Structured deep learning for visual localization and recognition

Wanli Ouyang （欧阳万里）
wanli.ouyang@sydney.edu.au

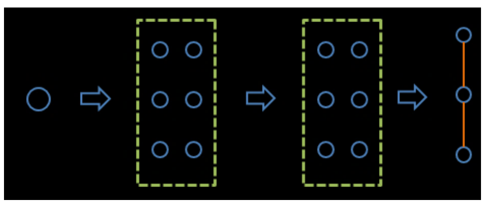
The Chinese University of Hong Kong
The University of Sydney

Outline

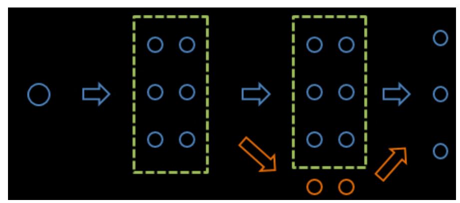
Introduction

Structured deep learning

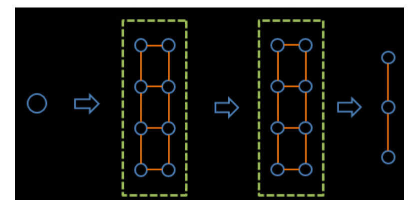
Structured output



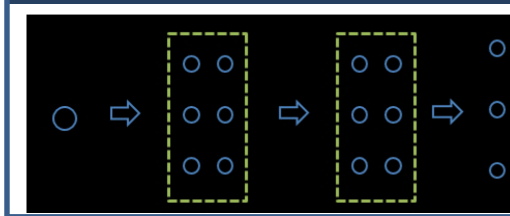
Structured Hidden factors



Structured features



Back-bone model design

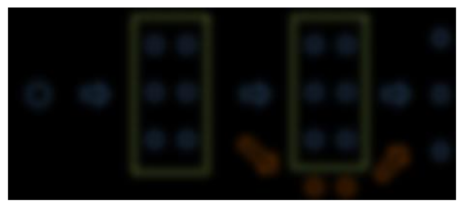


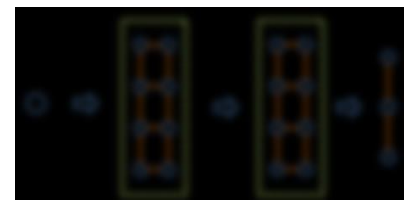
Conclusion

Outline

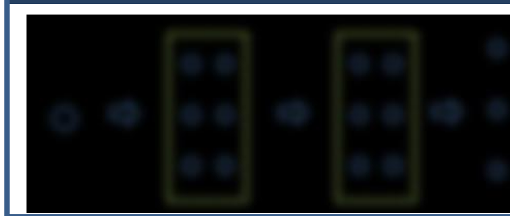
Introduction

Structured deep learning





Back-bone model design



Conclusion

Object recognition

Cat

Object detection

Child | Tooth brush | Woman | Tooth brush

Object recognition

Cat

Action recognition

Application

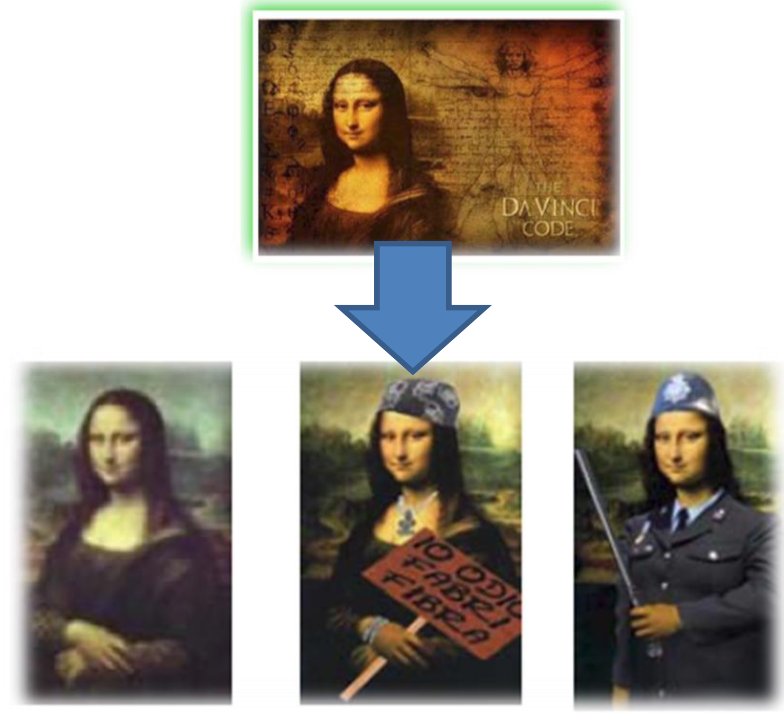
- Automotive safety and automatic car driving

Application

- Automotive safety and automatic car driving
- Robotics and Human-computer interaction

Application

- Automotive safety and automatic car driving
- Robotics and Human-computer interaction
- Internet of Things

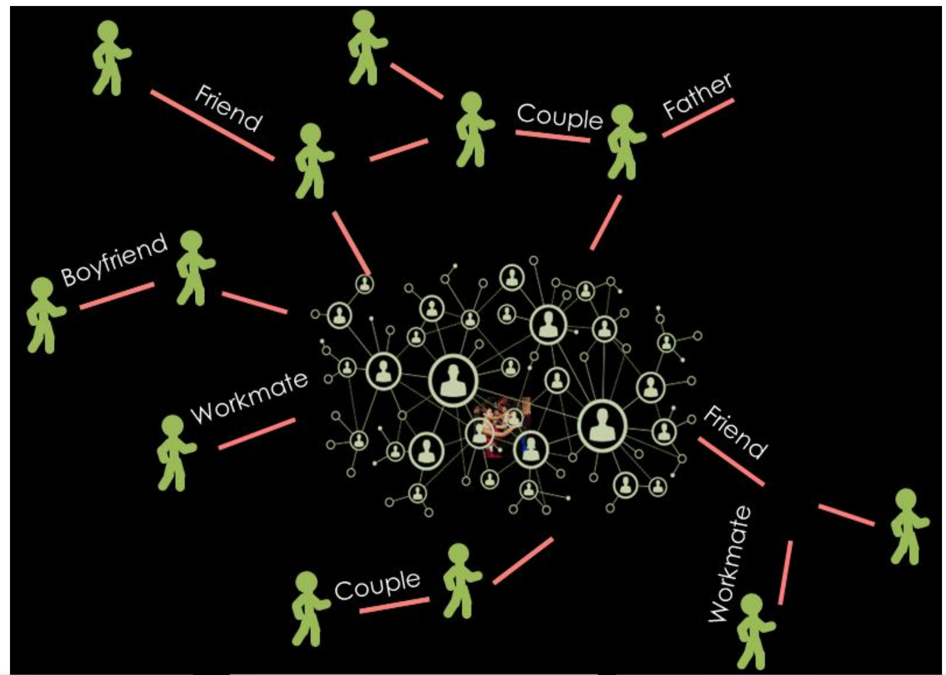


Application

- Automotive safety and automatic car driving
- Robotics and Human-computer interaction
- Internet of Things
- Public safety and smart city

Application

- Automotive safety and automatic car driving
- Robotics and Human-computer interaction
- Internet of Things
- Public safety and smart city
- Social network



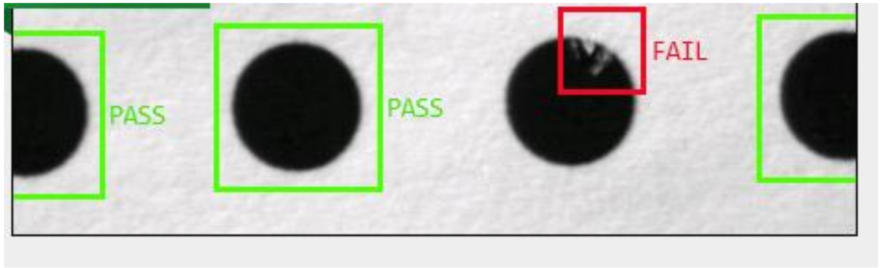

Family

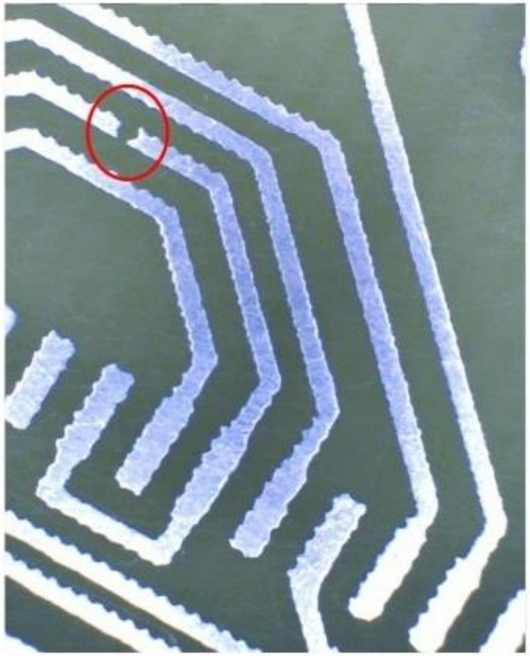

Workmate


Father

Application

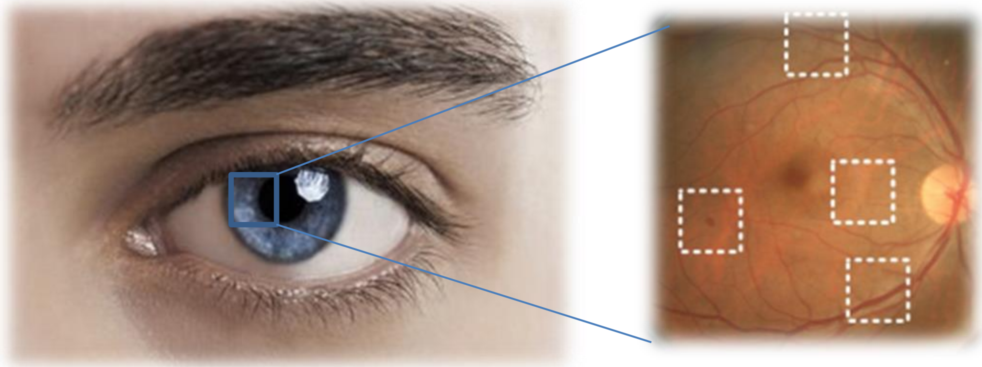
- Automotive safety and automatic car driving
- Robotics and Human-computer interaction
- Internet of Things
- Public safety and smart city
- Social network
- Industrial production



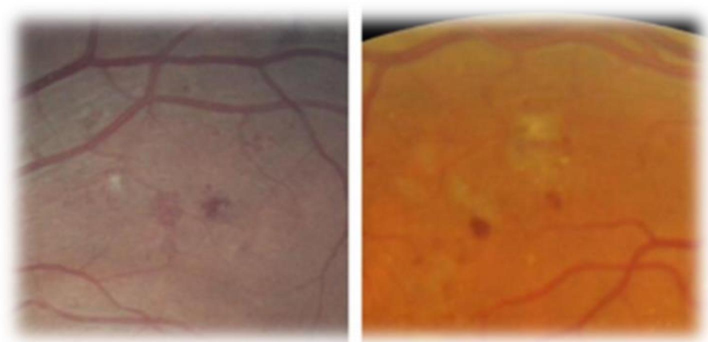


Application

- Automotive safety and automatic car driving
- Robotics and Human-computer interaction
- Internet of Things
- Public safety and smart city
- Social network
- Industrial production
- Bio-medical imaging



