x1 conv	—	8192
fc6	—	4096
fc7	—	4096



Training Steps

- Initialize TPN based on the pre-trained C3D model
- Use the generated proposals to initialize recognition network
- Use the weights tuned by recognition network to update TPN
- Use tuned weights and proposals from TPN for final recognition network

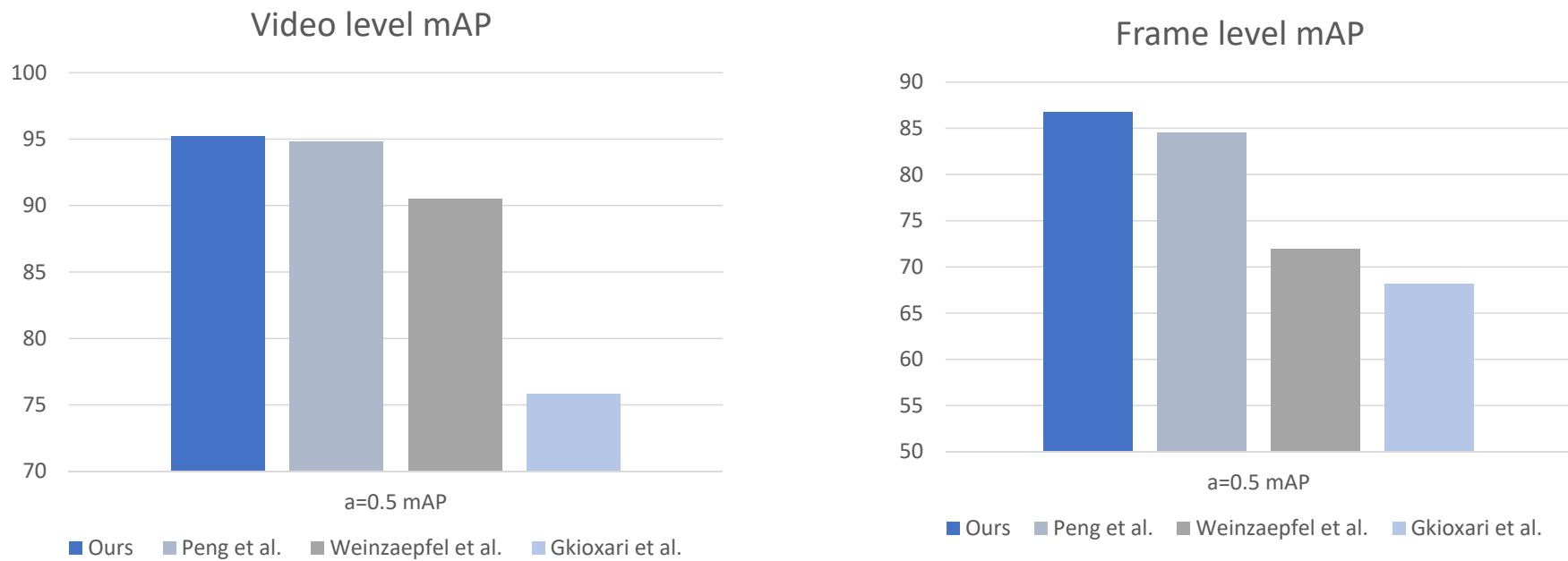
Implementation Details

- Clip : $400 \times 300 \times 8$ ($h \times w \times d$)
- Learning Rate: Initialized at 10^{-3} and
 - decreased to 10^{-4} after 60k iteration.
- Weight Decay: 0.0005
- Batch Size: 4

Negative Mining

- Untrimmed Videos contain positive and negative clips
- Initialize the TPN by using only positive clips.
- Apply the trained model on the whole training video (positive and negative clips)
- Select boxes in negative clips with highest scores as hard negatives
- In updating TPN procedure, we choose
 - 32 boxes, which have IoU with any ground truth greater than 0.7 as positive samples
 - Randomly pick another 16 samples as negative
 - Select 16 samples from hard negative pool as negative.

Experimental results for UCF-Sports



P. Weinzaepfel et al. *Learning to track for spatio-temporal action localization*, ICCV 2015

G. Gkioxari et al. *Finding action tubes*, CVPR 2015

X. Peng et al. Multi-region two-stream r-cnn for action detection, ECCV 2016

UCF-Sports -- Skip Pooling

Feature From	Frame mAP $\alpha = 0.5$	Frame mAP $\alpha = 0.2$	Video mAP $\alpha = 0.5$
C_5	74.9	91.6	77.9
$C_5 + C_1$	81.2	94.5	92.1
$C_5 + C_2$	86.7	95.2	94.8
$C_5 + C_3$	85.8	95.2	91.7
$C_5 + C_4$	77.6	91.3	81.2

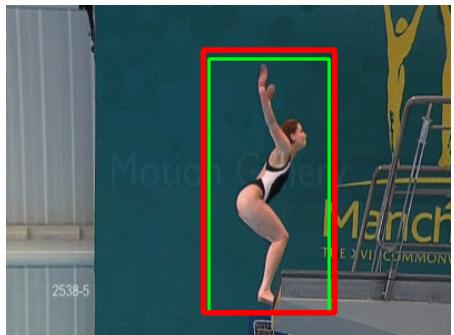
Detection Results for UCF-Sports



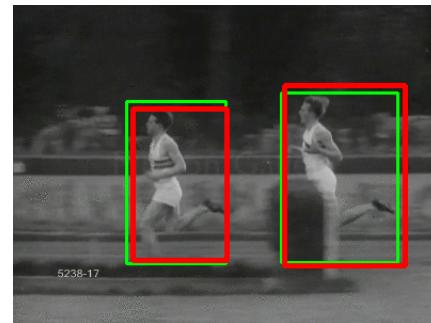
Horse Riding



Golf Swing



Diving



Running

Red: Our detection
Green: Ground Truth

UCF-Sports – Frame mAP

	Diving	Golf	Kick	Lifting	Riding	Run	Skate.	Swing	Swing Bench	Walk
Gkioxari et al.	75.8	69.3	54.6	99.1	89.6	54.9	29.8	88.7	74.5	44.7
Wenzaepfel et al.	60.71	77.55	65.26	100.0	99.53	52.6	47.14	88.88	62.86	64.44
Peng et al.	96.12	80.47	73.78	99.17	97.56	82.37	57.43	83.64	98.54	75.99
Ours	84.38	90.79	86.48	99.77	100.00	83.65	68.72	65.75	99.62	87.79

P. Weinzaepfel et al. *Learning to track for spatio-temporal action localization*, ICCV 2015

G. Gkioxari et al. *Finding action tubes*, CVPR 2015

X. Peng et al. Multi-region two-stream r-cnn for action detection, ECCV 2016

Action Detection for J-HMDB



BrushHair



Claping



Picking

Red: Our detection
Green: Ground Truth



Kicking



Golf



Climbing Stairs

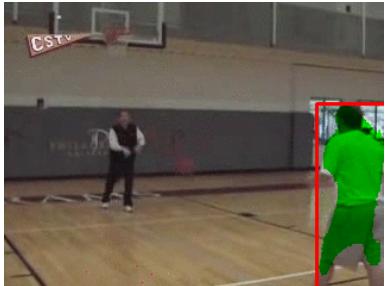
Action Detection for J-HMDB

	f.-mAP ($\alpha=0.5$)	v.-mAP ($\alpha=0.2$)	v.-mAP ($\alpha=0.5$)
Gkioxari et al.	36.2	-	53.3
Wenzaepfel et al.	45.8	63.1	60.7
Peng et al.	58.5	74.3	73.1
Ours w/o skip	47.9	66.9	58.6
Ours	84.38	90.79	86.48

- P. Weinzaepfel et al. *Learning to track for spatio-temporal action localization*, ICCV 2015
G. Gkioxari et al. *Finding action tubes*, CVPR 2015
X. Peng et al. Multi-region two-stream r-cnn for action detection, ECCV 2016

An End-to-end 3D Convolutional Neural Network for Action Detection and Segmentation in Videos

Action Segmentation Results on J-HMDB



Catch



Brush hair



Clap



Golf



Climb



Pick

Red: Bounding box detection
Green: Segmentation Map

Trained on JHMDB tested on UCF Sports



Red: Bounding box detection
Green: Segmentation map

Densely Annotated Video Segmentation (DAVIS'16)

- Video Object Segmentation Dataset
- 50 Videos
- 3455 Frames
- 480P: 854X480

Video Object Segmentation on Davis'16



Swan



HorseJump-High

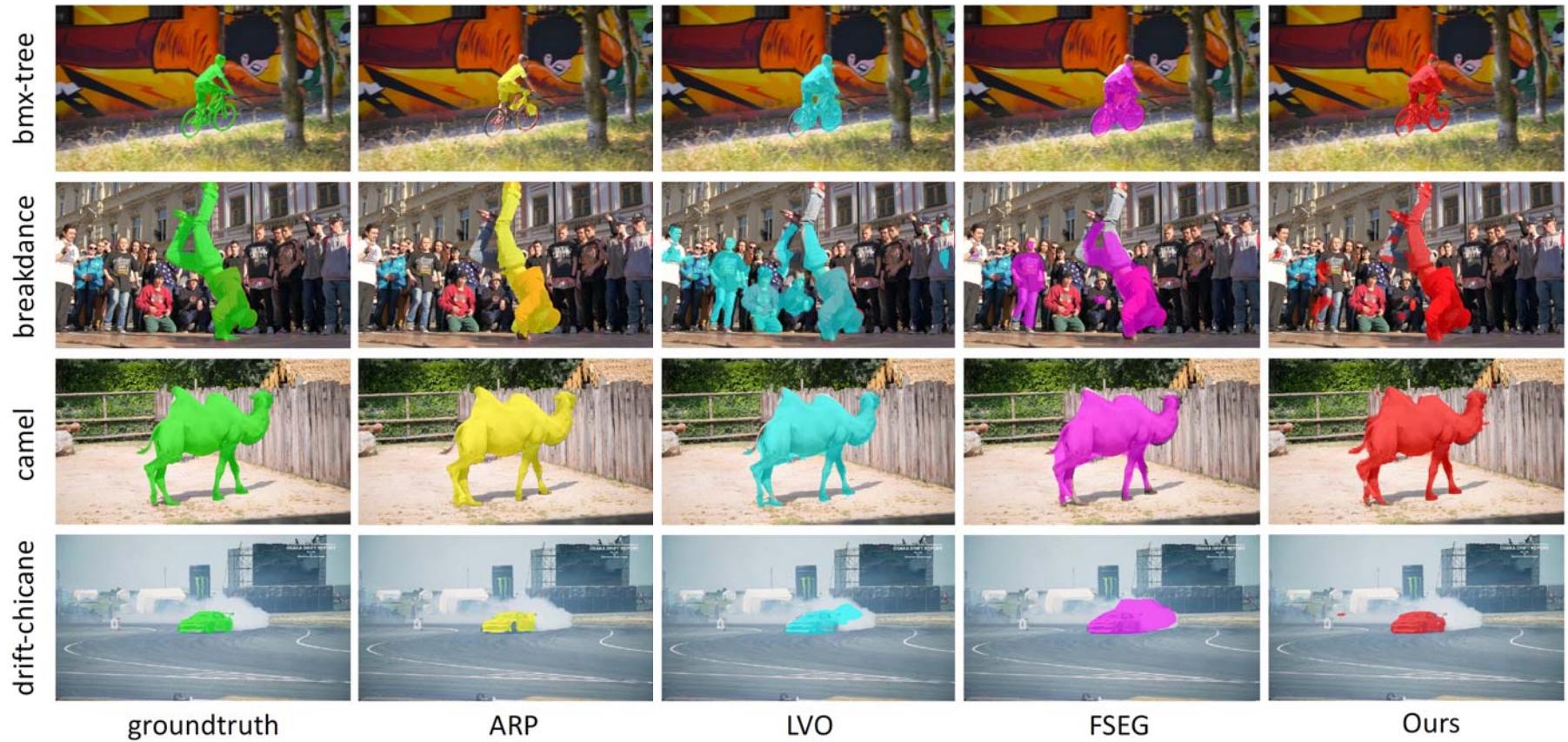


Libby

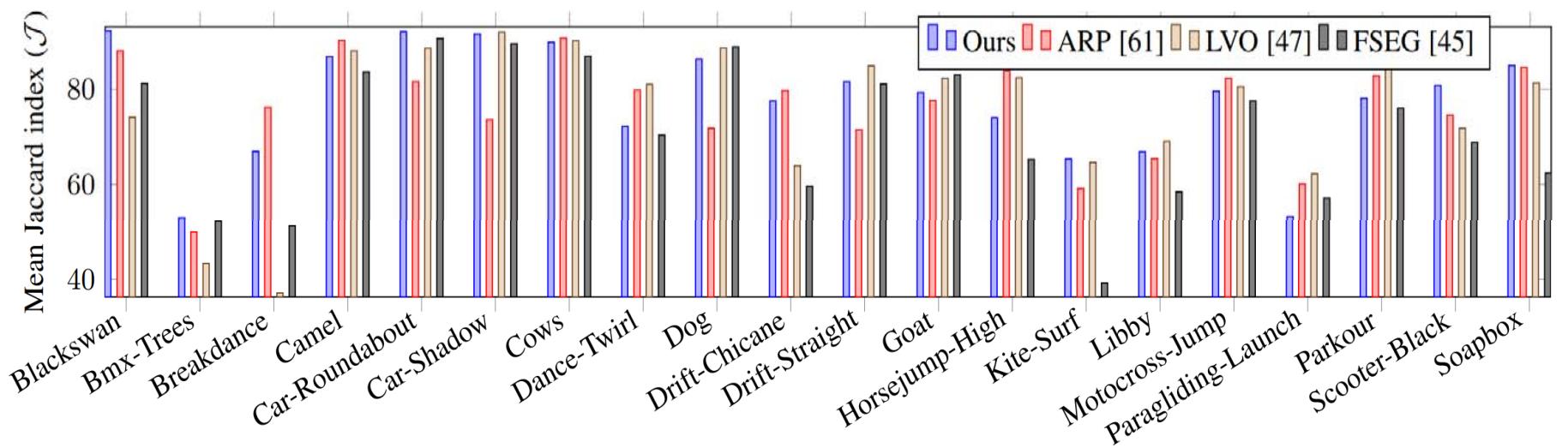


Car-roundabout

Comparison of Different Approaches



Comparison of Mean Jaccard Index on DAVIS'16



Qualitative Results on DAVIS'16

	Measure	ARP	FSEG	LMP	FST	CUT	Ours
\mathcal{J}	Mean↑	76.2	70.7	70.0	55.8	55.2	77.6
	Recall↑	91.1	83.5	85.0	64.9	57.5	95.2
	Decay↓	7.0	1.5	1.3	0.0	2.2	2.3
\mathcal{F}	Mean↑	70.6	65.3	65.9	51.1	55.2	75.5
	Recall↑	83.5	73.8	79.2	51.6	61.0	94.7
	Decay↓	7.9	1.8	2.5	2.9	3.4	4.9
\mathcal{T}	Mean↓	39.3	32.8	57.2	36.6	27.7	22.0