Microaneurysms



Blot hemorrhages

Challenges -- person

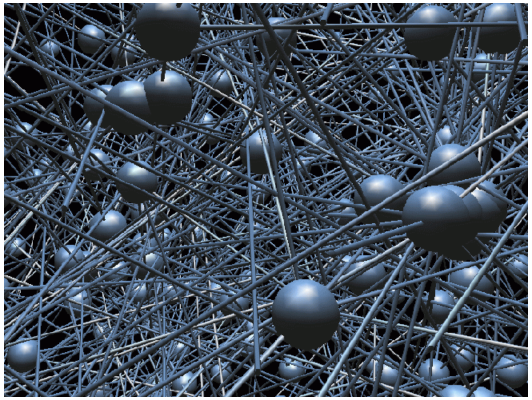
- Intra-class variation
  - Color

Challenges -- person

- Intra-class variation
  - Color
  - Occlusion

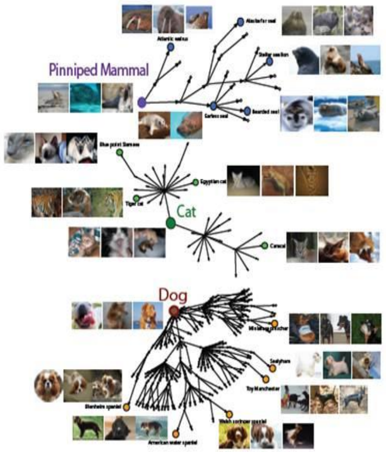
Challenges -- person

- Intra-class variation
  - Color
  - Occlusion
  - Deformation



Simulate brain activities and employ millions of neurons to fit billions of training samples. Deep neural networks are trained with GPU clusters with tens of thousands of processors

Hinton won ImageNet competition
Classify 1.2 million images into 1,000 categories
Beating existing computer vision methods by 20+%
Surpassing human performance



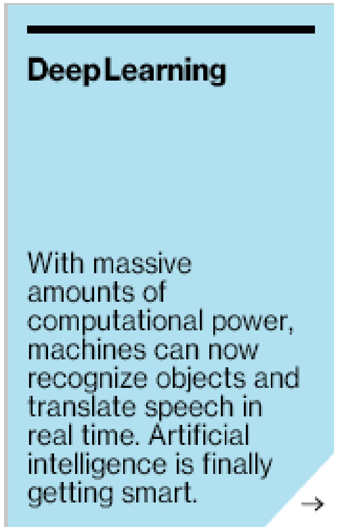
Deep learning

REVOLUTIONARY

Web-scale visual search, self-driving cars, surveillance, multimedia …

Hold records on most of the computer vision problems

MIT Tech Review
Top 10 Breakthroughs 2013
Ranking No. 1



ImageNet Large Scale Visual Recognition Challenge

Object detection

Child

Tooth brush

Woman

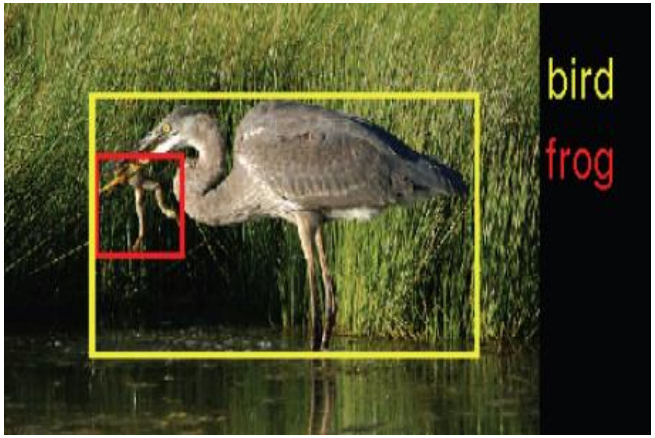
Tooth brush

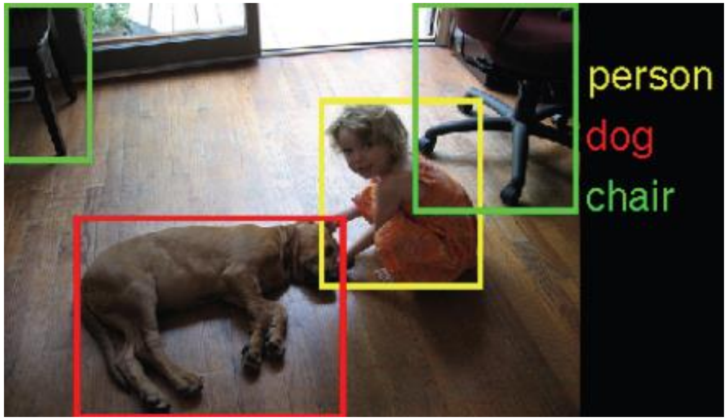
Object recognition

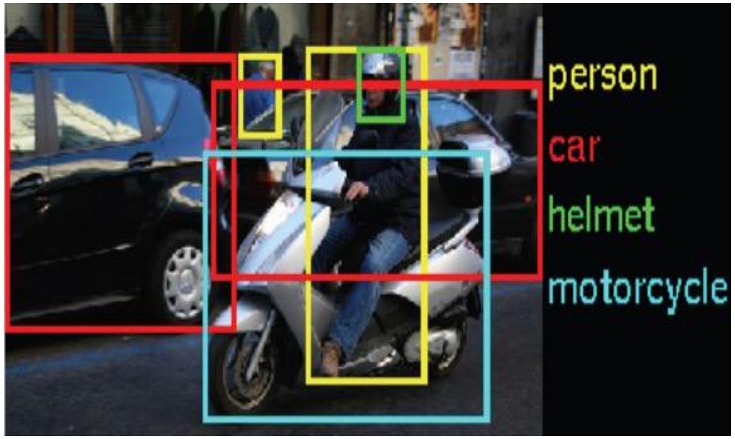
Cat

ImageNet Object Detection Task

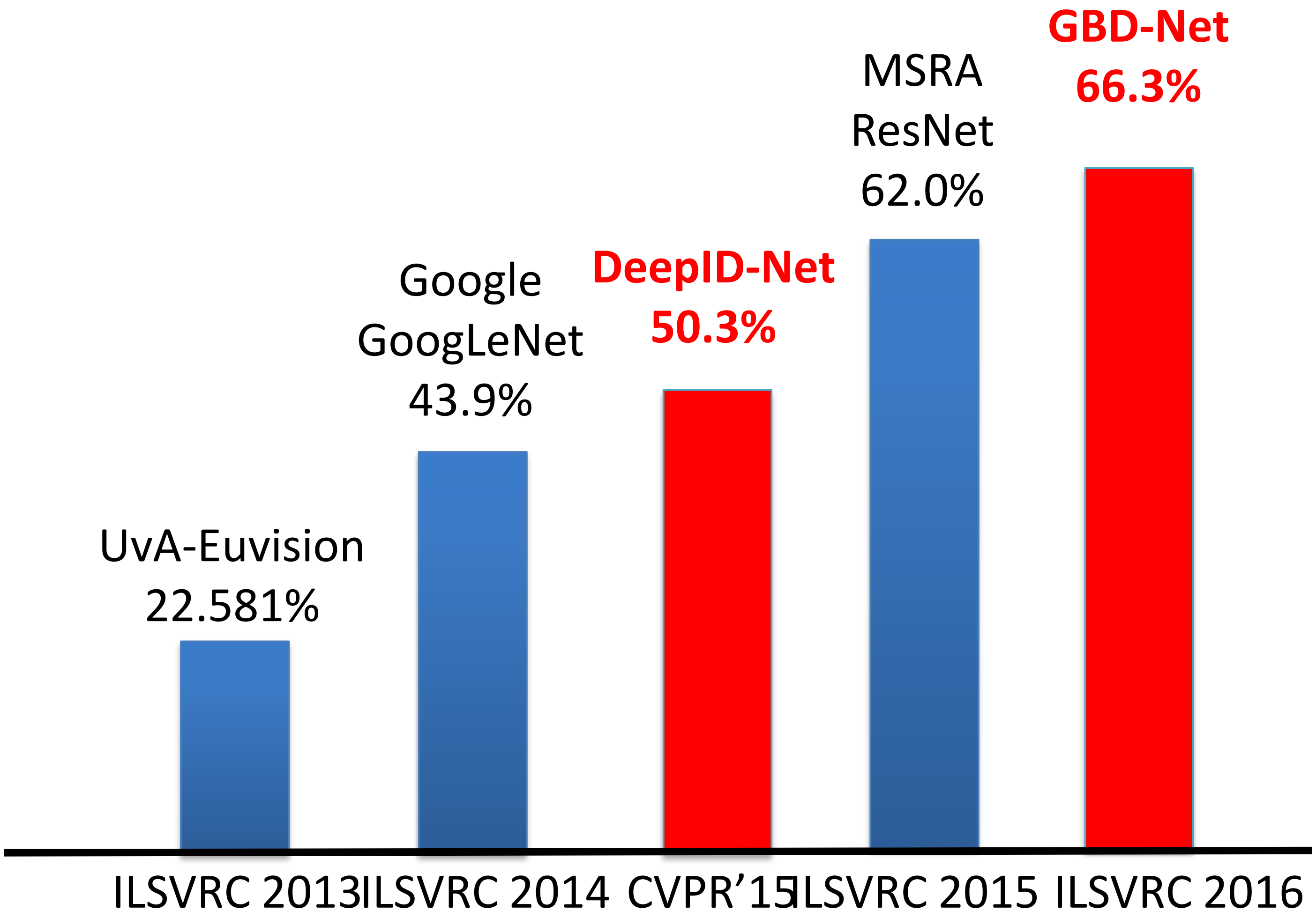
- 200 object classes
- ~500,000 training images, 60,000 test images







Mean Averaged Precision (mAP)



W. Ouyang and X. Wang, et al. "DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection," CVPR15, TPAMI17

X. Zeng, W. Ouyang, J. Yan, etc, "Crafting gbd-net for object detection," ECCV16, TPAMI 2017

Our team at ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

	2014	2015	2016
Object detection	2nd (Google 1st)		1st
Video object detection/tracking		1st	1st

Our team at Common Object in Context (COCO)

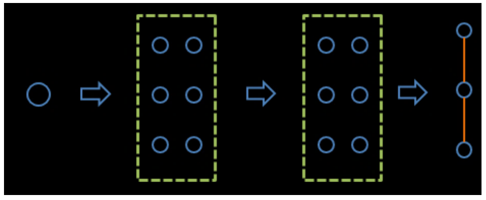
	2018
Object detection and instance segmentation	1st

Outline

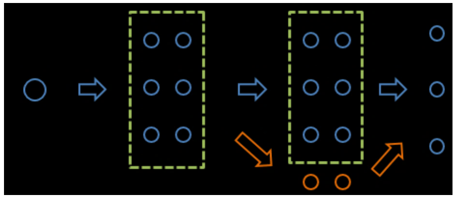
Introduction

Structured deep learning

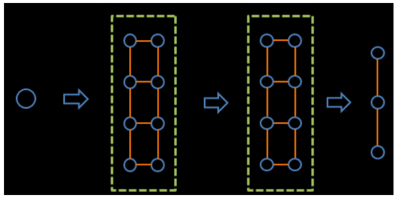
Structured output



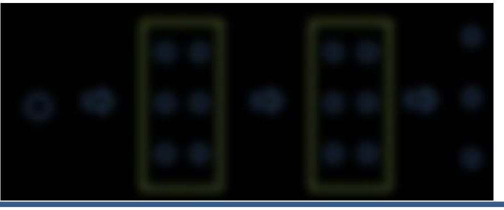
Structured Hidden factors



Structured features

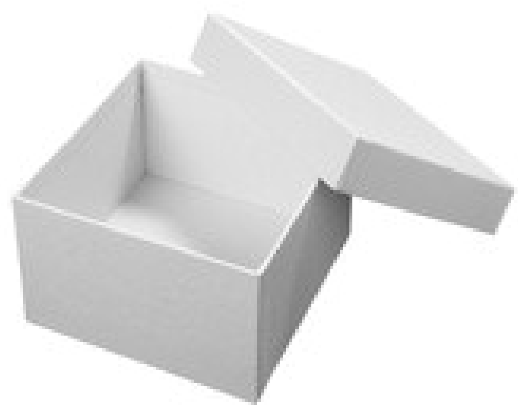


Back-bone model design

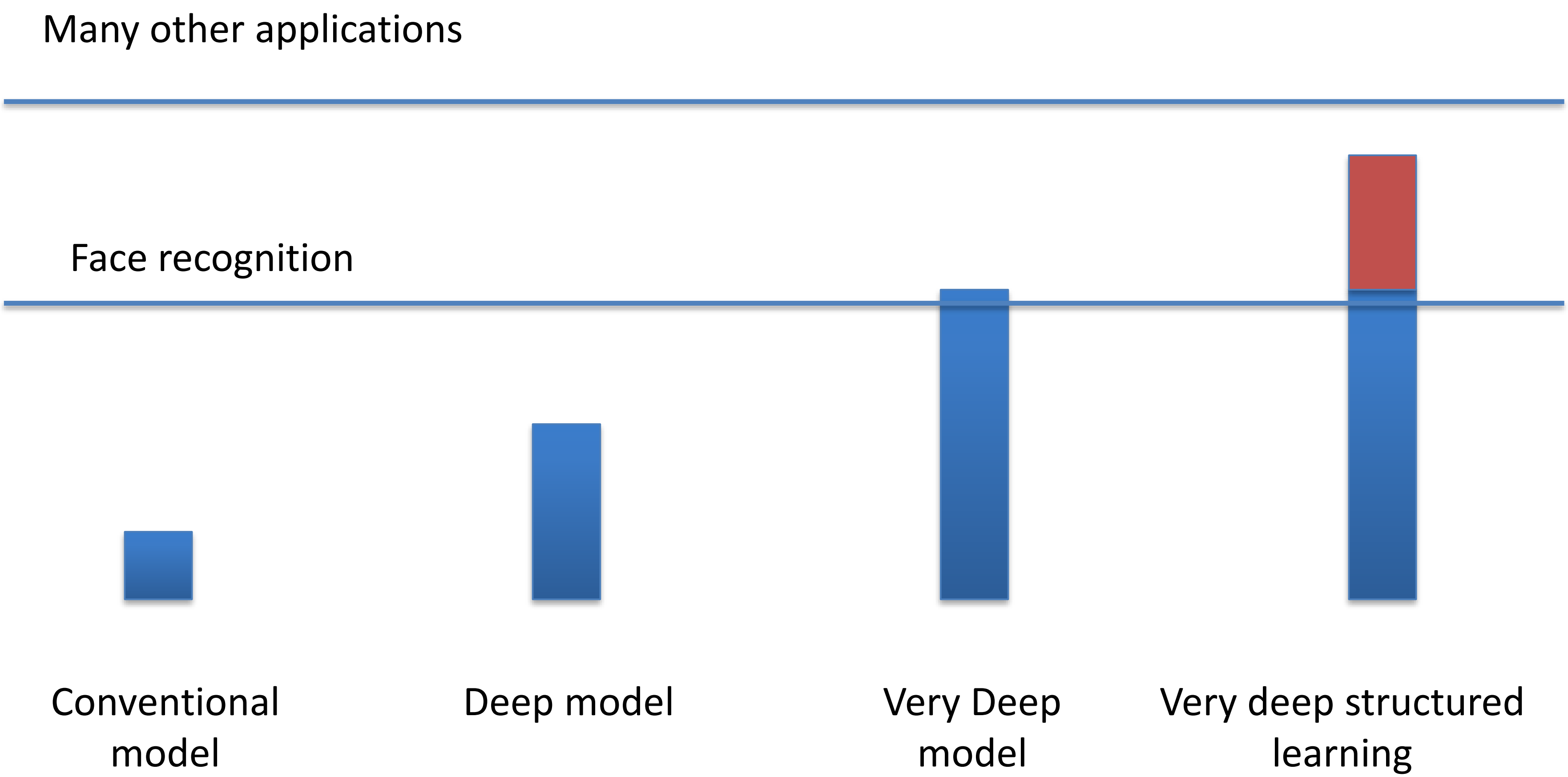


Conclusion

Is deep model a black box?