Action Detection on THUMOS'13

	f.-mAP ($\alpha=0.5$)	v.-mAP ($\alpha=0.05$)	v.-mAP ($\alpha=0.1$)	v.-mAP ($\alpha=0.2$)	v.-mAP ($\alpha=0.3$)
Weinzaepfel et al.	35.84	54.3	51.7	46.8	37.8
Peng et al.	39.63	54.5	50.4	42.3	32.7
Ours	41.37	54.7	51.3	47.1	39.2

P. Weinzaepfel et al. *Learning to track for spatio-temporal action localization*, ICCV 2015

X. Peng et al. Multi-region two-stream r-cnn for action detection, ECCV 2016

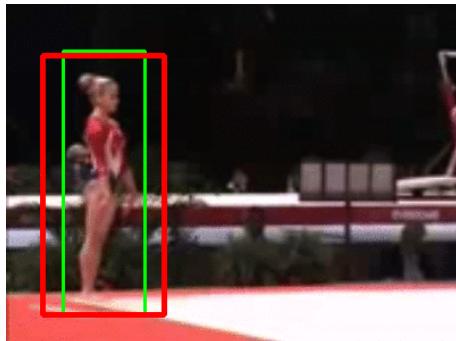
Experiments for THUMOS'13



Biking



HorseRiding



FloorGymnastics



CliffDiving

Red: Our detection
Green: Ground Truth

T-CNN for Action Detection in Videos

Rui Hou, Chen Chen, Mubarak Shah

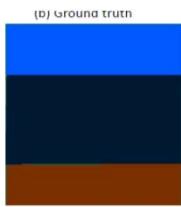
ICCV-2017

<http://crcv.ucf.edu/papers/iccv17/T-CNN-camera-ready.pdf>

Contents



- road
- sea
- sky
- sand



Sematic Segmentation



Facial Attributes Detection



Human Re-Identification



Target Detection in WAMI



Anomaly Detection

Diving



Human Action Localization



Single Blank:

He ___ up the steps of the stand and away. (Runs)

Video Fill In The Blank



Reading The Mind

Video Fill In the Blank using LR/RL LSTMs with Spatial-Temporal Attentions

Amir Mazaheri, Dong Zhang, and Mubarak Shah

ICCV 2017

<http://crcv.ucf.edu/papers/iccv17/PID4929115.pdf>

Problem Definition

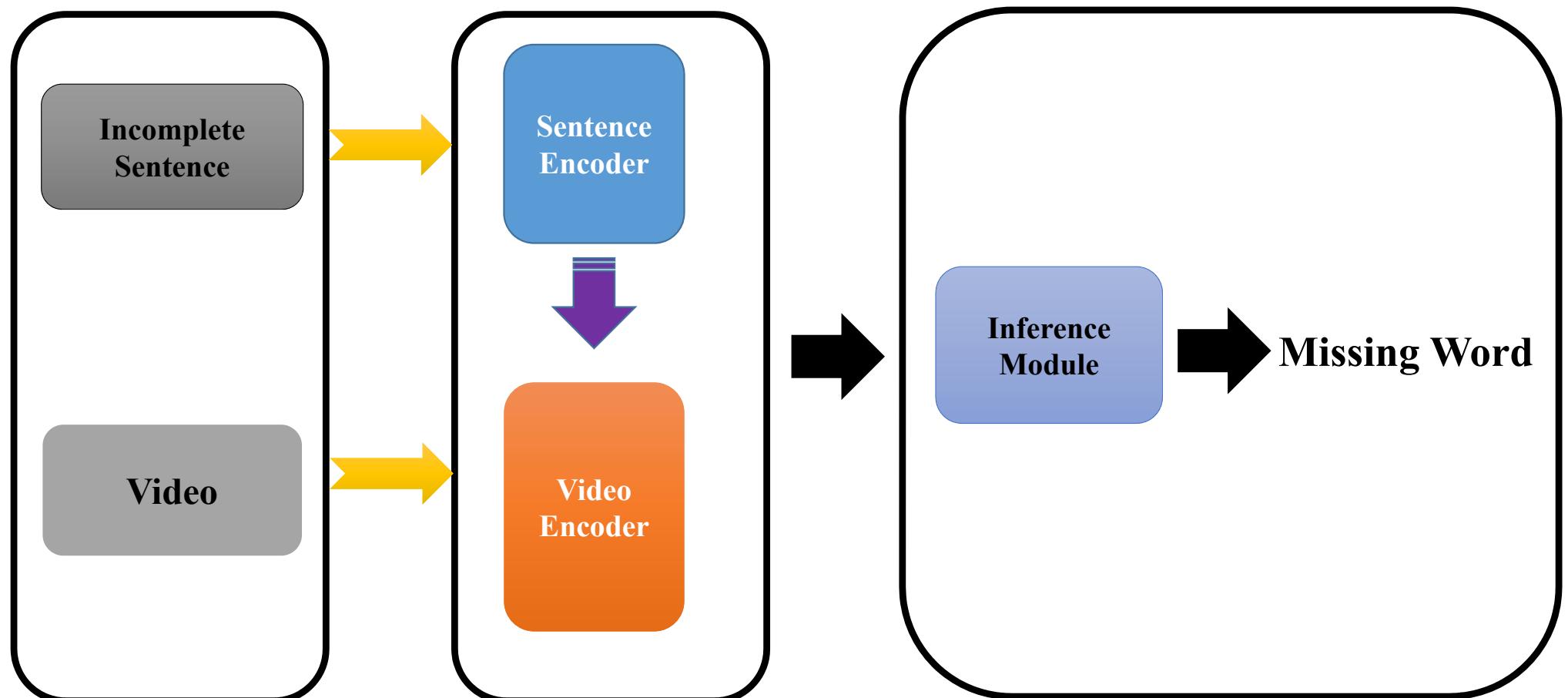
- **Input:**
 - A video
 - An Incomplete Sentence
- **Output:**
 - Missing word(s)



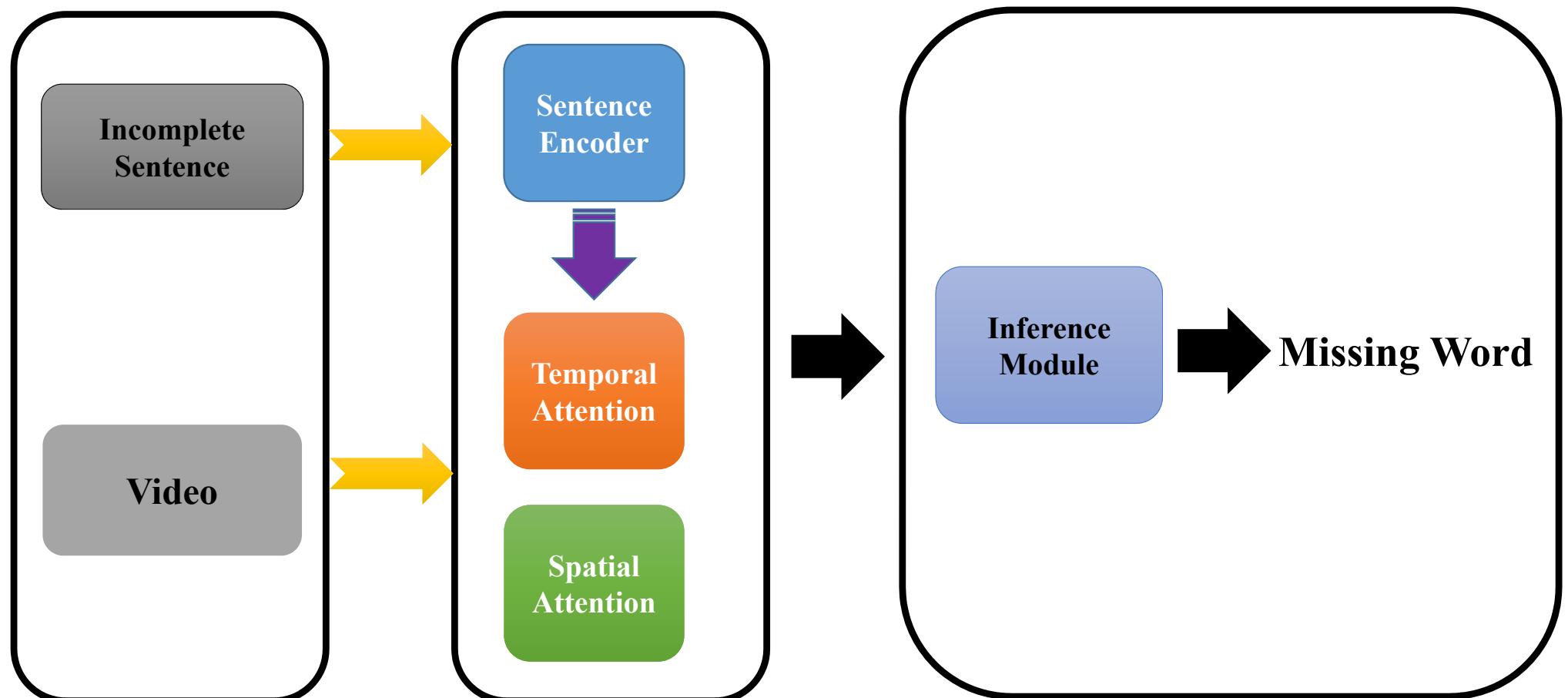
Single Blank:

He ___ up the steps of the stand and away.