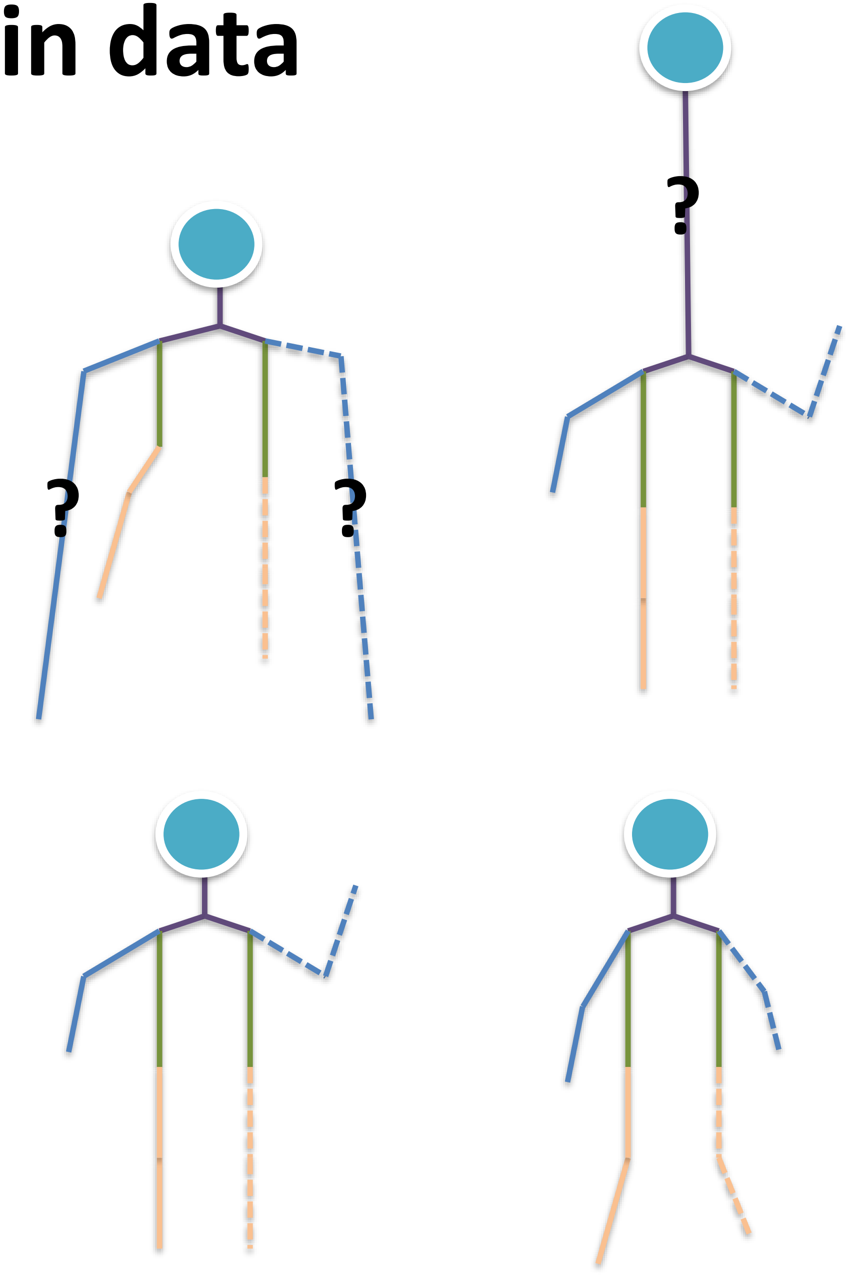


Performance vs practical need

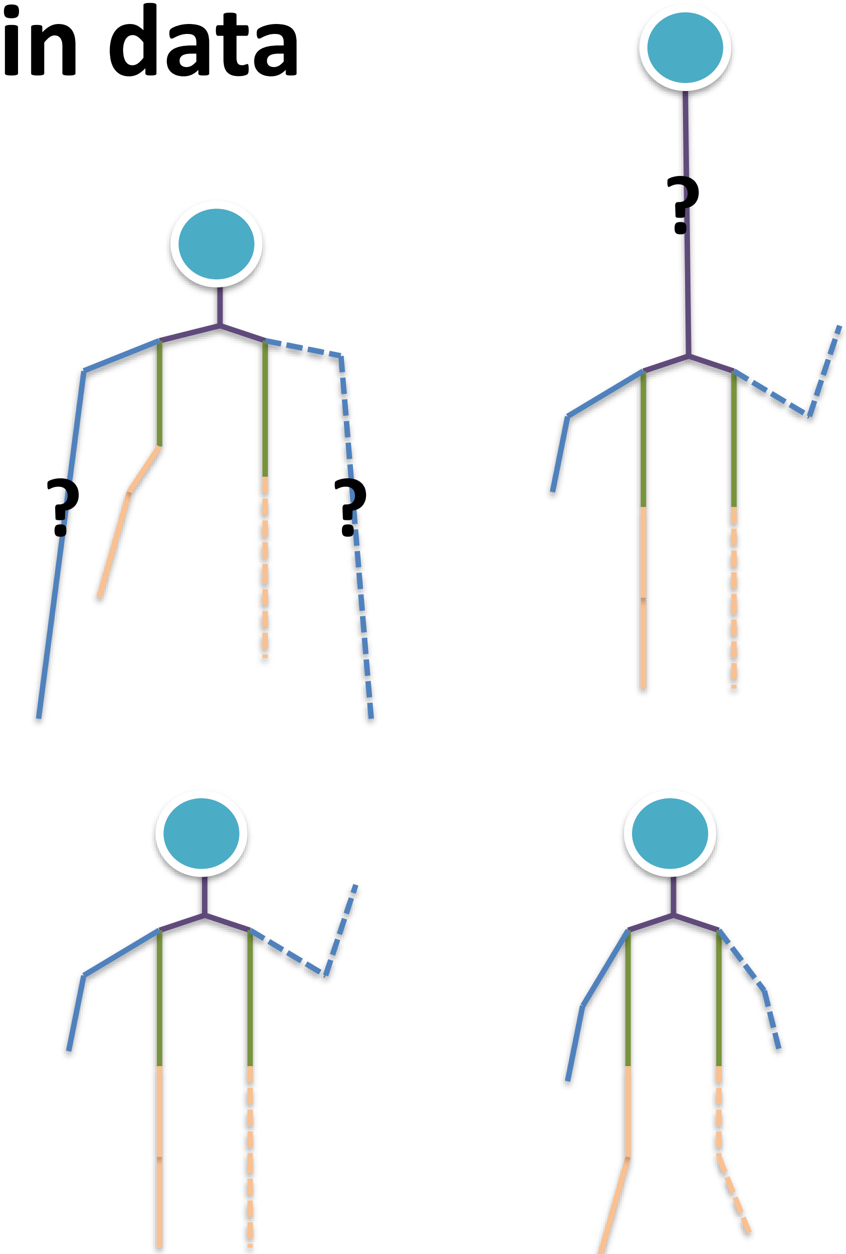


Structure in data

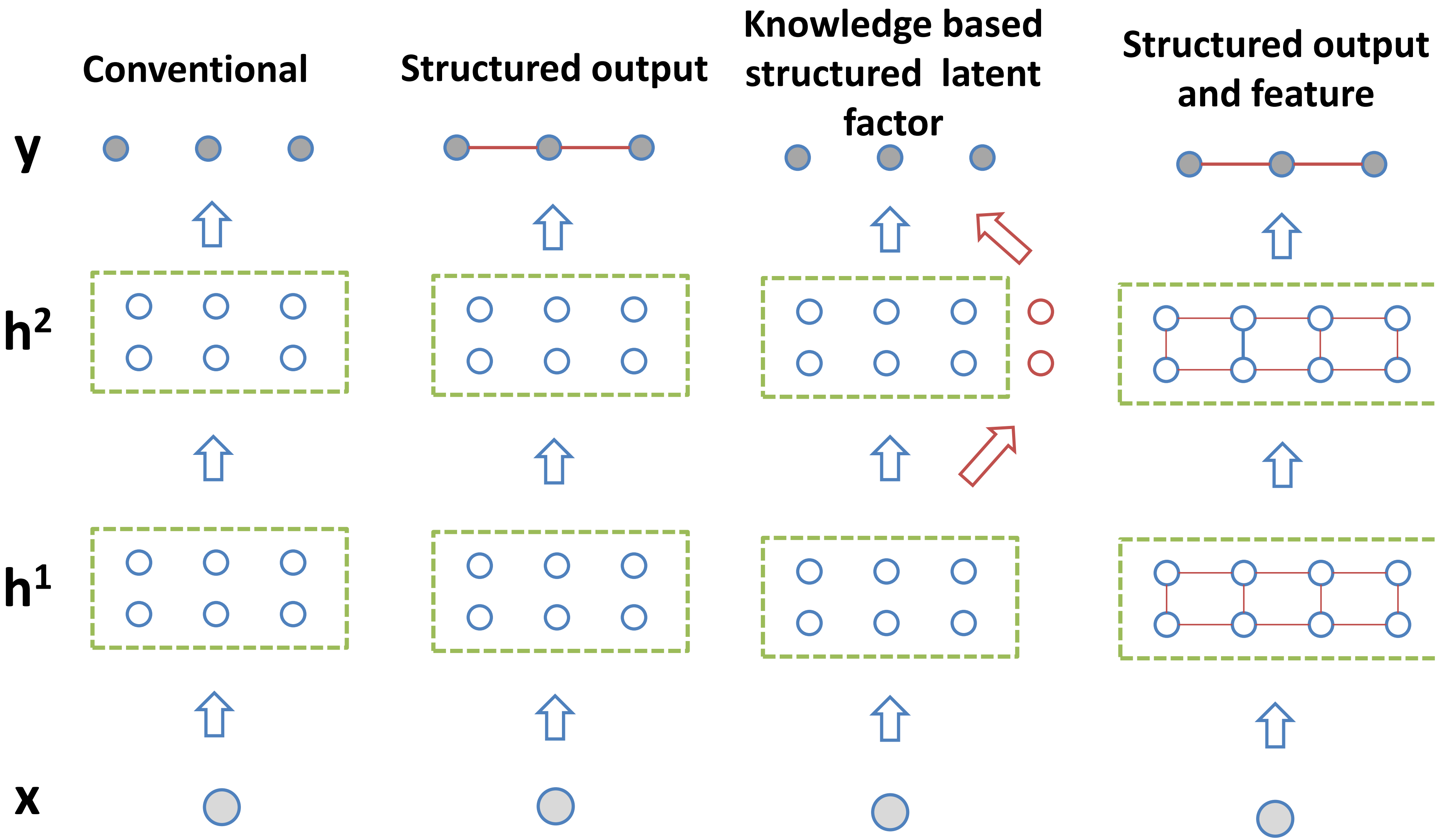
Structure in data



Structure in data



Model structures among neurons

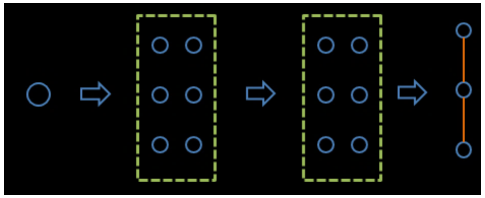


Outline

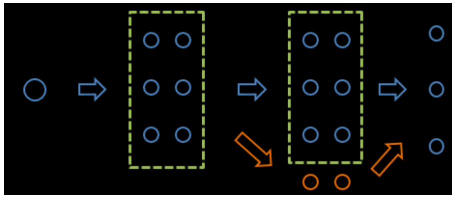
Introduction

Structured deep learning

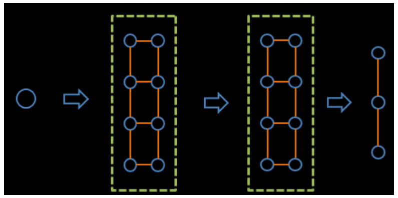
Structured output



Structured Hidden factors



Structured features



Back-bone model design

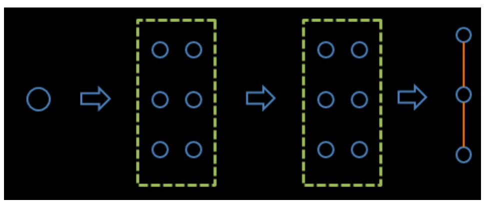
Conclusion

Outline

Introduction

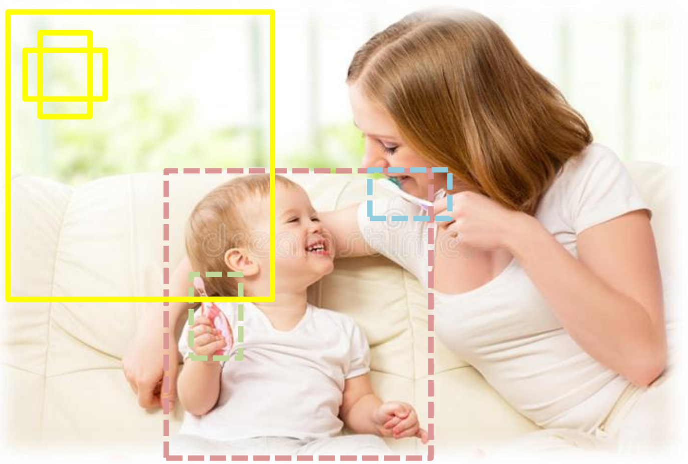
Structured deep learning

Structured output



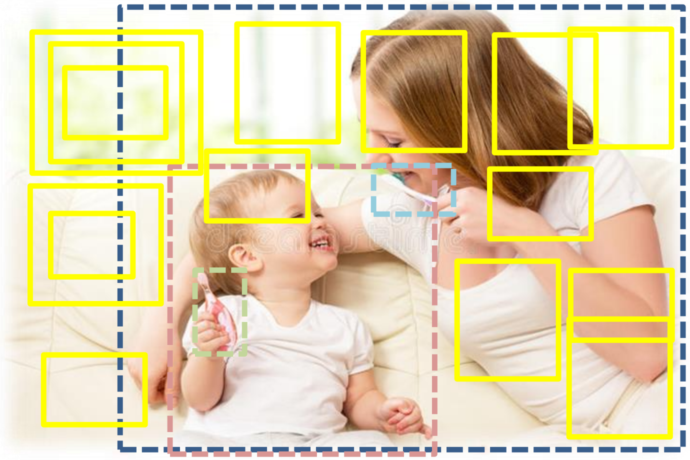
Object detection

- Sliding window
- Variable window size



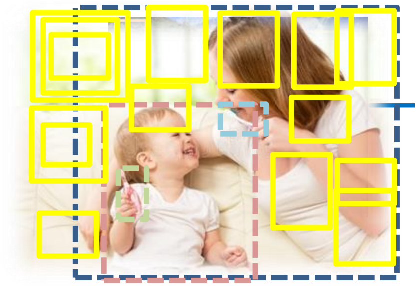
Motivation

- Much more negative samples than positive samples
- Easy to tell some regions do not contain any object

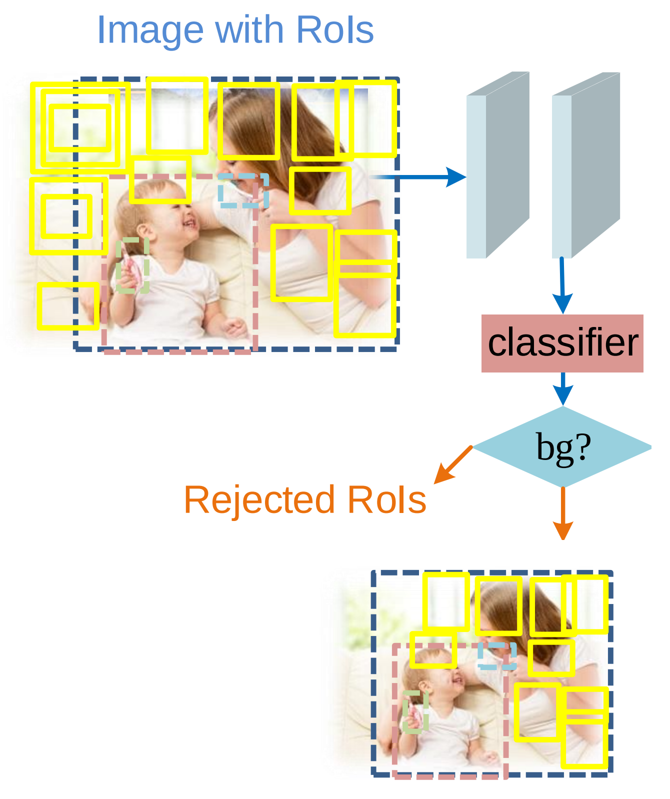


Cascade Network

Image with RoIs

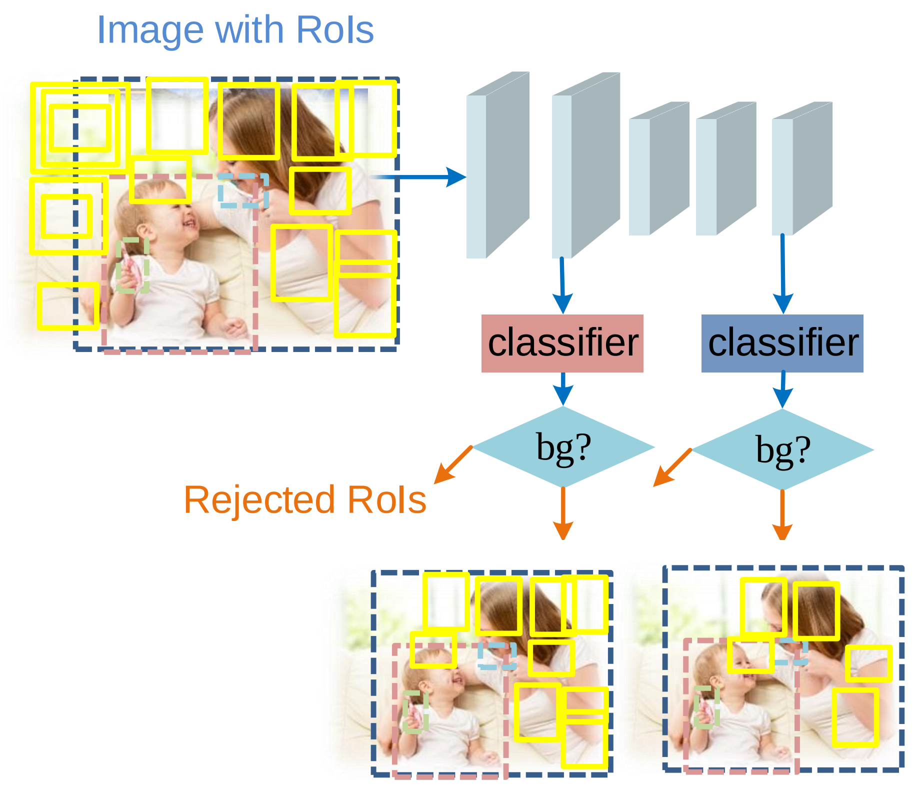


Cascade Network



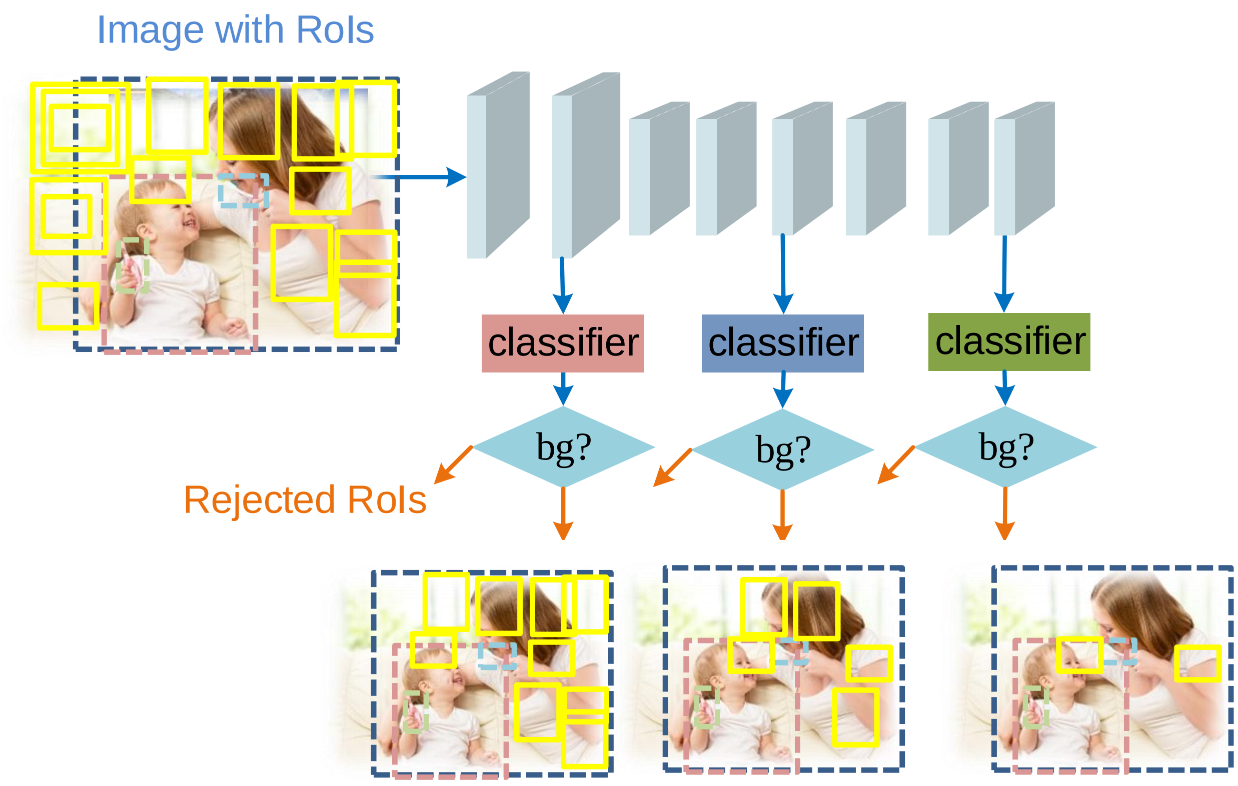
Wanli Ouyang, Kun Wang, Xin Zhu, Xiaogang Wang. "Chained Cascade Network for Object Detection", *Proc. ICCV*, 2017.

Cascade Network



Wanli Ouyang, Kun Wang, Xin Zhu, Xiaogang Wang. "Chained Cascade Network for Object Detection", *Proc. ICCV*, 2017.

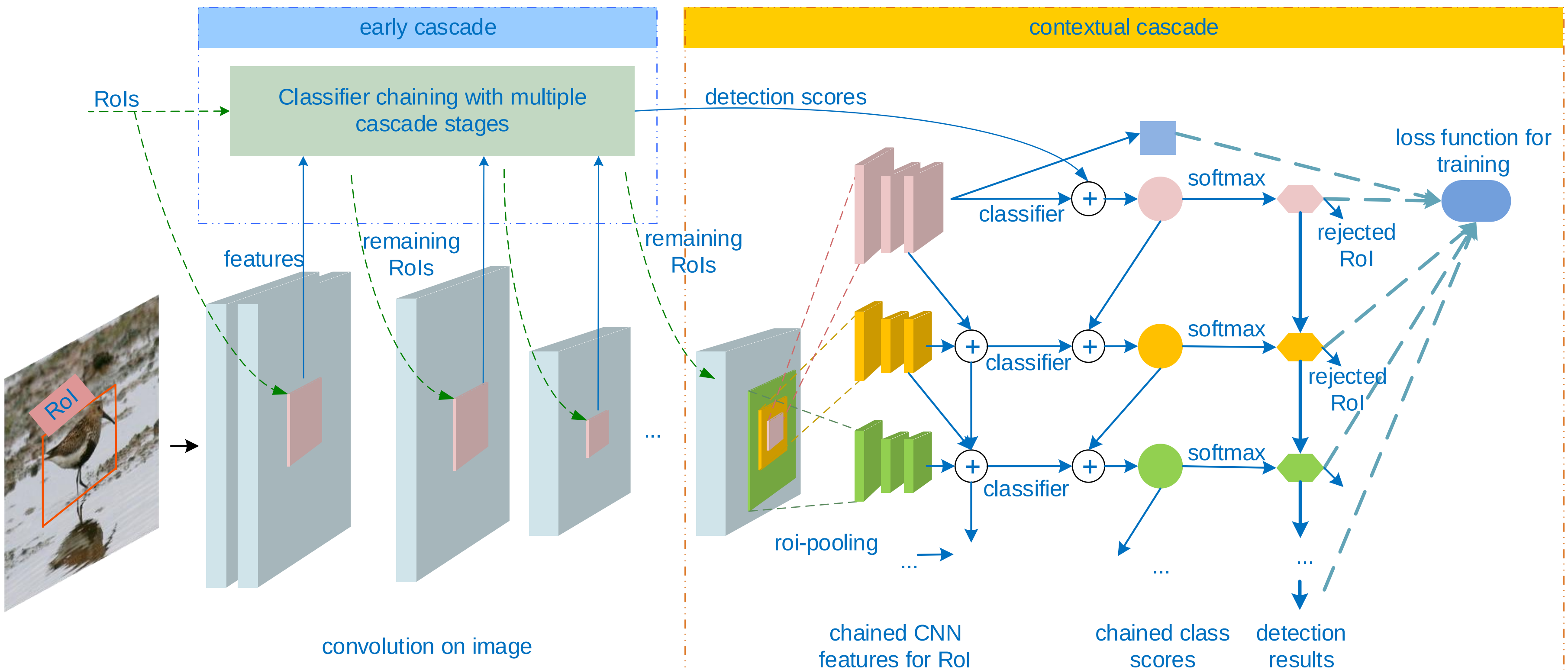
Cascade Network



Wanli Ouyang, Kun Wang, Xin Zhu, Xiaogang Wang. "Chained Cascade Network for Object Detection", *Proc. ICCV*, 2017.