Block Diagram



Block Diagram



Sentence Encoder



Someone stops his _____ and kisses her on the head

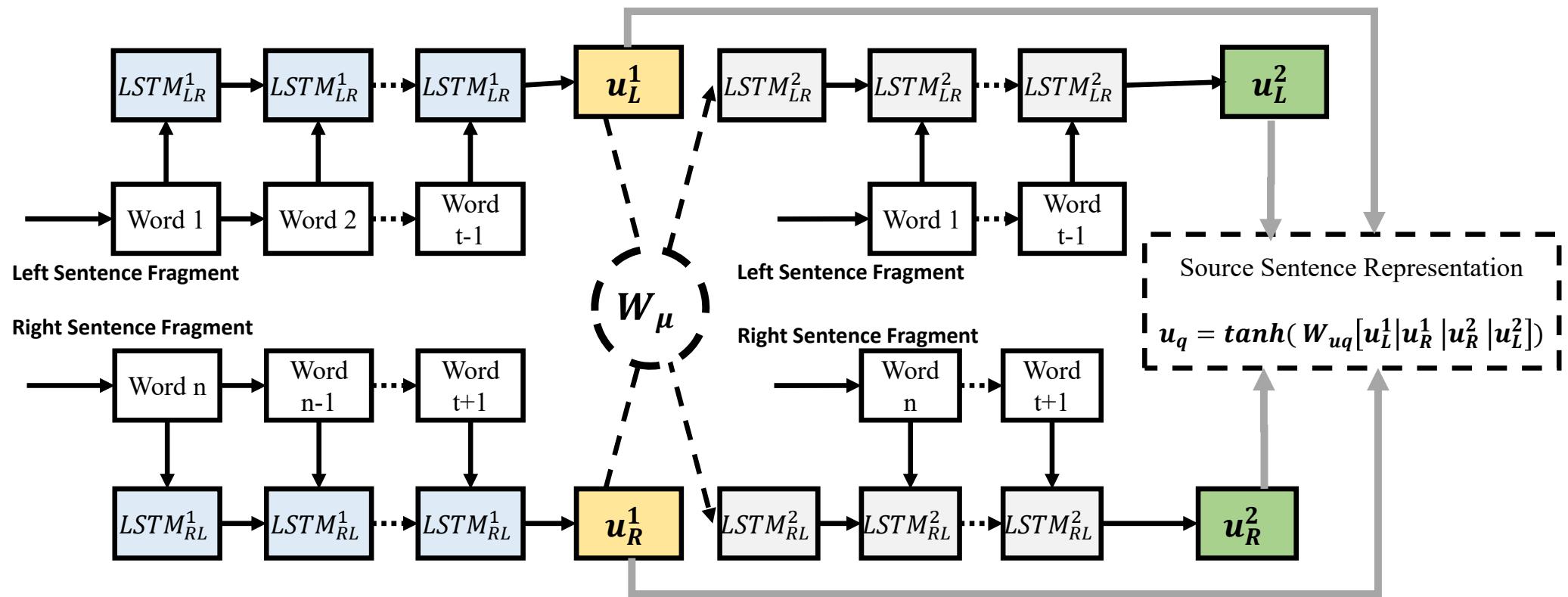


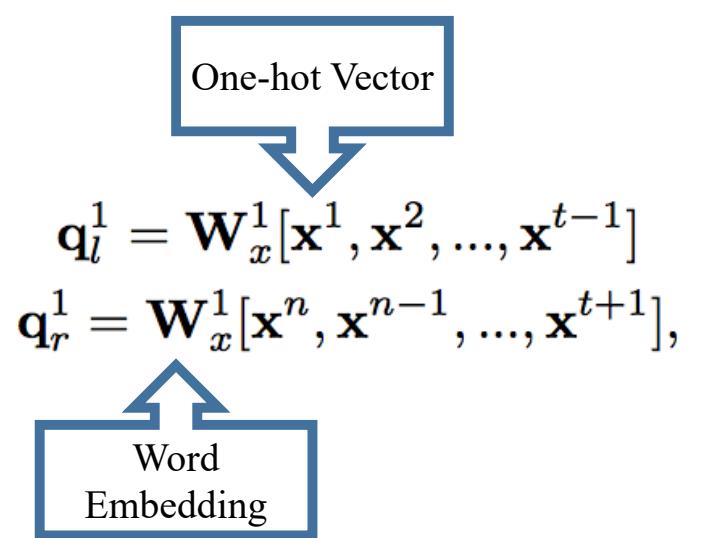
Encoding by LSTM



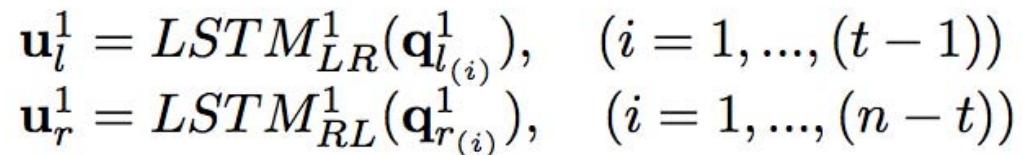
Encoding by LSTM

Sentence Encoder



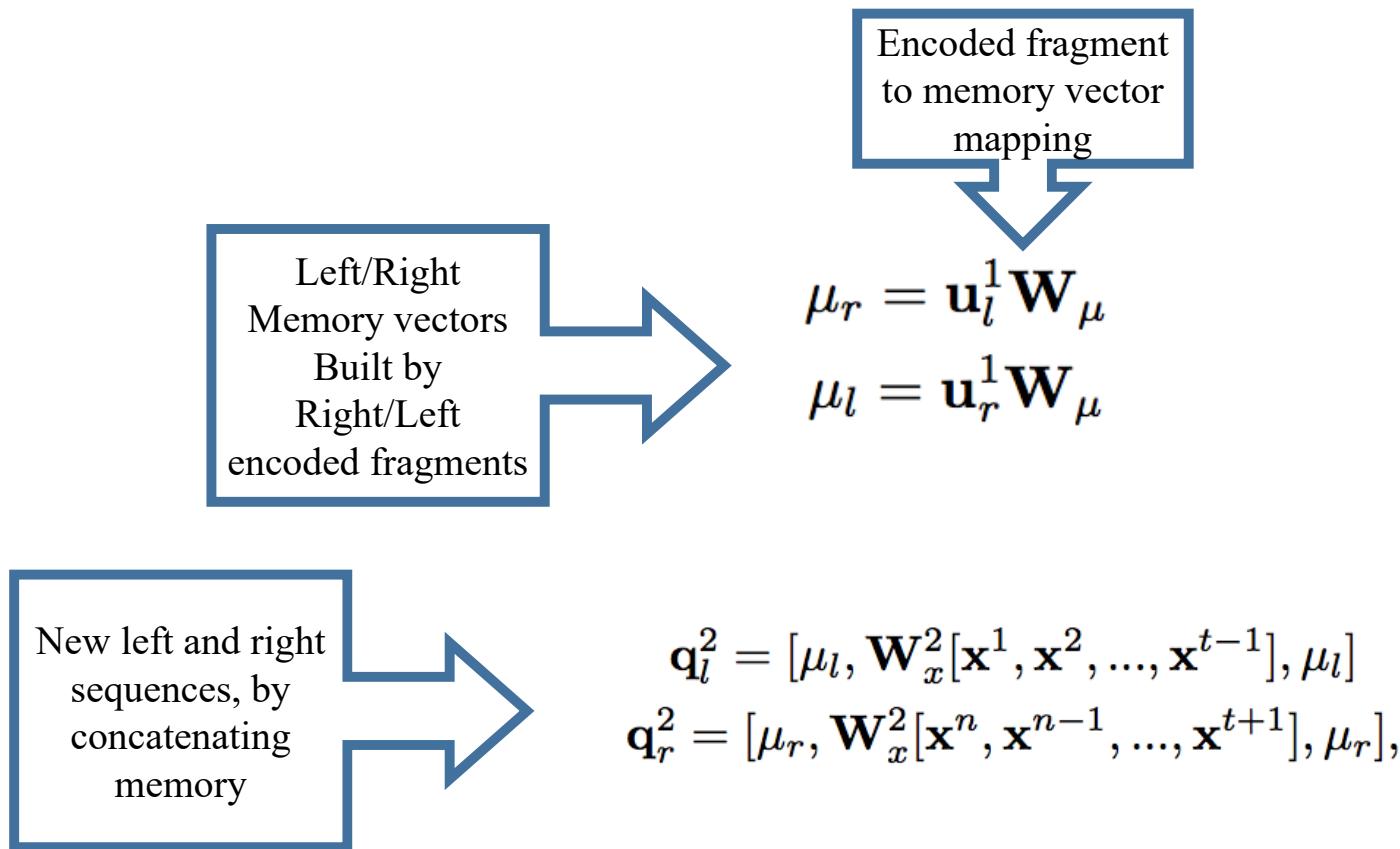


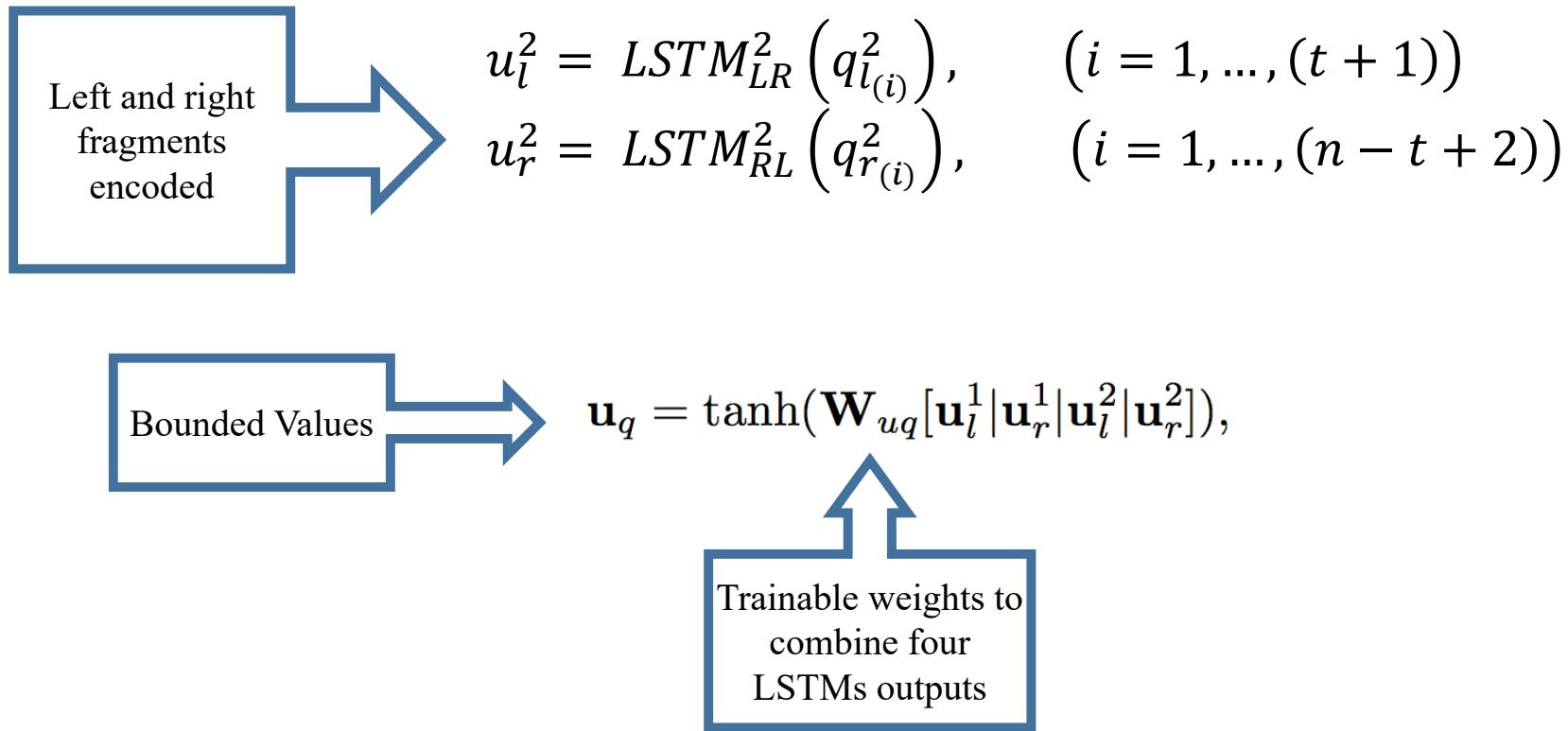
Left and right
Encoded
Fragments



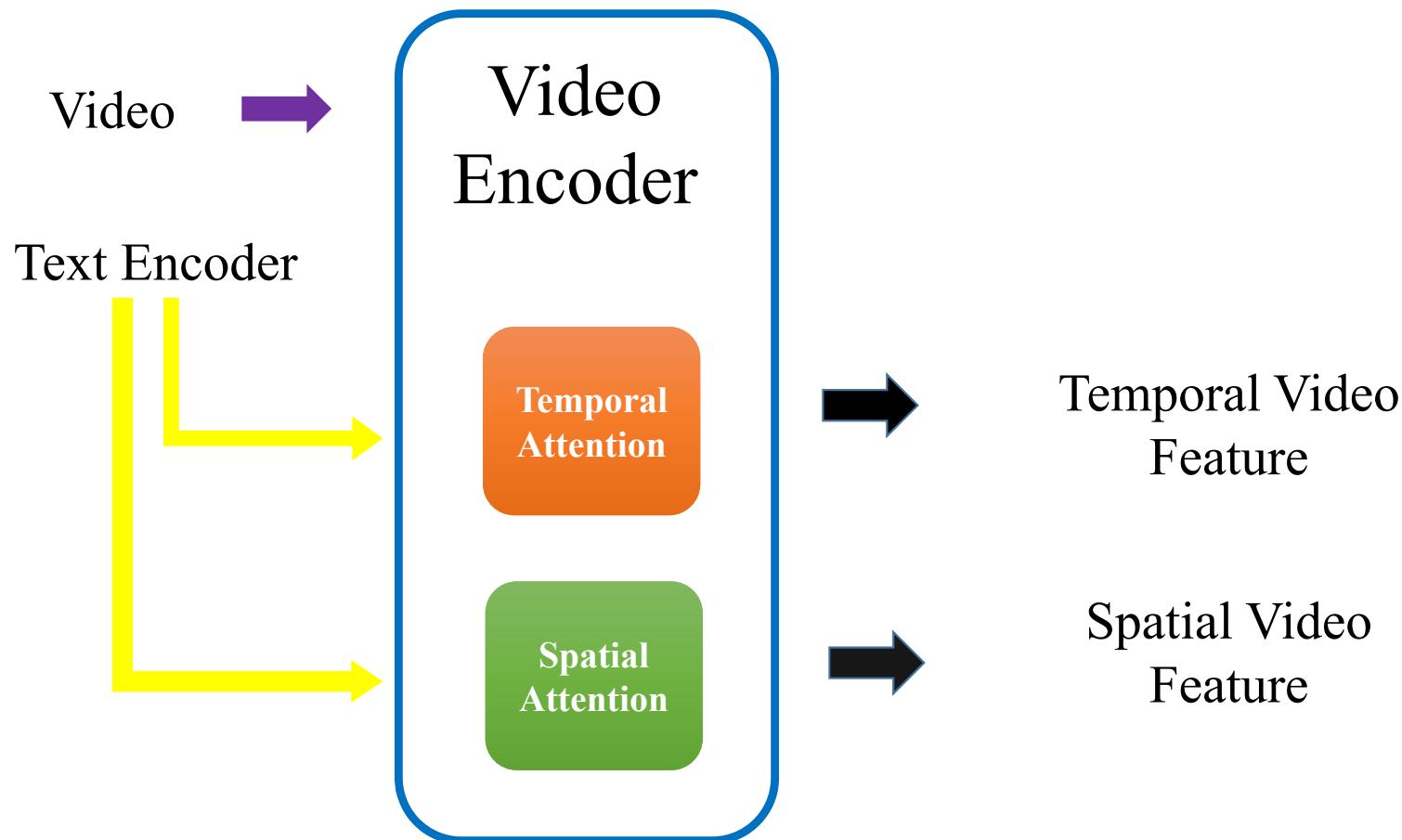
A large blue arrow points from the "Left and right Encoded Fragments" box to the first term of the equation $\mathbf{u}_l^1 = LSTM_{LR}^1(\mathbf{q}_{l(i)}^1)$.

$$\mathbf{u}_l^1 = LSTM_{LR}^1(\mathbf{q}_{l(i)}^1), \quad (i = 1, \dots, (t - 1))$$
$$\mathbf{u}_r^1 = LSTM_{RL}^1(\mathbf{q}_{r(i)}^1), \quad (i = 1, \dots, (n - t))$$