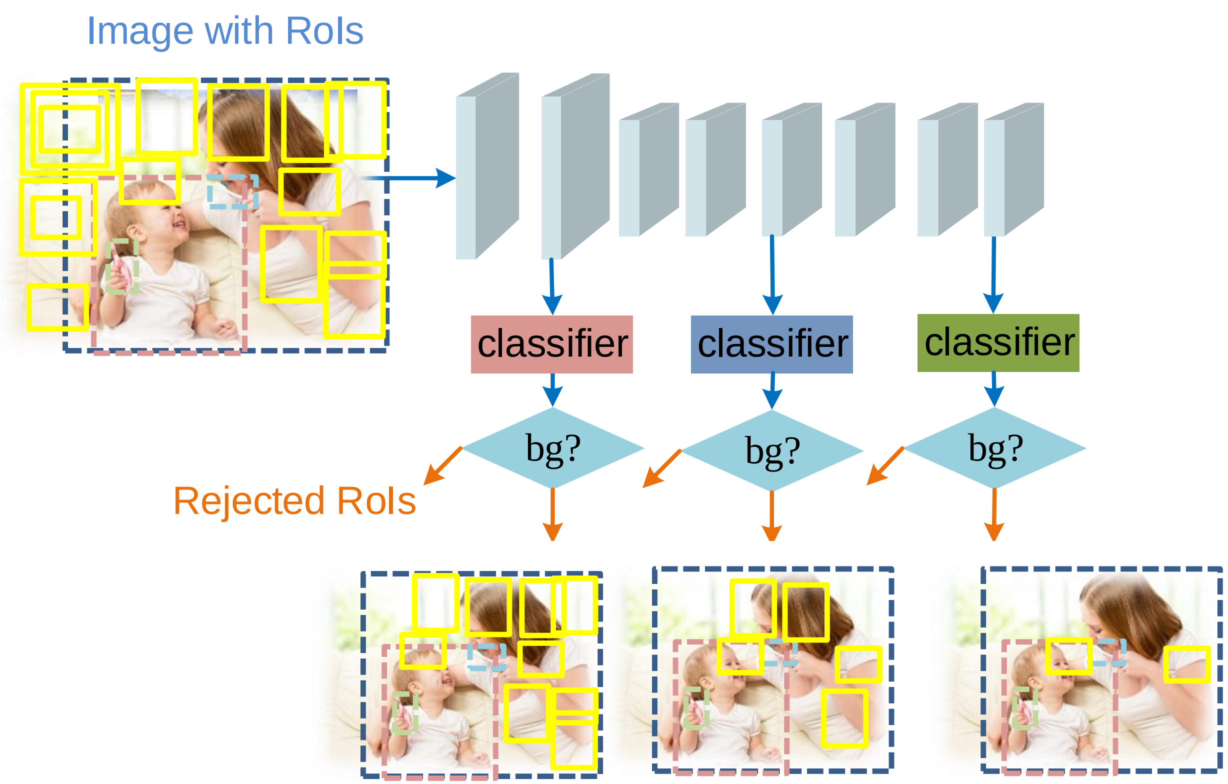
Model structures among classifiers at different stages

- Build up cascade at several stages in one network



Wanli Ouyang, Kun Wang, Xin Zhu, Xiaogang Wang. "Chained Cascade Network for Object Detection", *Proc. ICCV*, 2017.

Model structures among classifiers at different stages



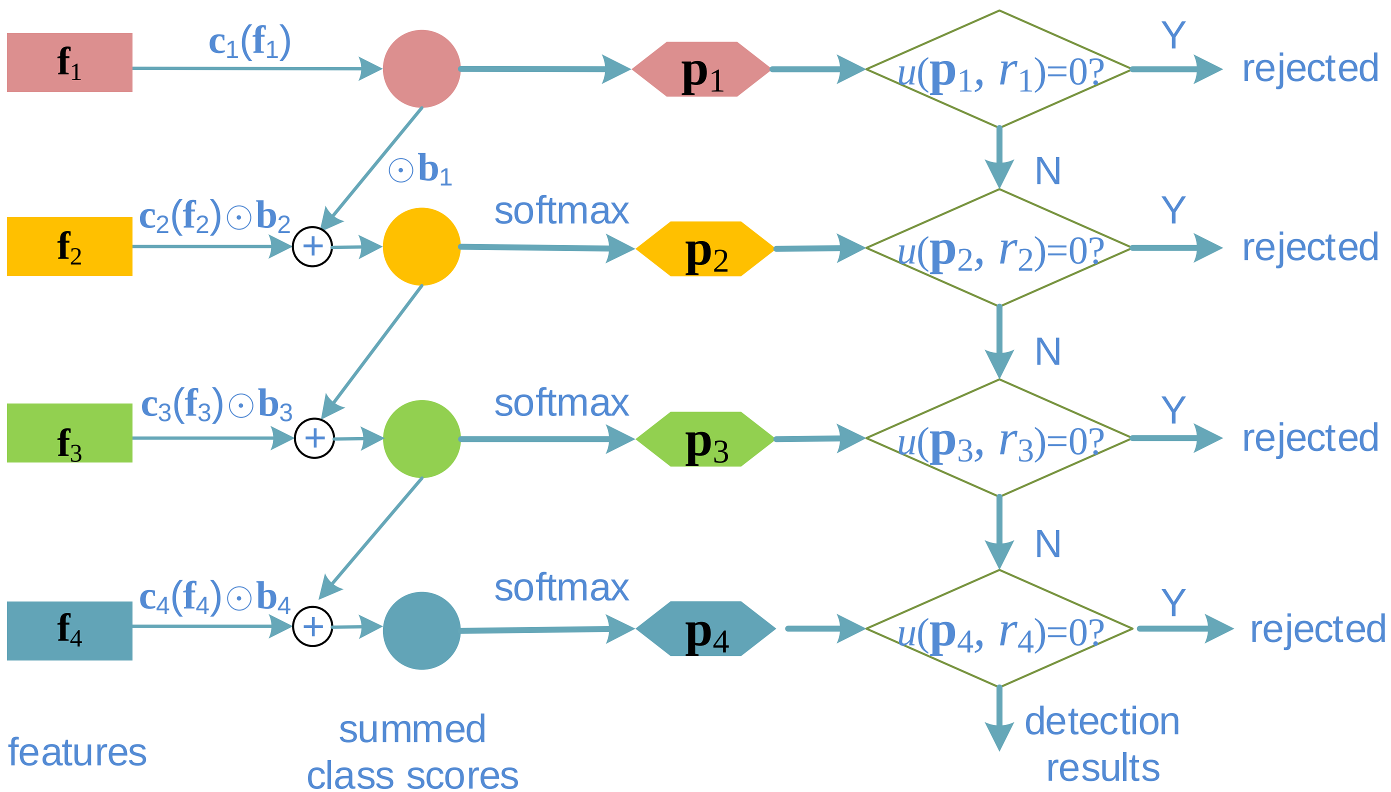
axe, axe, tooth brush

tooth brush, tooth brush, tooth brush

Wanli Ouyang, Kun Wang, Xin Zhu, Xiaogang Wang. "Chained Cascade Network for Object Detection", *Proc. ICCV*, 2017.

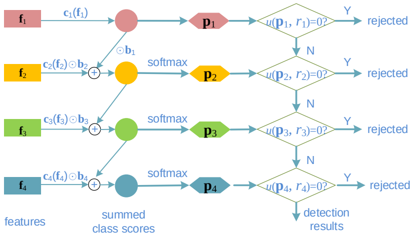
Model structures among classifiers at different stages with different context

- Build up structure among classifiers $c_i(*)$ at different stages



Wanli Ouyang, Kun Wang, Xin Zhu, Xiaogang Wang. "Chained Cascade Network for Object Detection", *Proc. ICCV*, 2017.

Experimental results



- Build up structure among classifiers $c_i(*)$ at different stages

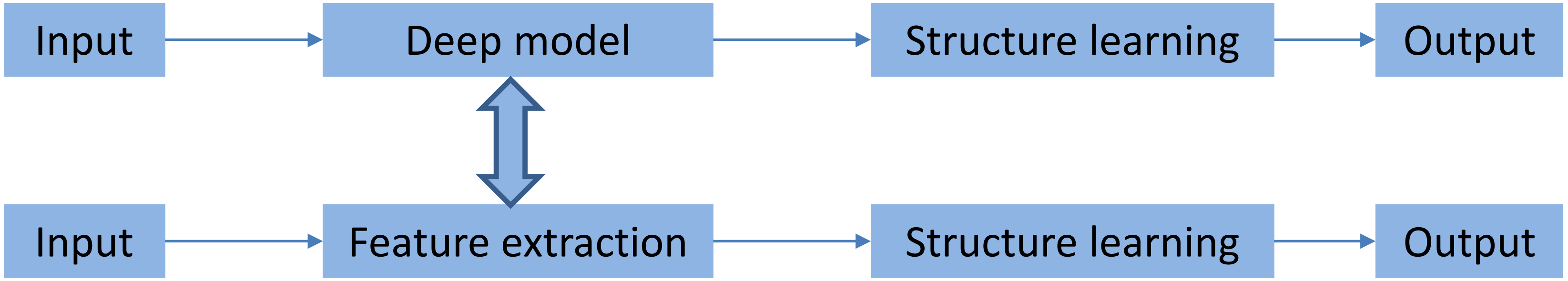
cascade?		√
chaining classifier?		√
mAP	49.4	50.9

ImageNet val2 detection mean average precision (%) with different setting on classifier chaining.

Wanli Ouyang, Kun Wang, Xin Zhu, Xiaogang Wang. "Chained Cascade Network for Object Detection", *Proc. ICCV*, 2017.

Structured output

- Treat deep model as feature extractor
- Jointly learn feature and structured output
  - Structure layer capture the structured information that cannot be modeled by conventional deep model, e.g. relationship between cascaded classifiers
  - Conventional deep model need not be influenced by the problem that can be well solved by structured model, e.g. need not be influenced by the huge amount of easy negative data



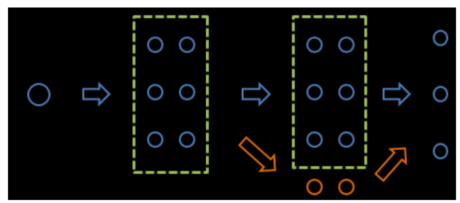
Wanli Ouyang, Kun Wang, Xin Zhu, Xiaogang Wang. "Chained Cascade Network for Object Detection", *Proc. ICCV*, 2017.

Outline

Introduction

Structured deep learning

Structured Hidden factors



Back-bone model design

Conclusion

Outline

- Introduction
- Learning features
  - Learning Feature Pyramids (ICCV17)
- Learning
  - Structure of output
  - Structured Hidden factors
    - Joint deep learning for pedestrian detection (ICCV13)
    - Deep-ID Net for object detection (T-PAMI16)
    - Mutual Learning Mutual Visibility Relationship for pedestrian detection (IJCV16)
  - Structure of features
- Conclusion