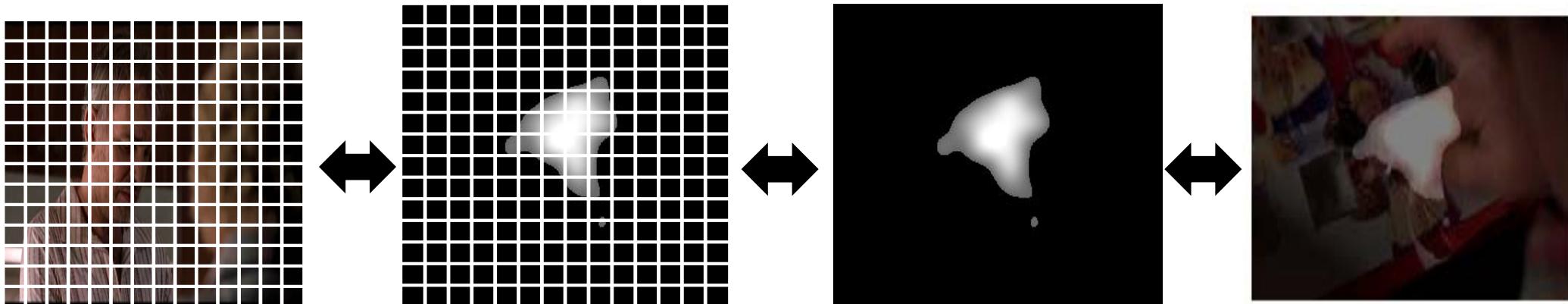


Video Encoding



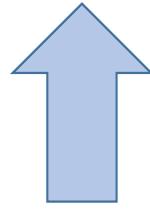
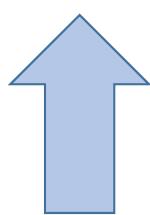
Spatial Attention Module

Input video Regions Scores Attention map



Someone watches out of the corner of his eye as the kid finds a cheap _____ inside.
(Answer: Sweet)

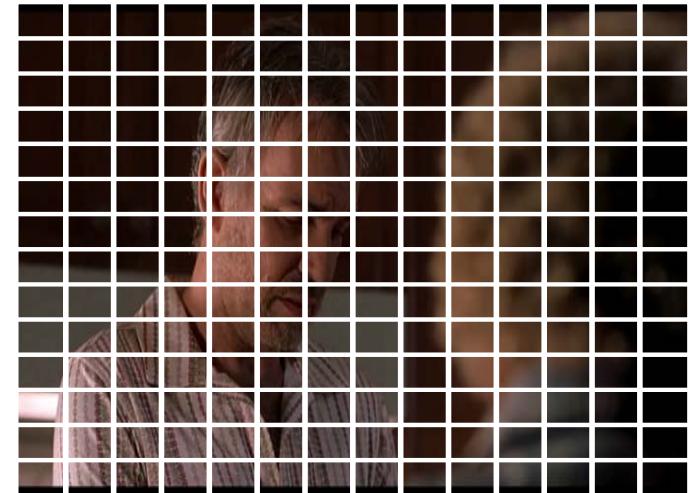
Temporal Attention Module



Someone watches out of the corner of his eye as the kid finds a cheap _____ inside. (Answer: Sweet)

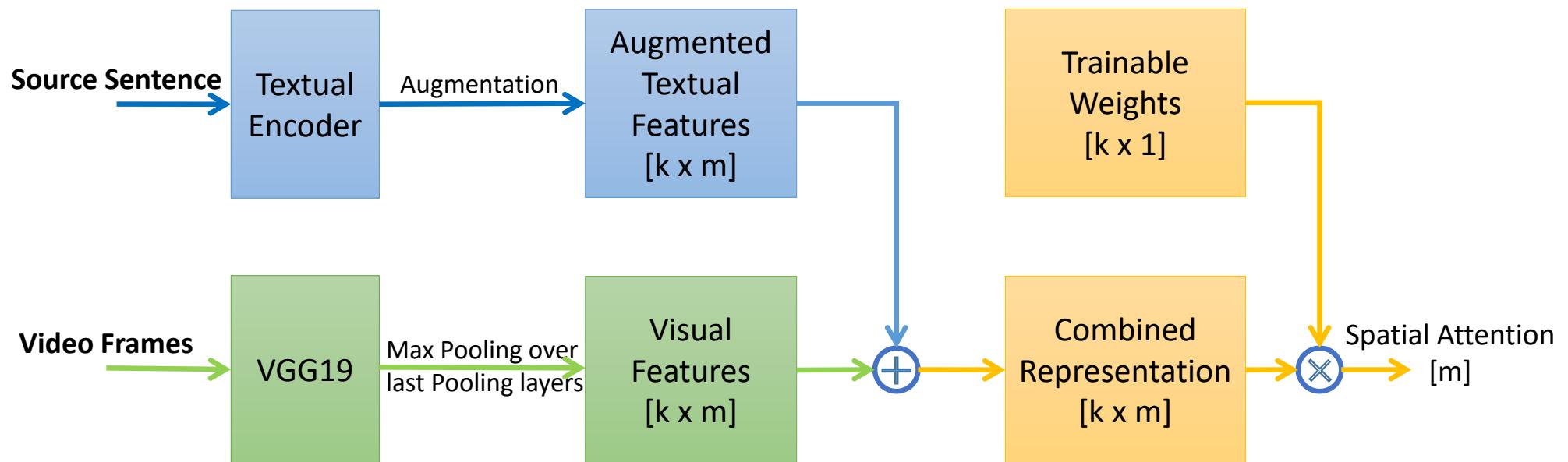
Spatial Attention Module

- Which **regions** of the frames to look?
- VGG-19 pre-trained network **last convolution** layer
- $14 \times 14 \times 512$
 - **196** regions and **512** dimensional feature vectors
- Each region corresponds to a **32 × 32 pixels** patch



Someone watches out of the corner of his eye as the kid finds a cheap _____ inside. (Answer: **Sweet**)

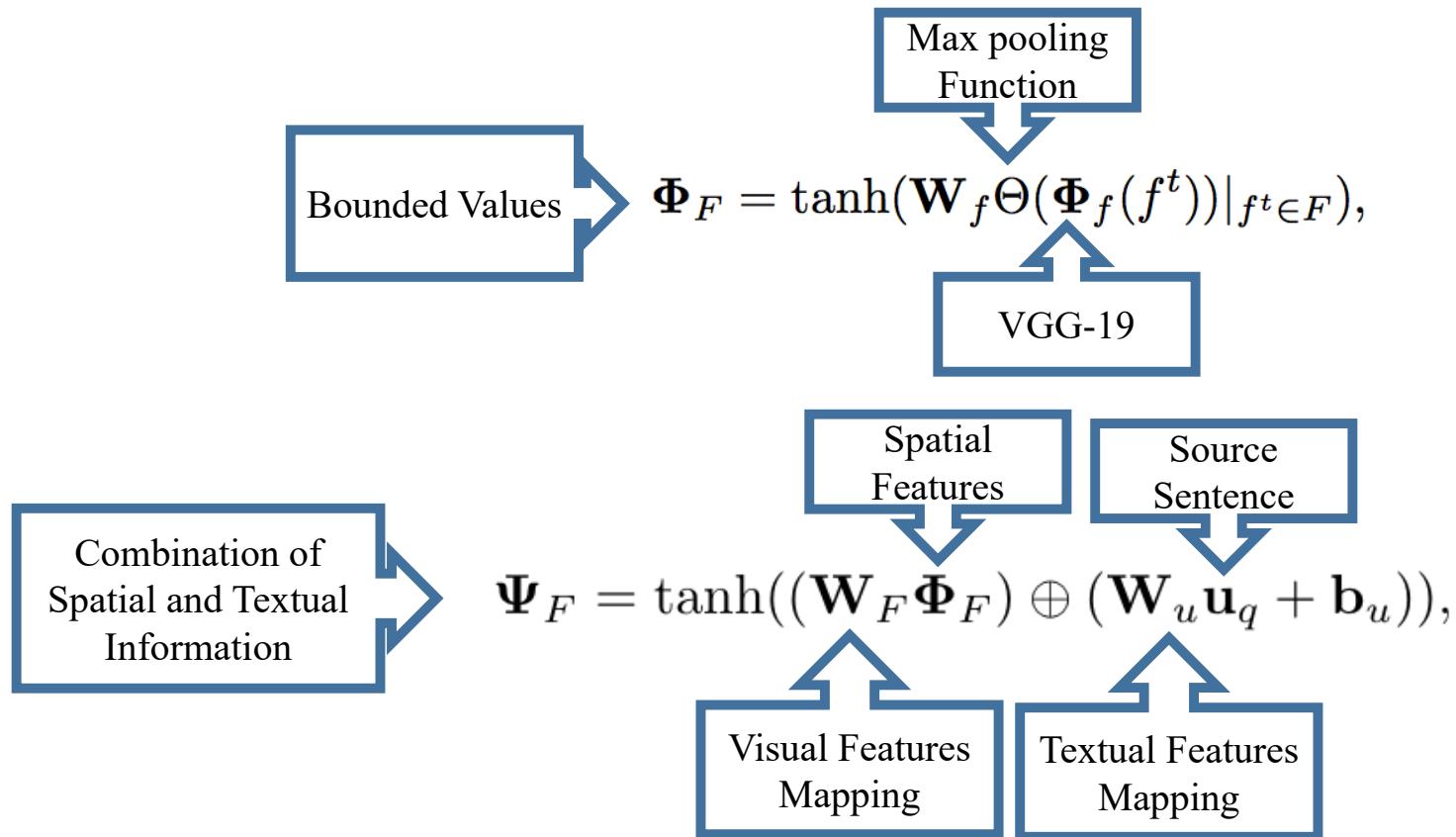
Spatial Attention – Block Diagram



m : Number of regions (196)

k : Textual and visual joint representation length

Spatial Attention Networks



Spatial Attention Networks

$$\mathbf{p}_{sp} = \text{softmax}(\Psi_F^T \mathbf{w}_{sp}),$$

Bounded Value

$$\mathbf{u}_{sp} = \Phi_F \mathbf{p}_{sp},$$

Attention as probability

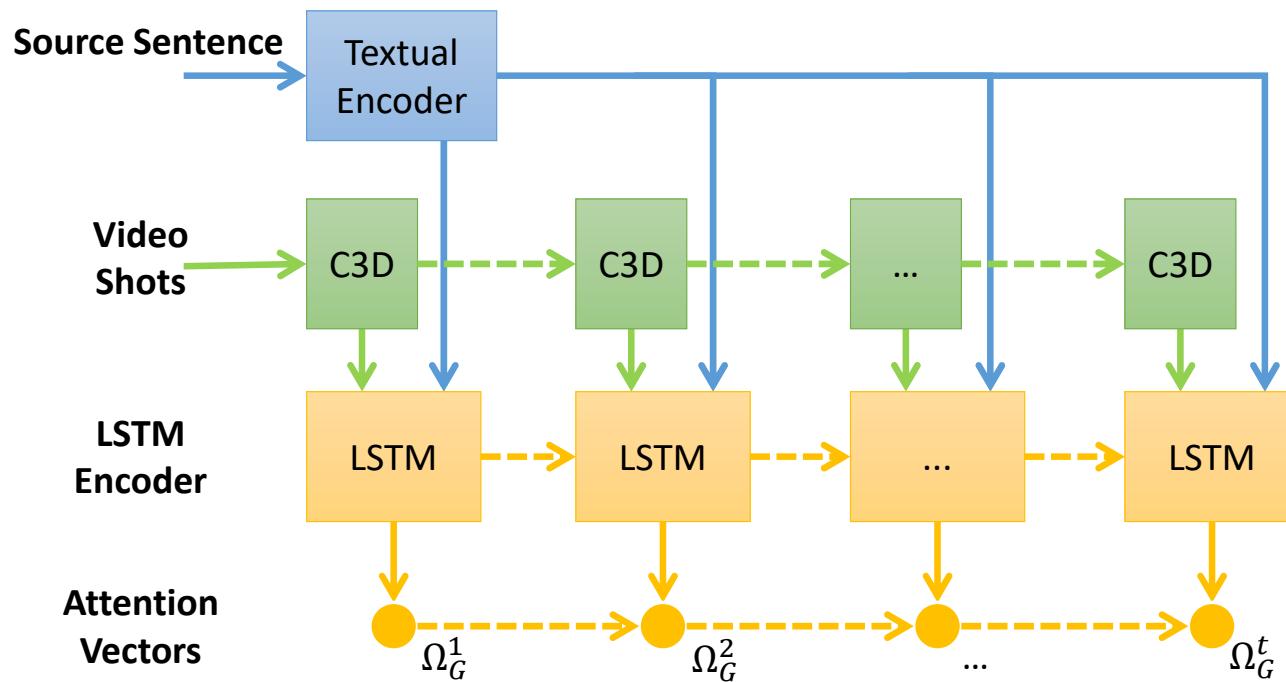
Weighted Average

Temporal Attention Model

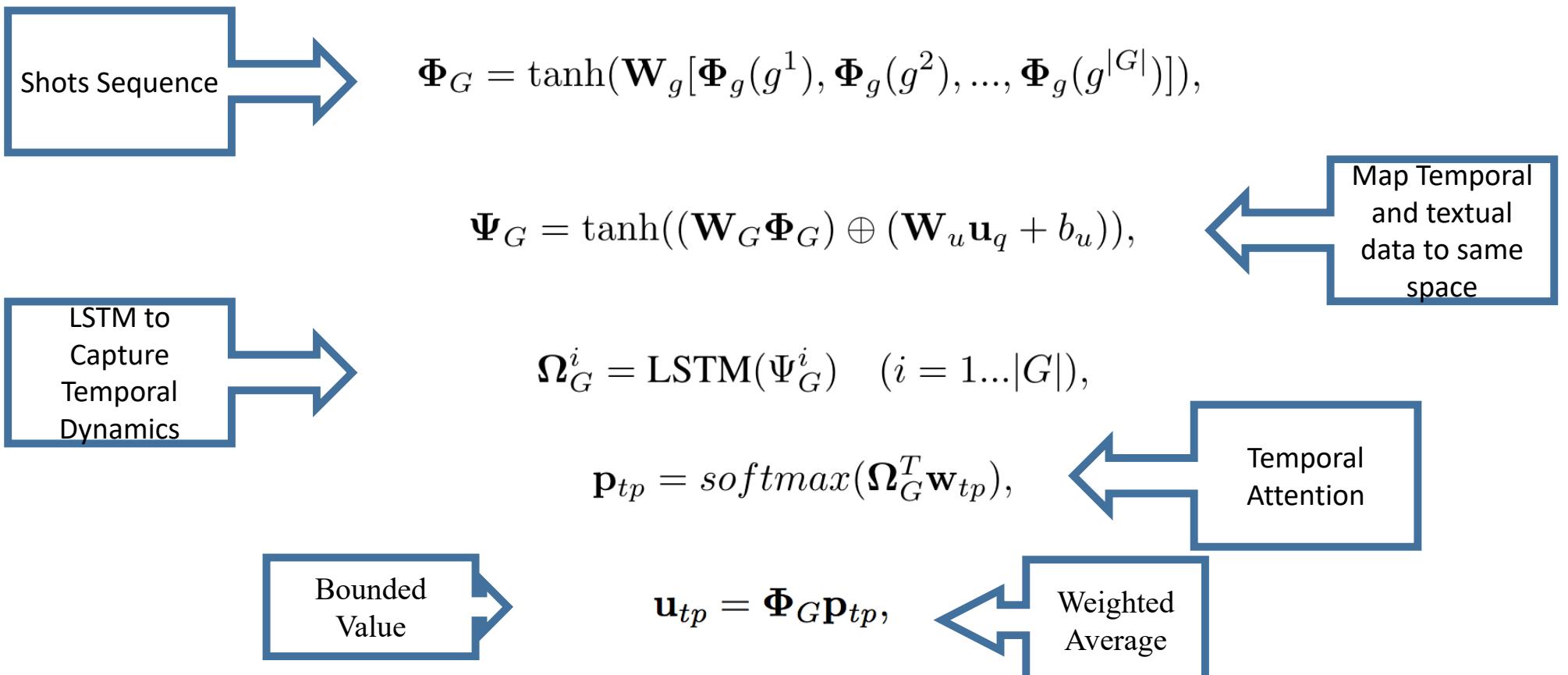
- C3D Features for each 16 frames
 - C3D is a temporal-convolutional network
 - Pre-trained on SPORT1M dataset
 - Encodes short shots (16 frames)
- Videos are longer than 16 frames
 - Sequence of 16 frames shots
- Which shot is more important?



Temporal Attention Network



Temporal Attention Network



Inference Module

$$\mathbf{u} = [\mathbf{u}_q + \mathbf{u}_{sp} + \mathbf{u}_{tp}],$$

All bounded
Same
dimension



$$P_{blank} = softmax(\mathbf{W}_{blank}\mathbf{u}),$$

$$\hat{b} = \arg \max_{b \in \beta} P_{blank}(b)$$

Experiments

- Large Scale Movie Description Challenge 2016 (LSMDC16)
- About **360,000** samples
- Dataset is built upon movies audio descriptions
 - Complicated sentences
 - Large dictionary of words (21,000)
- Videos vary in length
 - 2 – 60 seconds

Training

- End to End training
 - All modules trained together
- Adagrad Optimizer
 - Any adaptive solver
- Cross-categorical Loss Function
- Batch-normalization before all non-linearity
- 8-10 hours training time on TitanX GPU

Quantitative Results

- Human Accuracy ~ 0.68
- Text Only Methods (Blind Test)
- Video Only (Lazy Student)
- Video + Text Methods
- Our Method

Method	Accuracy
Text Only	
Random Guess	0.006
LSTM Left Sentence	0.155
LSTM Right Sentence	0.165
BiLSTM	0.320
Our Sentence Encoding	0.367
<i>Human [27]</i>	0.302

Qualitative Results



Someone grabs her arm, pulls her close and
_____ her a lingering kiss.

Answer: gives

GT: gives



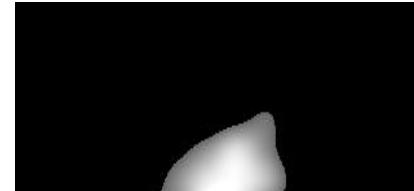
Qualitative Results



Someone watches out of the corner of his eye as the kid finds a cheap _____ inside.

Answer: Sweet

GT: Sweet



Qualitative Results



Someone defensively grabs a picture frame and presses ____ back to a wall by a white-trim doorway.

Answer: her

GT: her



Qualitative Results

Temporal Attention



Someone grabs her arm, pulls her close and ____ her a lingering kiss.

Spatial Attention

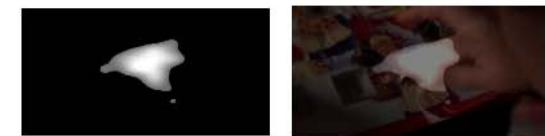


GT: gives

Ours: **gives**



Someone watches out of the corner of his eye as the kid finds a cheap ____ inside.



GT: sweet

Ours: **sweet**



Someone defensively grabs a picture frame and presses ____ back to a wall by a white-trim doorway



GT: her

Ours: **her**

Qualitative Results - Failures



Someone stops his _____ and kisses her on the head.



GT: daughter

Ours: **jacket**

Fill in Multiple Blanks

- More than one blank in sentence
 - More applications
 - Higher level of alignment between text and video
- More Challenging problem !
 - Each left and right fragments can be fragmented by themselves



Someone _____ his serious _____ from the contract.

	Left Fragment	Right Fragment
Blank 1	Someone	His serious
Blank 2	His serious	From the contract

Subdivision(cutting) method



	Left Fragment	Right Fragment
Blank 1	Someone	His serious from the contract
Blank 2	Someone his serious	From the contract

Masking method



LR/RL LSTMs with Spatial and Temporal Attention Models

Blank 1 = Lifts

Blank 2 = Gaze

Results

Method	Accuracy
Baselines	
Random Guess	0.006
Left LSTM (Masking)	0.104
Bi-LSTM (Masking)	0.156
2Videos + Textual (Subdivision) [14]	0.136
2Videos + Textual (Masking) [14]	0.177



Someone _____ and _____ his _____.

Cutting:

2Videos – Textual Encoding (**turns, faces, gaze**)
Our Model (**smiles, raises, wife**)

Masking:

2Videos – Textual Encoding (**smiles, shakes, gaze**)
Our Model (**smiles, shakes, head**)
Ground Truth: (**smiles, shakes, head**)



She _____ off _____ shoes and them aside.

Cutting:

2Videos – Textual Encoding	(walks, her, sets)
Our Model	(tosses, her, sets)

Masking:

2Videos – Textual Encoding	(takes, her, sets)
Our Model	(takes, her, tosses)
Ground Truth:	(takes, her, flings)

Conclusion

- Discussed the VFIB problem
- Novel sequence encoder for fragmented inputs
- Detailed formulation for spatial and temporal attention
- Quantitative and Qualitative Results
- Multiple Blanks cases

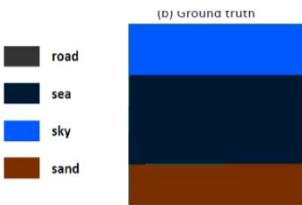
Video Fill In the Blank using LR/RL LSTMs with Spatial-Temporal Attentions

Amir Mazaheri, Dong Zhang, and Mubarak Shah

ICCV 2017

<http://crcv.ucf.edu/papers/iccv17/PID4929115.pdf>

Contents



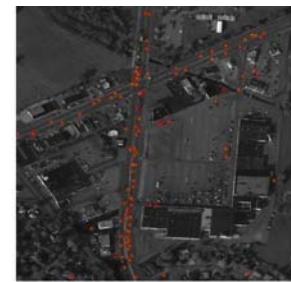
Sematic Segmentation



Facial Attributes Detection



Human Re-Identification



Target Detection in WAMI



Anomaly Detection

Diving



Human Action Localization



Single Blank:

He ___ up the steps of the stand and away. (Runs)

Video Fill In The Blank



Reading The Mind



UNIVERSITÀ
degli STUDI
di CATANIA



Deep Learning Human Mind for Automated Visual Classification

Concetto Spampinato, Simone Palazzo, Isaak
Kavasidis, Daniela Giordano

PeRCeVe Lab, University of Catania, Italy

Nasim Souly, Mubarak Shah

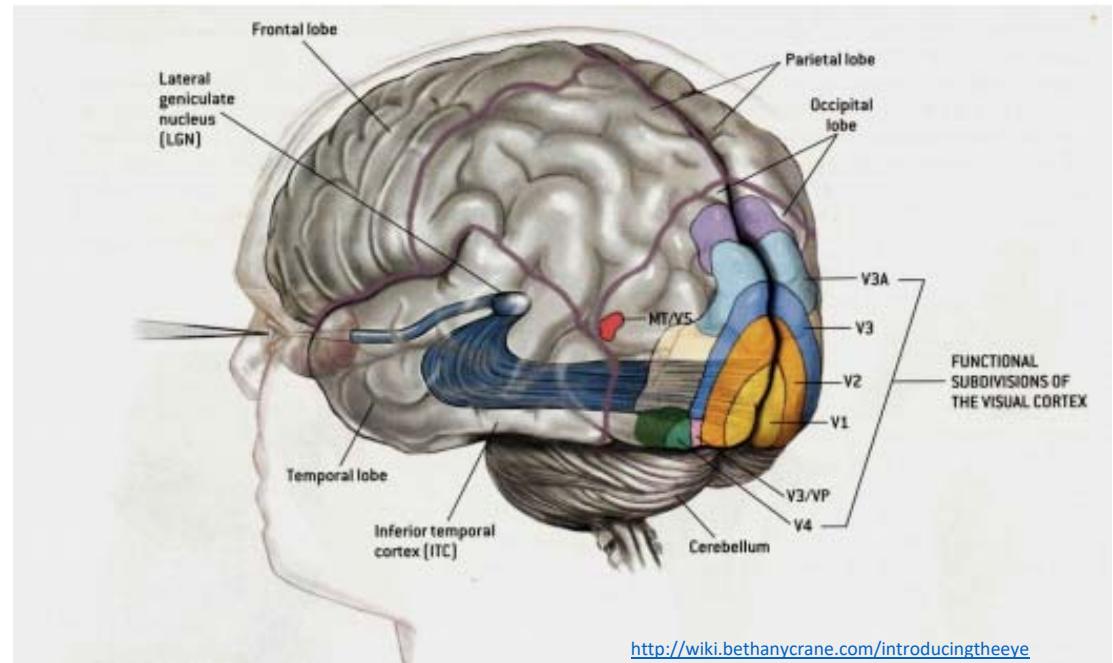
CRCV, University of Central Florida, USA

CVPR 2017

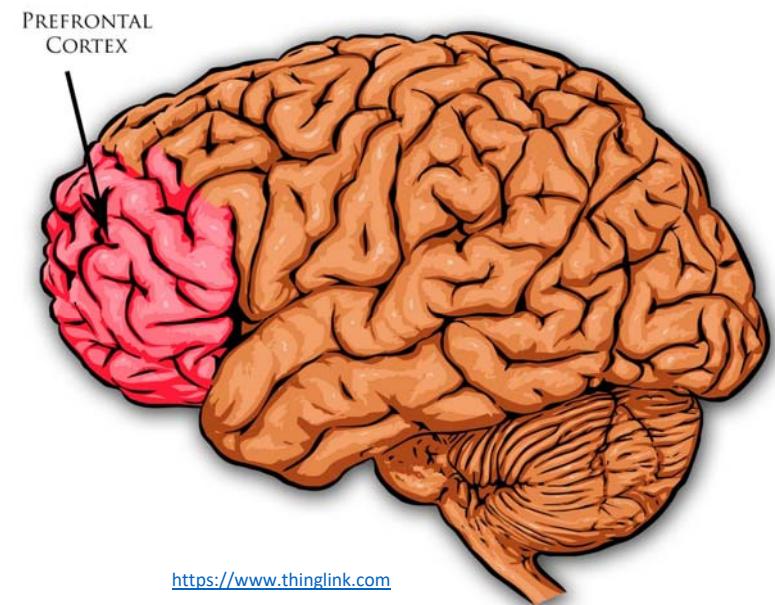
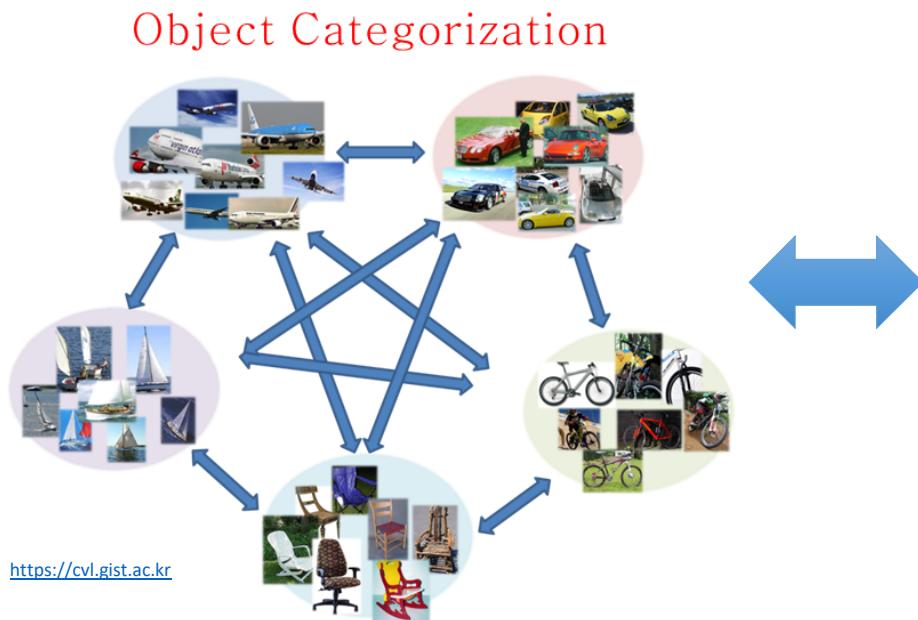
http://crcv.ucf.edu/papers/cvpr2017/cvpr_eeg_gen_2017_camera_ready.pdf

Motivation

- CNNs emulate human visual cortex
- What do CNNs miss?