Child | Tooth brush | Woman | Tooth brush

Object detection

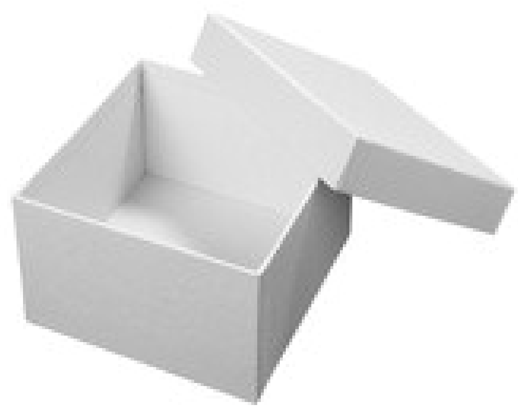
Challenges -- person

-
- Occlusion
- Deformation

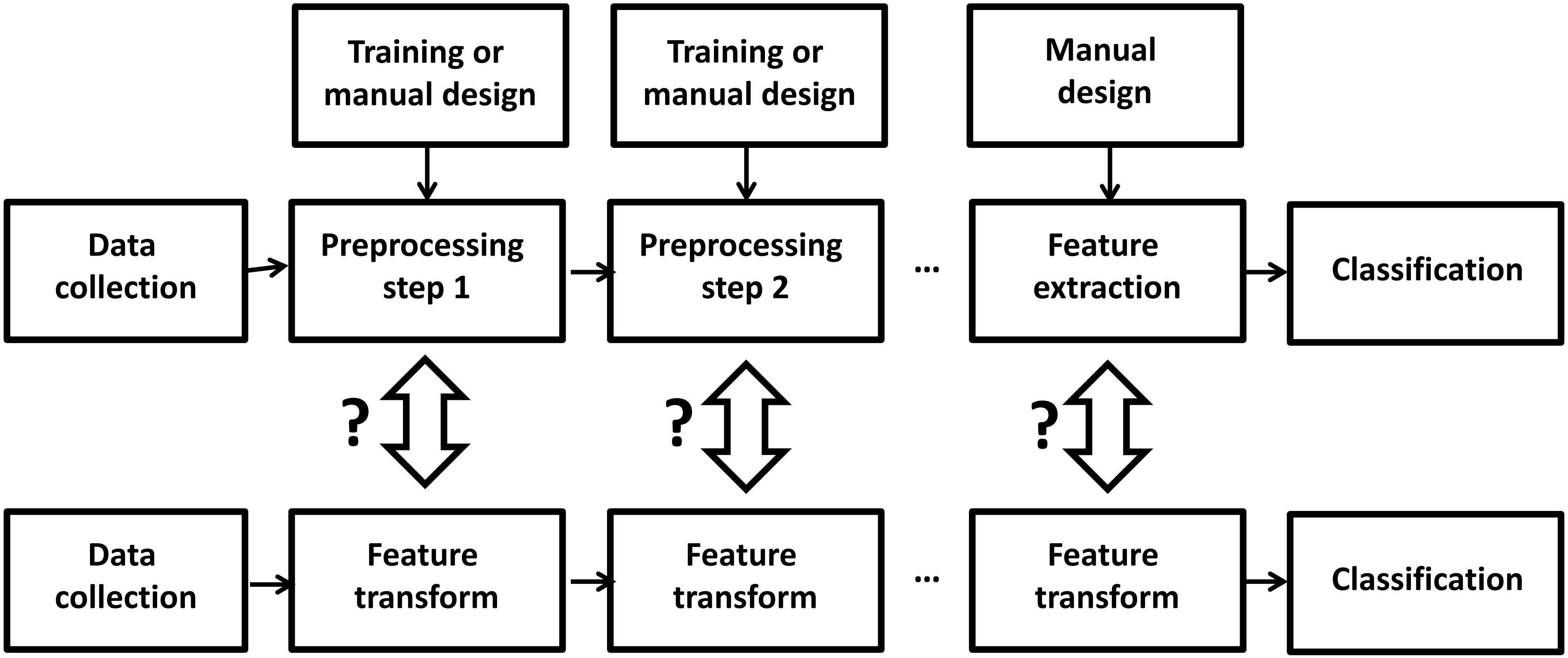
No annotation

Hidden

Is deep model a black box?



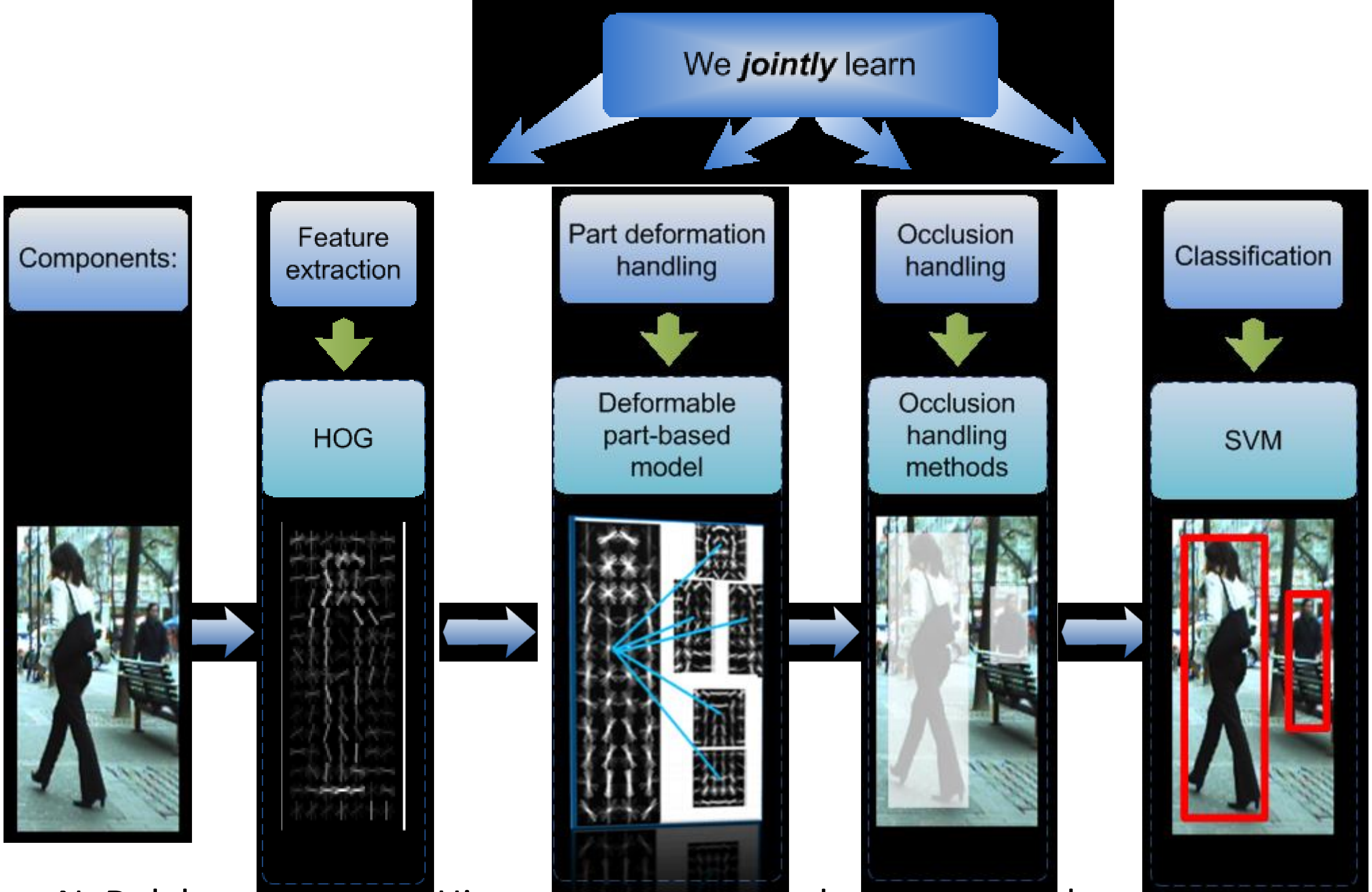
Joint Learning vs Separate Learning



End-to-end learning

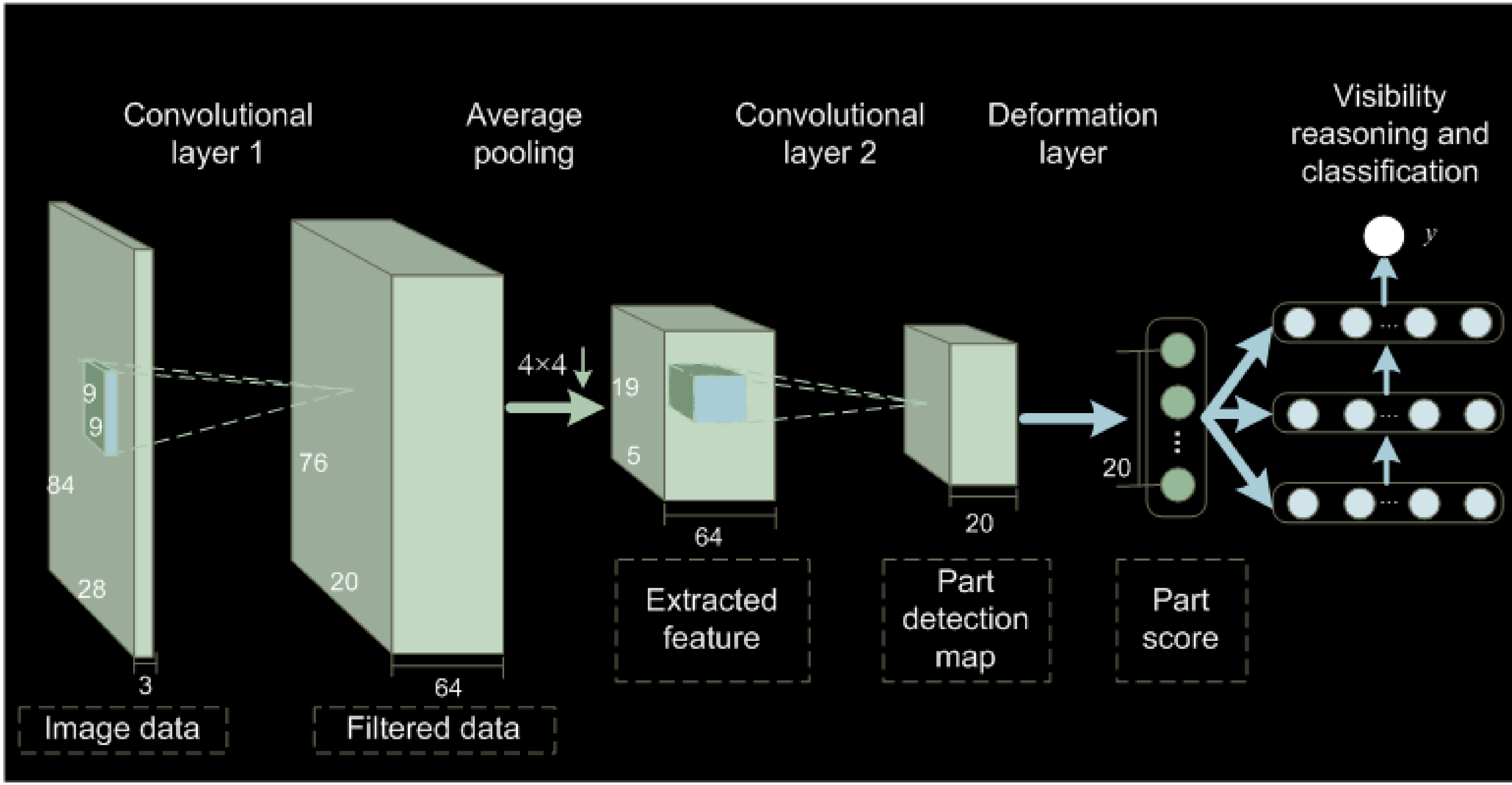
Deep learning is a framework/language but not a black-box model

Its power comes from joint optimization and increasing the capacity of the learner