- Visual classification in humans stands at the interface between perception (visual cortex) and conception (cognitive processes)



Harnessing visual-cognitive factors for visual categorization



Learning a feature space directly from human brain activity (EEG)



How?



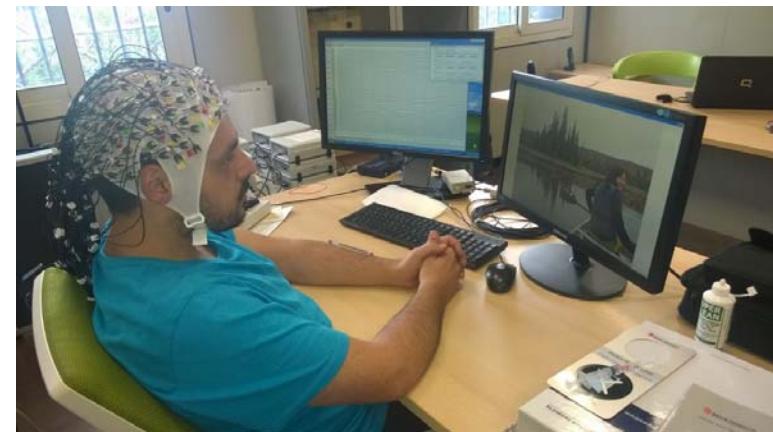
Brain activity in EEG recordings contain information about visual object categories [1, 2].

[1] T. Carlson, D. A. Tovar, A. Alink, and N. Kriegeskorte. Representational dynamics of object vision: the first 1000 ms. *Journal of Vision*, 13(10), 2013.

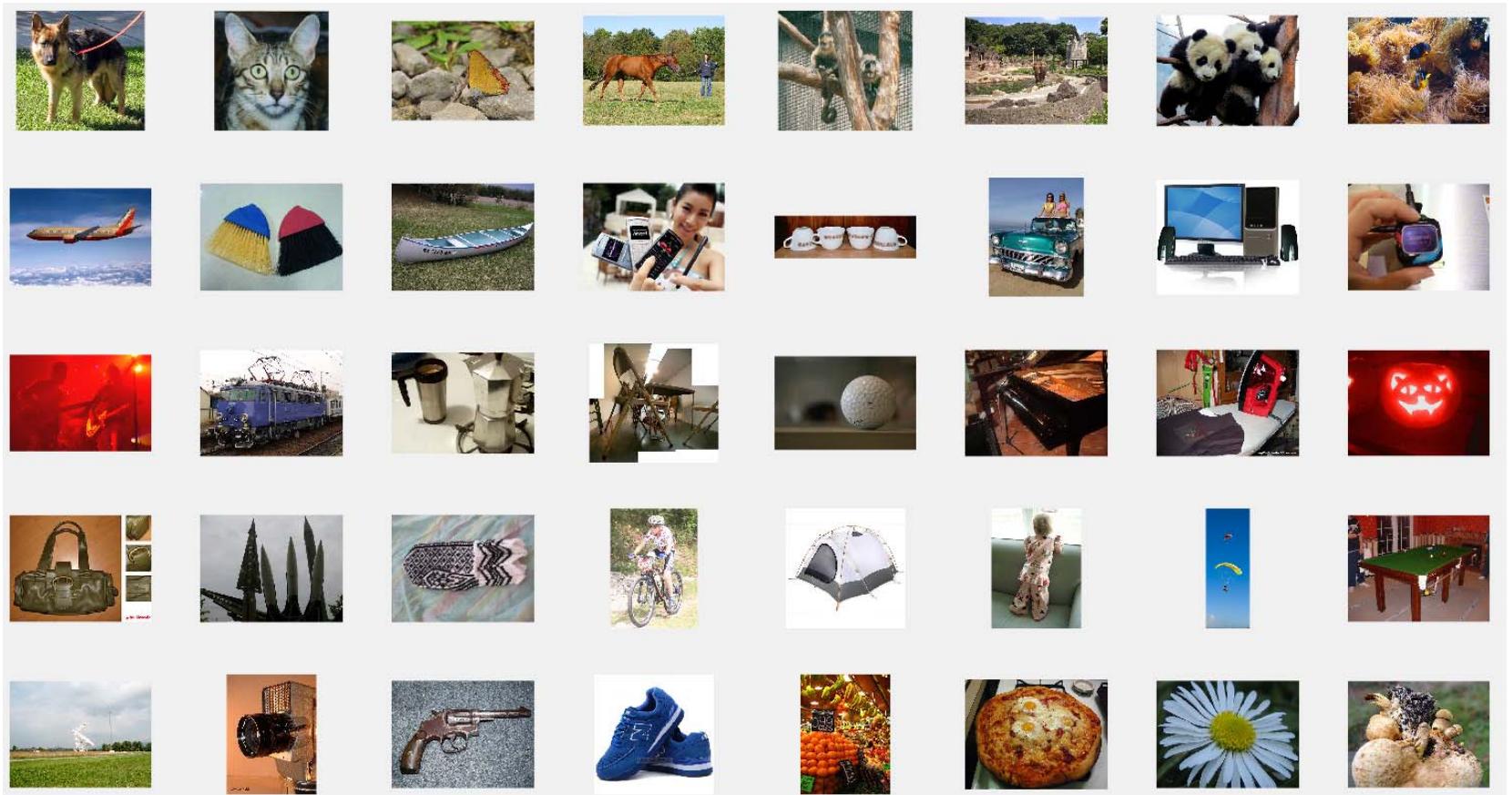
[2] T. A. Carlson, H. Hogendoorn, R. Kanai, J. Mesik, and J. Turrey. High temporal resolution decoding of object position and category. *Journal of Vision*, 11(10), 2011.

Dataset

- 6 subjects, asked to look at pictures while recording EEG
- 40 ImageNet classes
- Image duration 500 ms
- 128 channel EEG, 1000 Hz sampling rate, 16 bit resolution
- Dataset available:
 - http://perceive.dieei.unict.it/files/eeg_data_cvpr_2017.zip

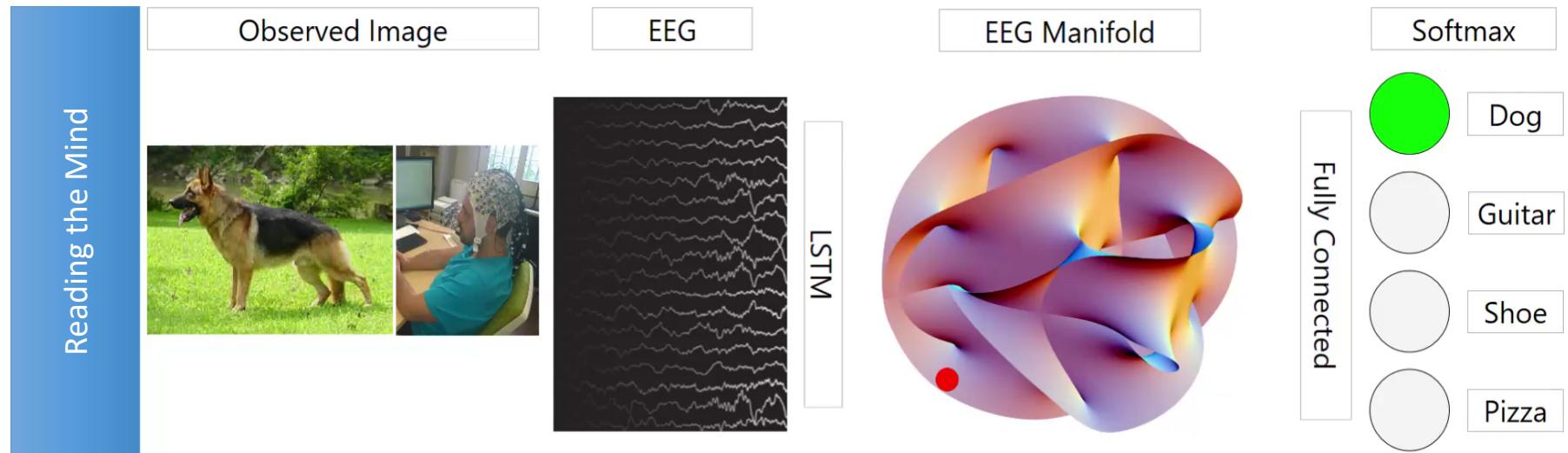


Dataset: 40 ImageNet Categories



- Can class-discriminative EEG features be extracted from raw EEG signals?

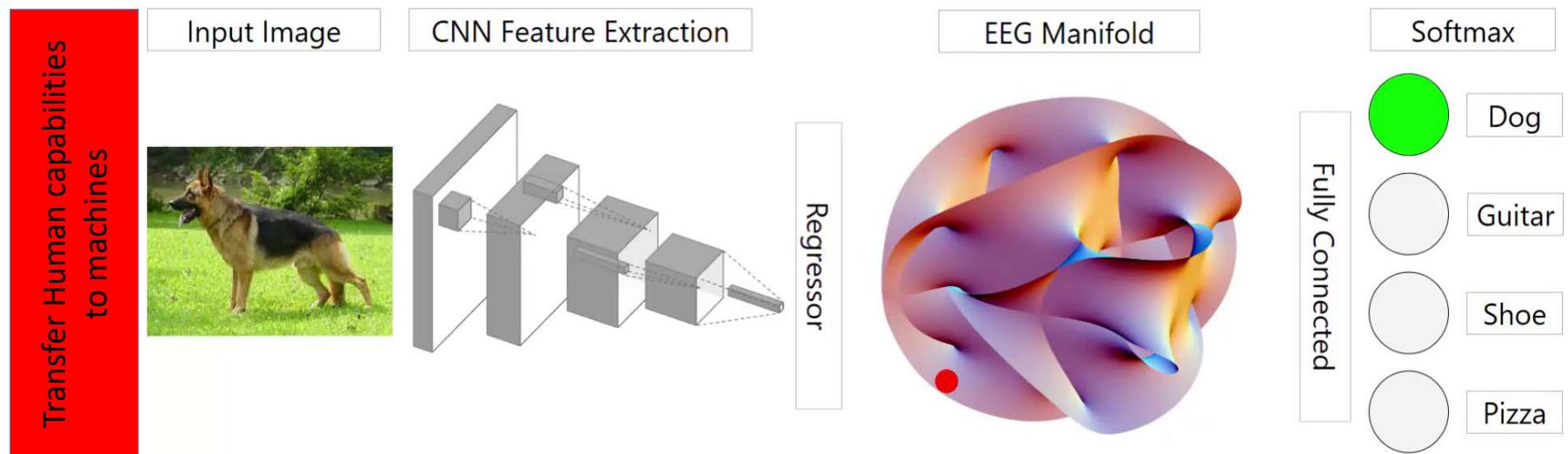
Approach



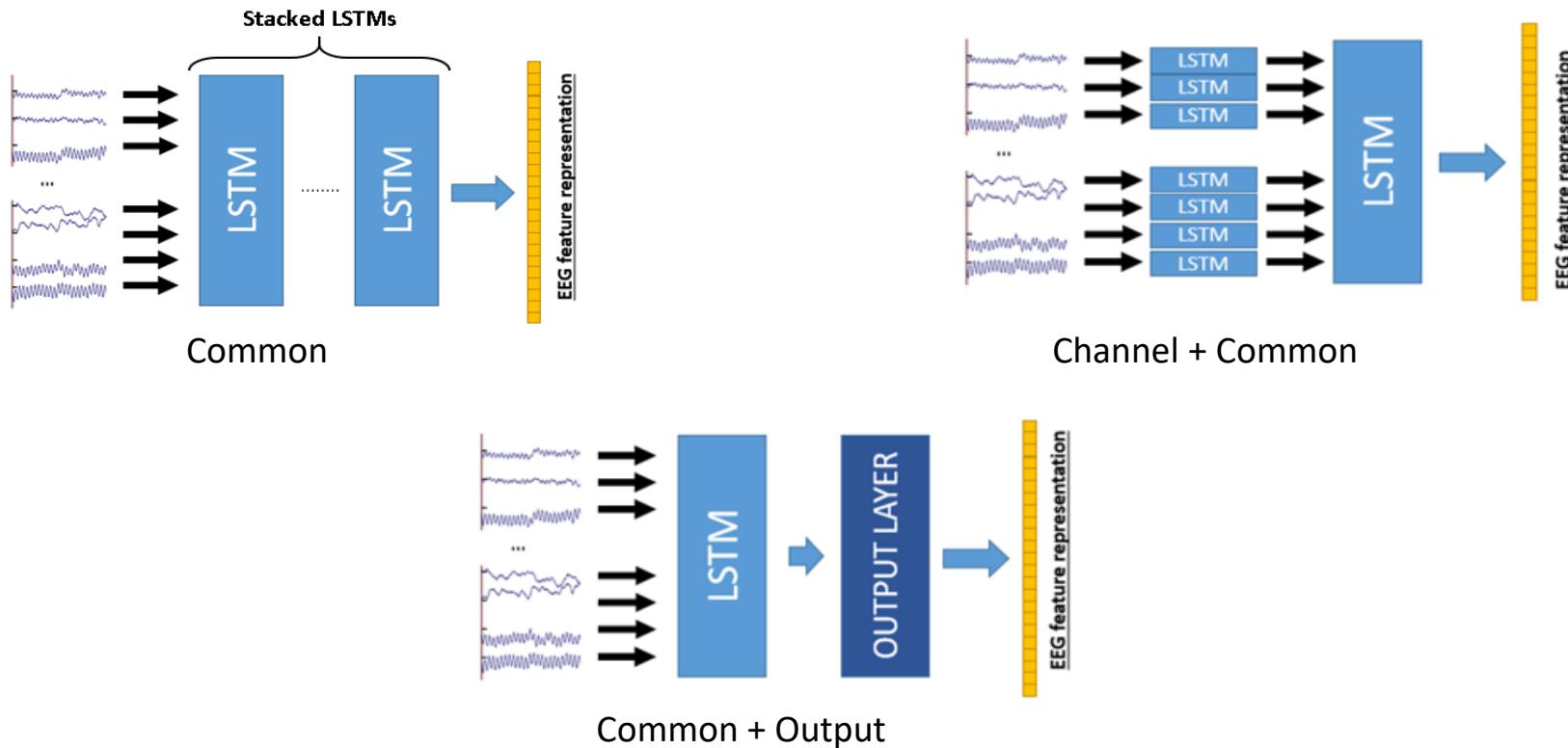
Human-Machine Computer Vision Systems

Bottom-up approach: brain-driven visual classifier

EEG features can be used for automated image classification



EEG Feature Encoding



EEG Classification Accuracy



Model	Details	Max VA	TA at max VA
Common	64 common	74.4%	73.9%
	128 common	77.3%	74.1%
	64,64 common	75.9%	72.5%
	128,64 common	79.1%	76.8%
	128,128 common	79.7%	78.0%

Deep Learning Human Mind for Automated Visual Classification



Results: Investigation of Cognitive Factors

- ❑ 50 – 120 ms: Feature extraction for object recognition [1]
- ❑ >120 ms: ???

Visualization time	Max VA	TA at max VA
40-480 ms	85.4%	82.9%
40-160 ms	81.4%	77.5%
40-320 ms	82.6%	79.7%
320-480 ms	86.9%	84.0%

[1] J. R. Heckenlively and G. B. Arden. Principles and practice of clinical electrophysiology of vision. MIT press, 2006.

Previous Methods

Work	Method	Object	Results
Bashivan et al. [1]	Combination of CNN and RNN	Cognitive load classification	90% accuracy over <u>4 classes</u>
Stober et al. [2]	CNNs	Classification of EEG signals evoked by songs	28% accuracy over <u>12 classes</u>
Kaneshiro et al. [3]	SVM	Classification of brain signals evoked by visual stimuli	29% accuracy over <u>12 classes</u>

[1] P. Bashivan, I. Rish, M. Yeasin, and N. Codella. Learning representations from EEG with deep recurrent-convolutional neural networks. ICLR 2016,

[2] S. Stober, A. Sternin, A. M. Owen, and J. A. Grahn. Deep feature learning for EEG recordings. ICLR 2016.

[3] B. Kaneshiro, et al. A Representational Similarity Analysis of the Dynamics of Object Processing Using Single-Trial EEG Classification. PloS One, 2015

Can EEG features be used for automated image classification?

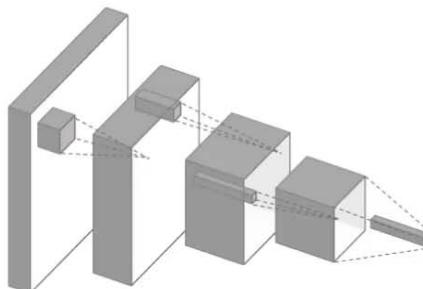
Network

Transfer Human capabilities
to machines

Input Image

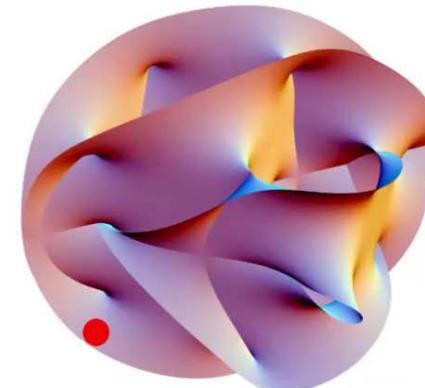


CNN Feature Extraction



Regressor

EEG Manifold



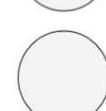
Softmax



Dog



Guitar



Shoe

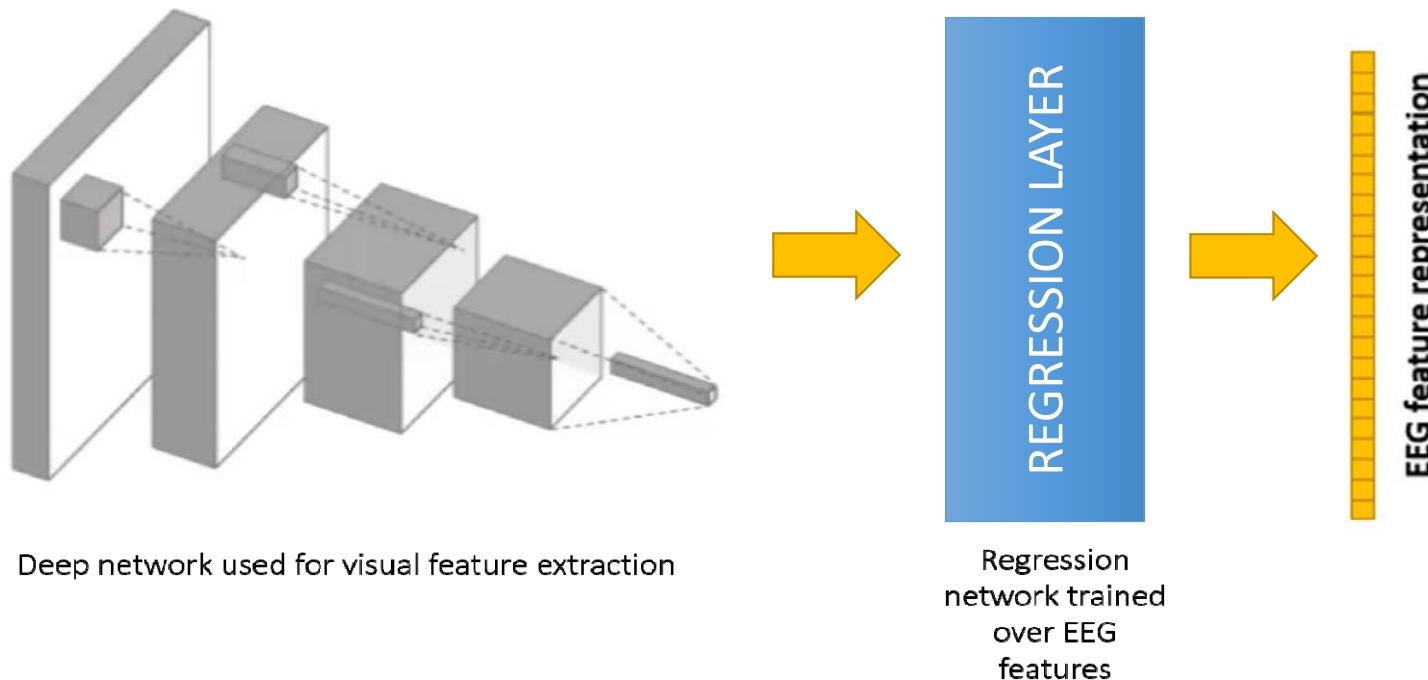


Pizza

Fully Connected

Human-Machine Computer Vision Systems

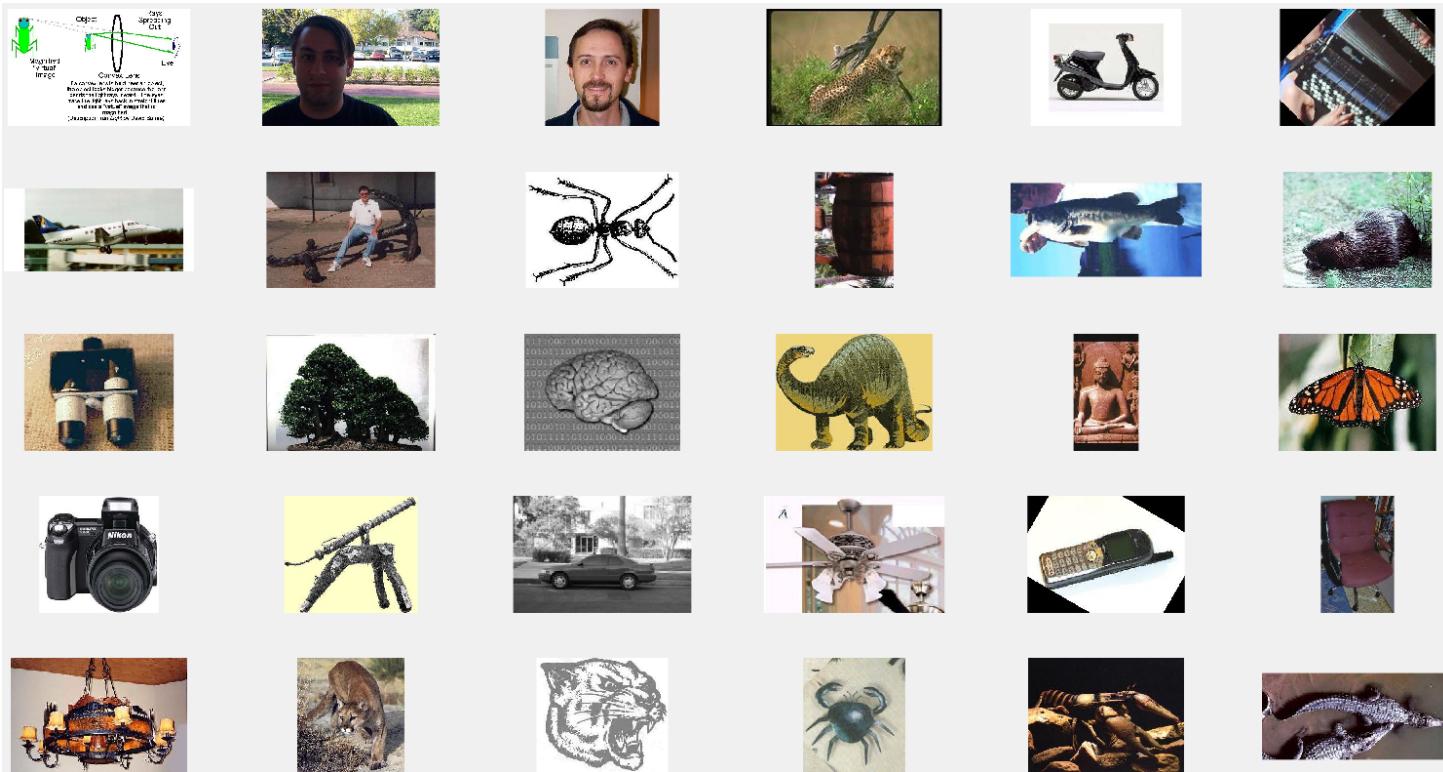
EEG feature regression architecture



- Regression accuracy:
 - Mean square error of regressed EEG features
 - Training: each image's target was the average EEG vector over the 6 subject

Feature set	AlexNet			GoogleNet			VGG		
	k-NN	Ridge	RF	k-NN	Ridge	RF	k-NN	Ridge	RF
Average	1.64	1.53	1.52	<u>0.62</u>	1.88	0.93	0.73	1.53	0.94

CALTECH 101 (30 classes)



RESULTS

GoogleNet	VGG	Our method
<i>92.6%</i>	<i>80.0%</i>	<i>89.7%</i>

Video Fill In the Blank using LR/RL LSTMs with Spatial-Temporal Attentions

Amir Mazaheri, Dong Zhang, and Mubarak Shah

ICCV 2017

<http://crcv.ucf.edu/papers/iccv17/PID4929115.pdf>



UNIVERSITÀ
degli STUDI
di CATANIA



Generative Adversarial Networks Conditioned by Brain Signals

S. Palazzo , C. Spampinato, I. Kavasidis, D. Giordano

PeRCeVe Lab, University of Catania, Italy

www.perceivelab.com

M. Shah

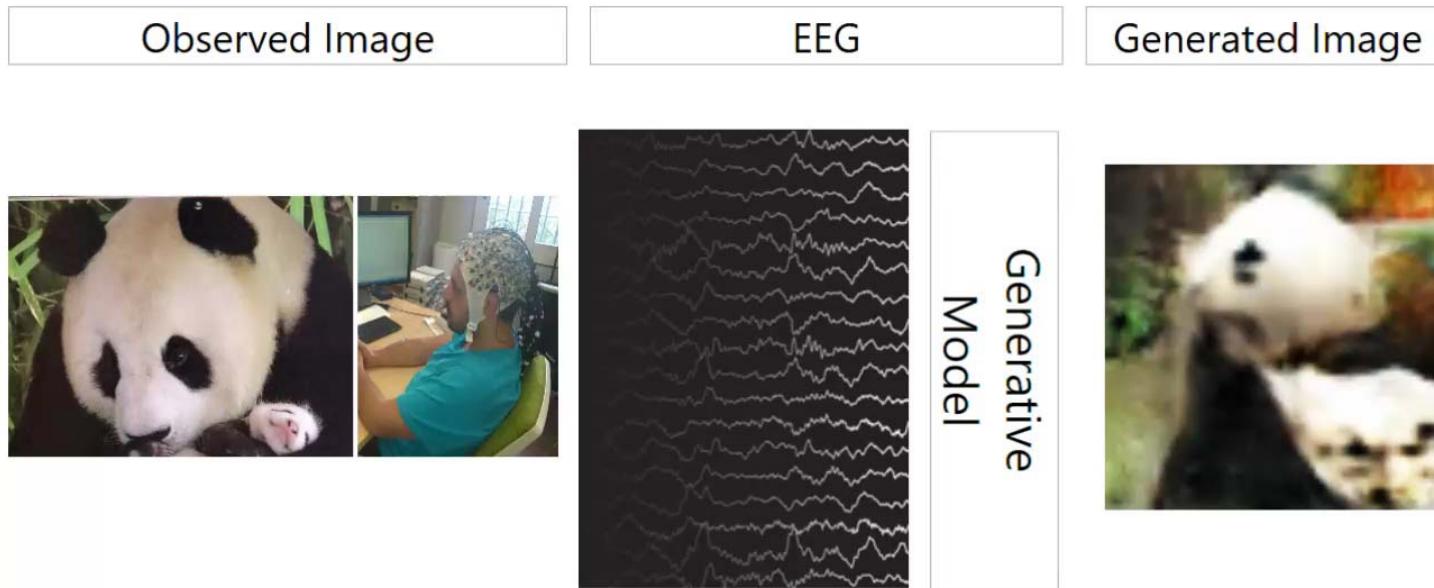
CRCV, University of Central Florida, USA

<http://crcv.ucf.edu>

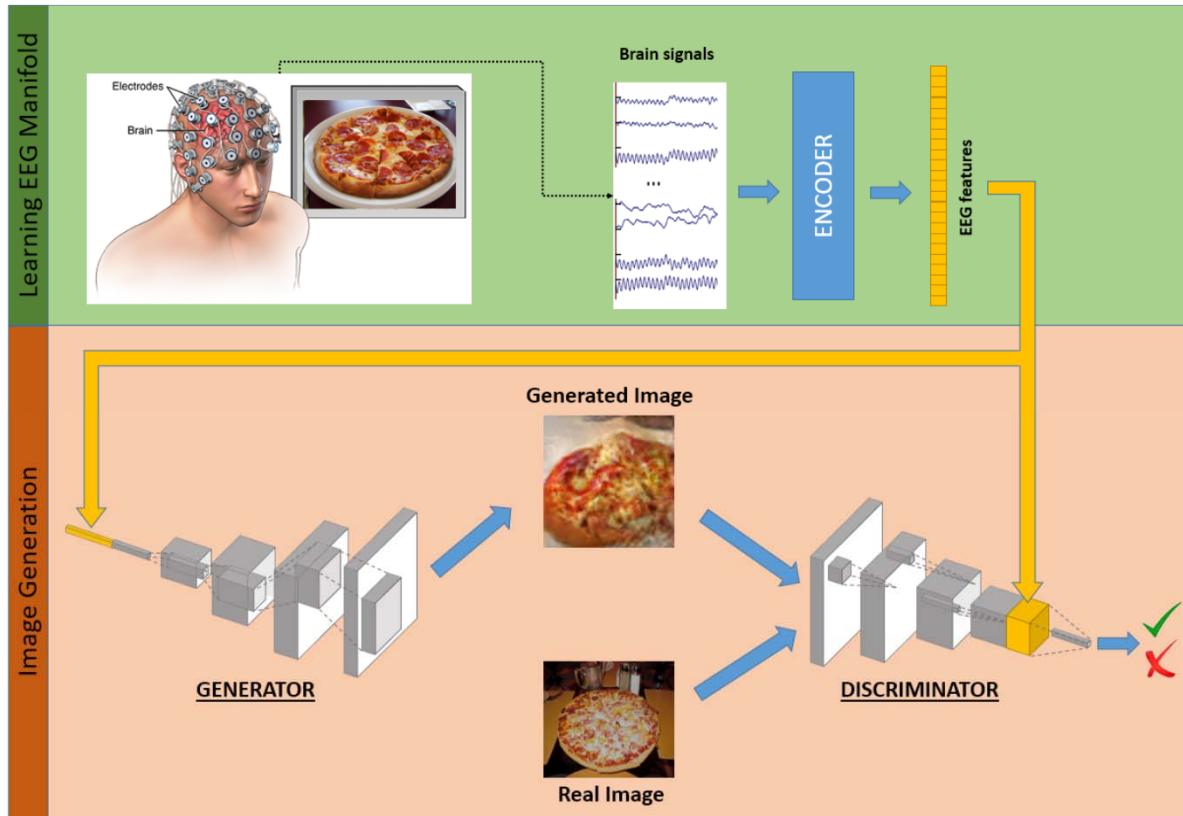
ICCV-2017

http://crcv.ucf.edu/papers/iccv17/egpaper_for_review.pdf

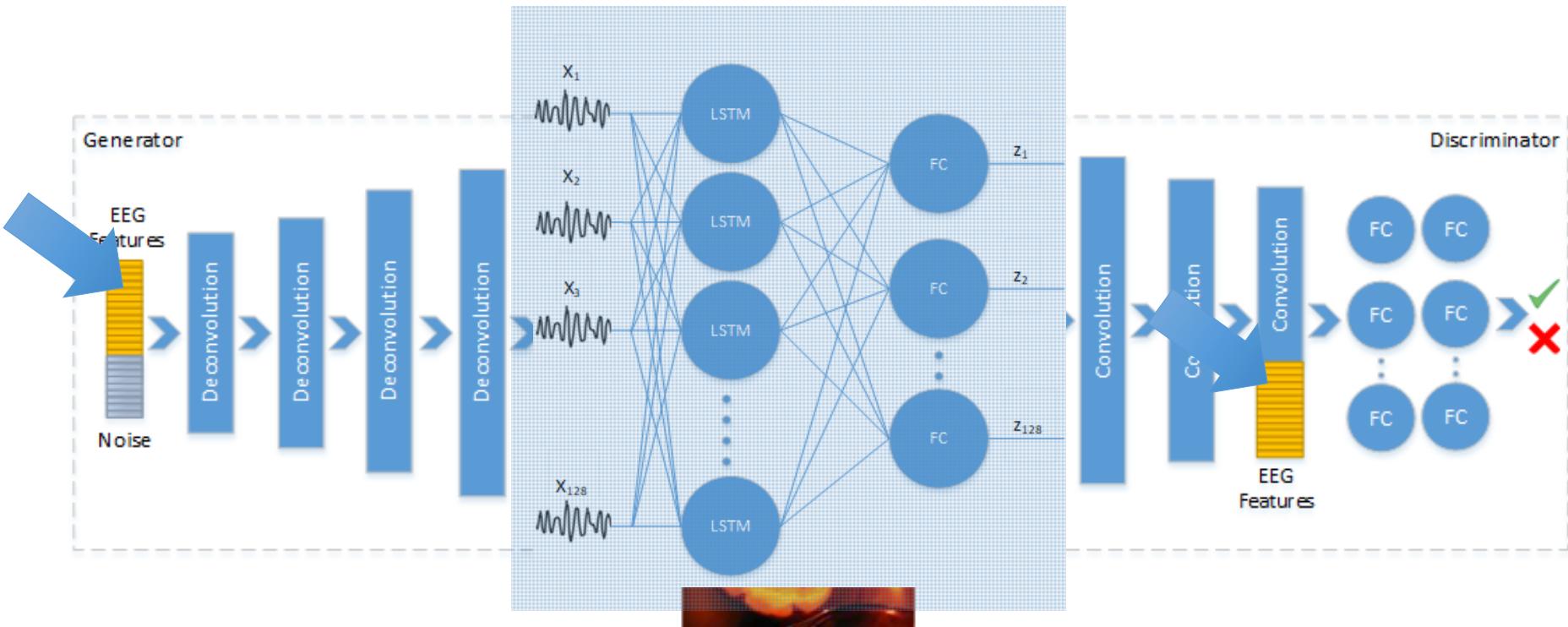
Generative models to generate images from brain signals



GAN (Generative Adversarial Network)



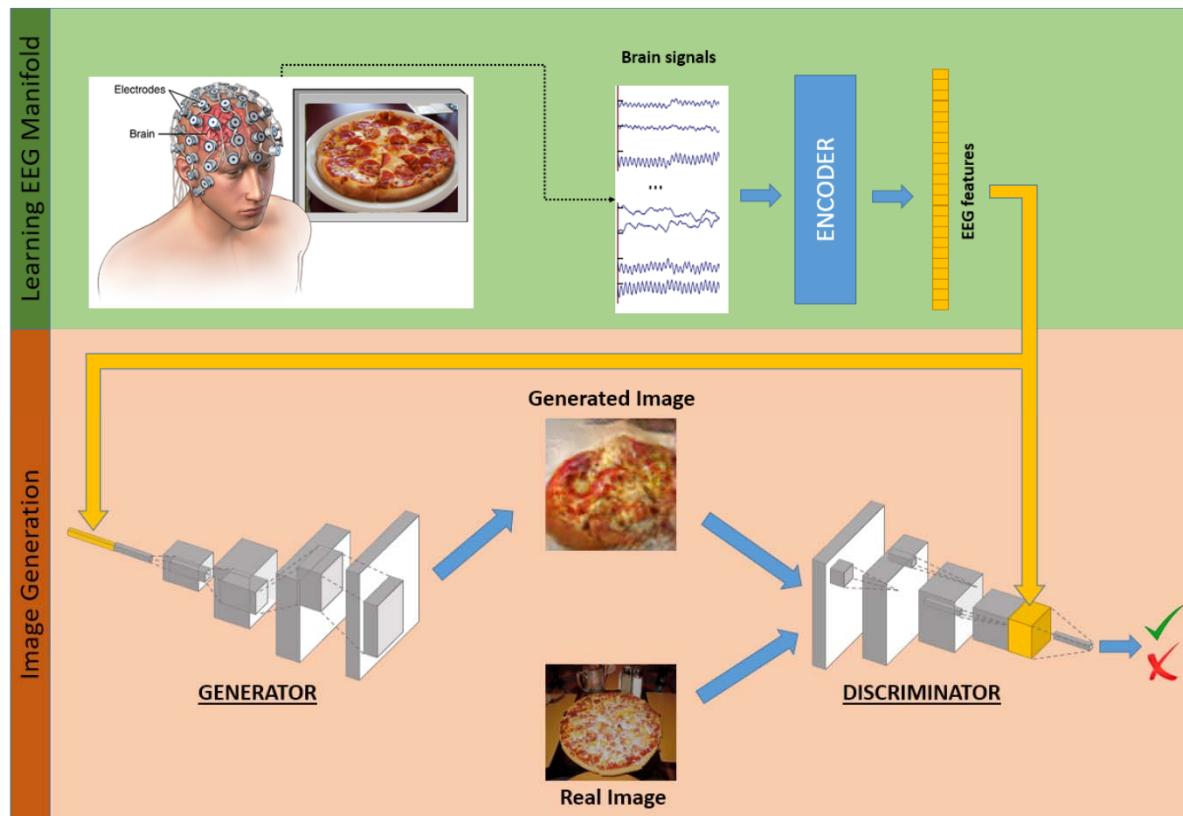
GAN (Generative Adversarial Network)



$$\mathcal{L}_D = -\log D(x_t|y_t) - \log(1 - D(x_c|y_w)) - \log(1 - D(x_w|y_w))$$

Human-Machine Computer Vision Systems

Bottom-up approach: brain-driven visual classifier



$$\mathcal{L}_G = - \log D(x_w|y_w)$$

$$\begin{aligned}\mathcal{L}_D = & - \log D(x_c|y_c) \\ & - \log (1 - D(x_c|y_w)) \\ & - \log (1 - D(x_w|y_w)).\end{aligned}$$

S. Palazzo, C. Spampinato, I. Kavasidis, M. Shah, "Generative Adversarial Networks Conditioned by Brain Signals" ICCV 2017

Generated Images



(a) Airliner



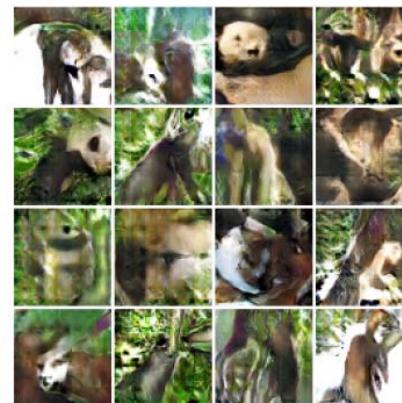
(b) Jack-o'-Lantern
 Figure 3. Good results



(c) Panda



(a) Banana



(b) Capuchin
 Figure 4. Bad results



(c) Bolete



UNIVERSITÀ
degli STUDI
di CATANIA



Generative Adversarial Networks Conditioned by Brain Signals

S. Palazzo , C. Spampinato, I. Kavasidis, D. Giordano

PeRCeVe Lab, University of Catania, Italy

www.perceivelab.com

M. Shah

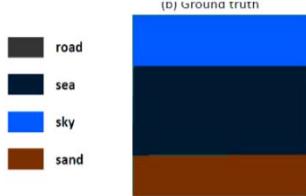
CRCV, University of Central Florida, USA

<http://crcv.ucf.edu>

ICCV-2017

http://crcv.ucf.edu/papers/iccv17/egpaper_for_review.pdf

Contents



Sematic Segmentation



Facial Attributes Detection



Human Re-Identification



Target Detection in WAMI



Diving



Anomaly Detection

Human Action Localization



Single Blank:

He ___ up the steps of the stand and away. (Runs)

Video Fill In The Blank



Reading The Mind

Contents

- PART-I: Deep Learning: A Short Overview
- **PART II: Computer Vision Employing Deep Learning**

Conclusions

- Deep Learning has been disruptive to Computer Vision
- Deep Computer Vision is being used
 - Self Driving Cars
 - Robotics
 - Health Care
 - Language and Vision
 - Sound and Vision
- Artificial General Intelligence
 - Alpha Go (zero)