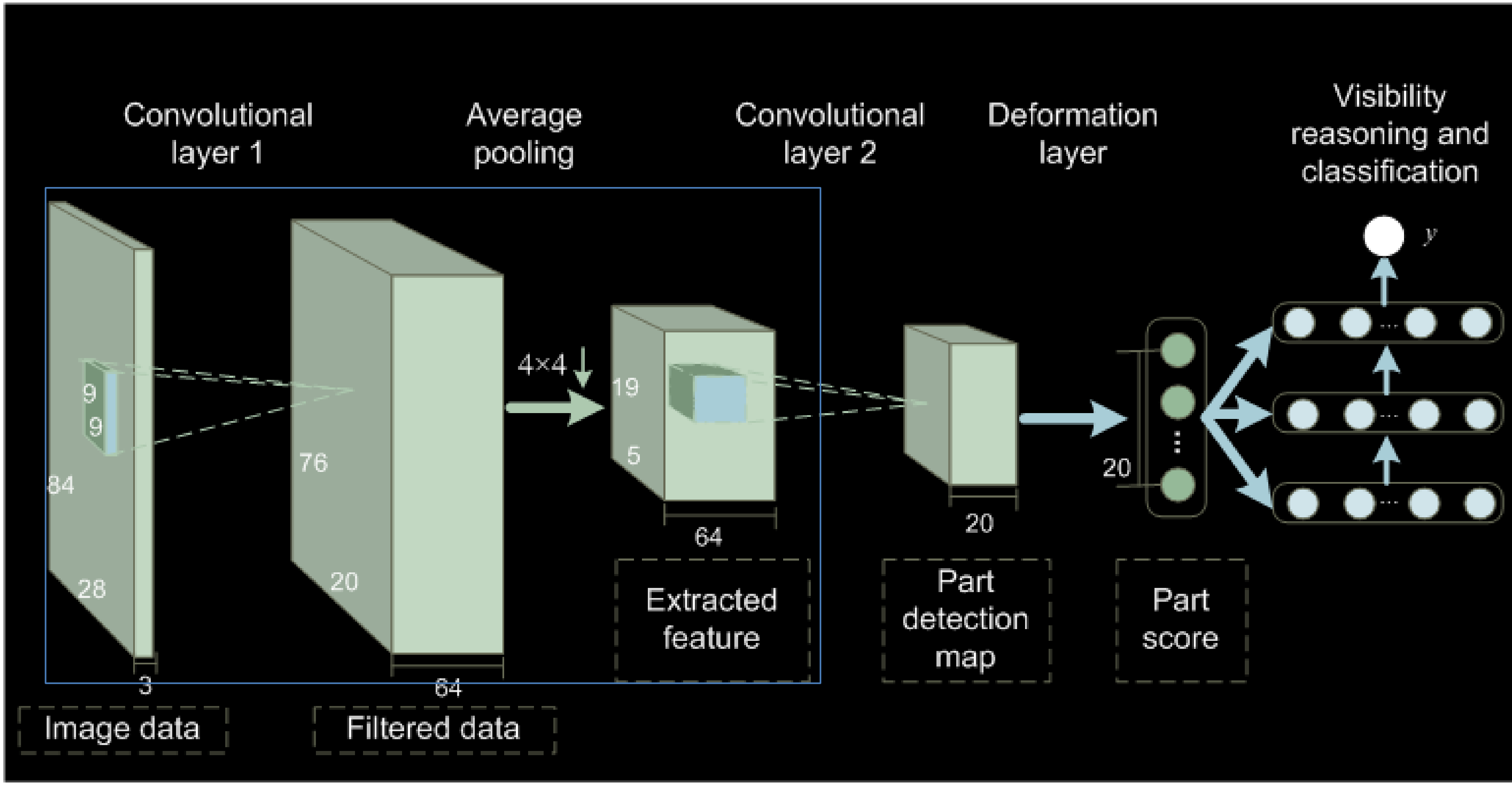
- N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. CVPR, 2005. (10,000+ citations)
- P. Felzenszwalb, D. McAlester, and D. Ramanan. A Discriminatively Trained, Multiscale, Deformable Part Model. CVPR, 2008. (4000+ citations)
- W. Ouyang and X. Wang. A Discriminative Deep Model for Pedestrian Detection with Occlusion Handling. CVPR, 2012.

Our Joint Deep Learning Model



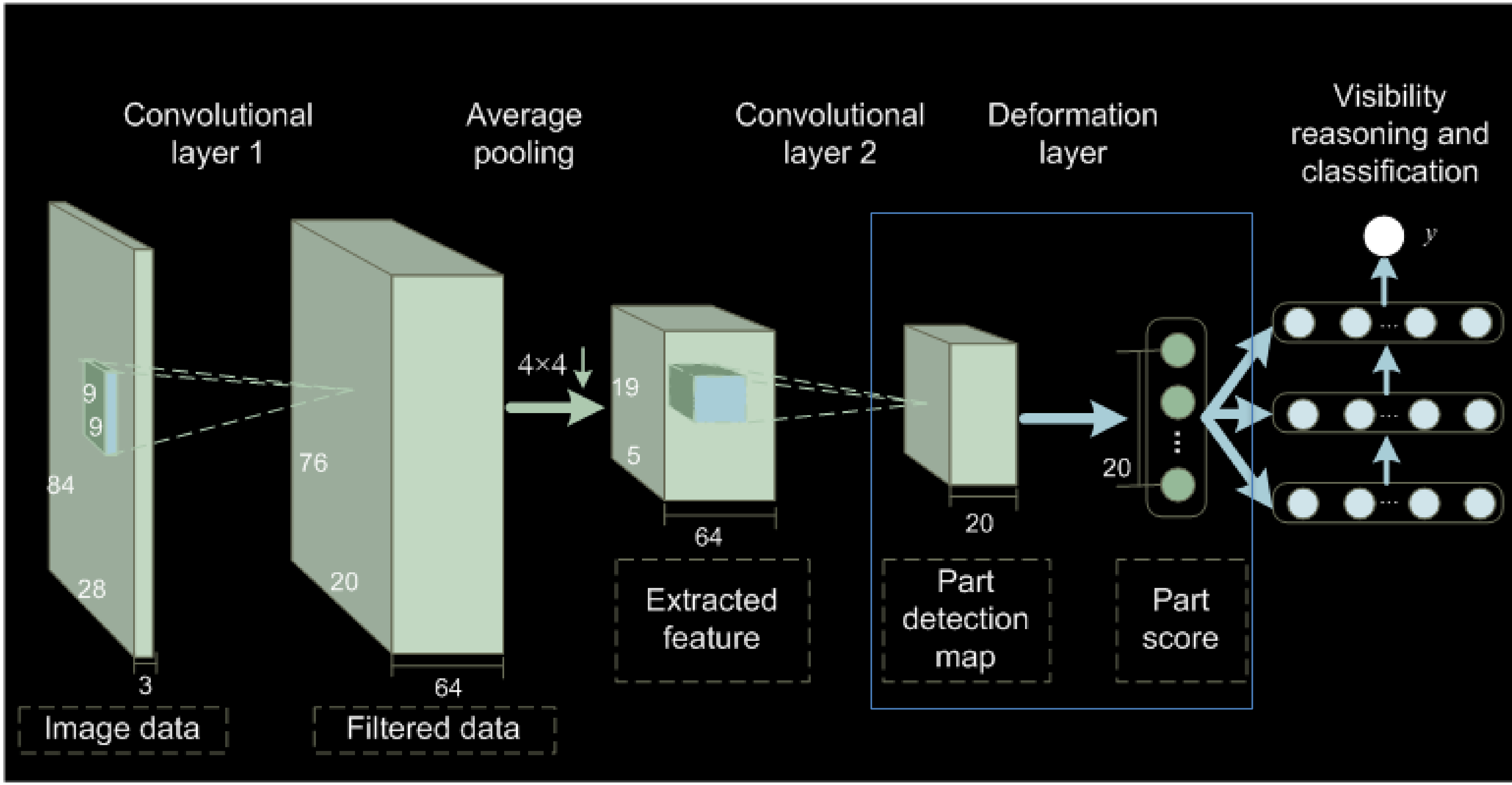
W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," Proc. ICCV, 2013.

Our Joint Deep Learning Model



W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," Proc. ICCV, 2013.

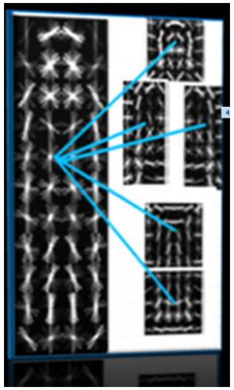
Our Joint Deep Learning Model



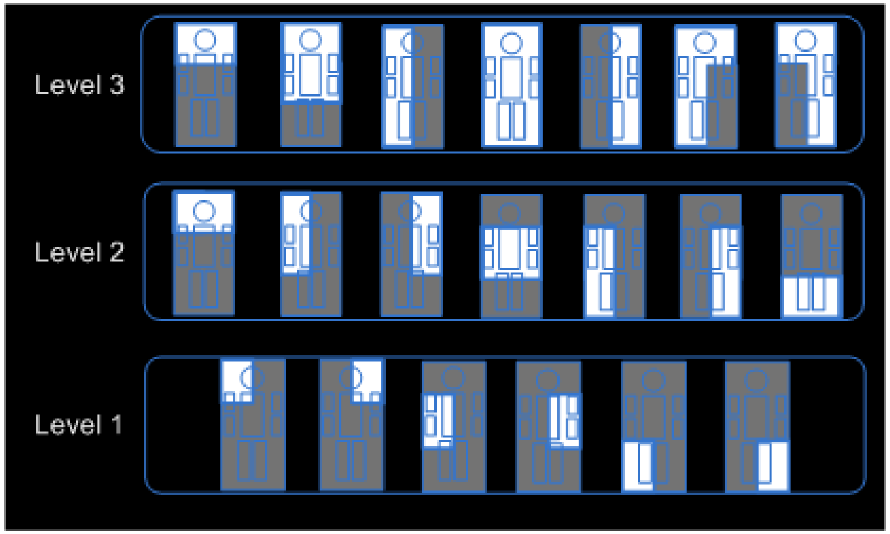
W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," Proc. ICCV, 2013.

Modeling Part Detectors

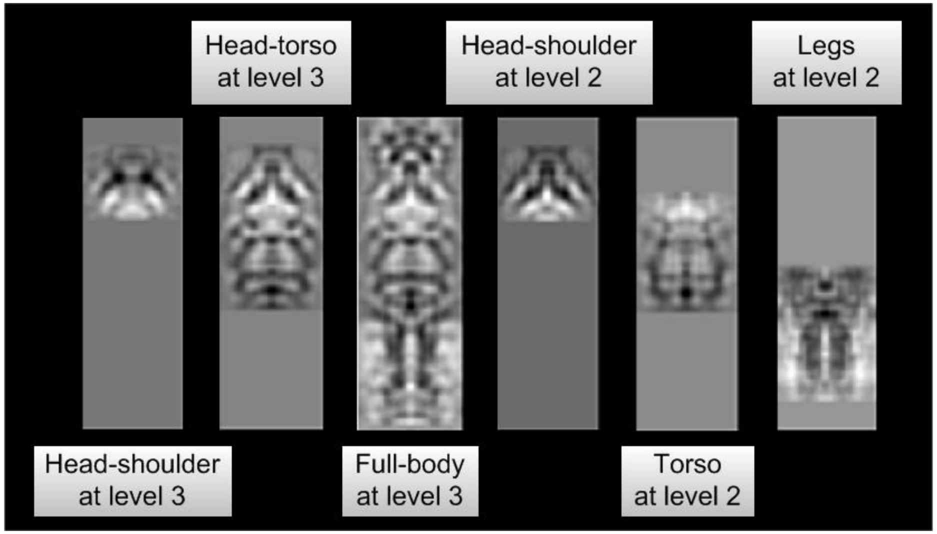
- Design the filters in the second convolutional layer with variable sizes



Part models learned from HOG



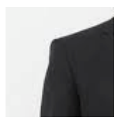
Part models

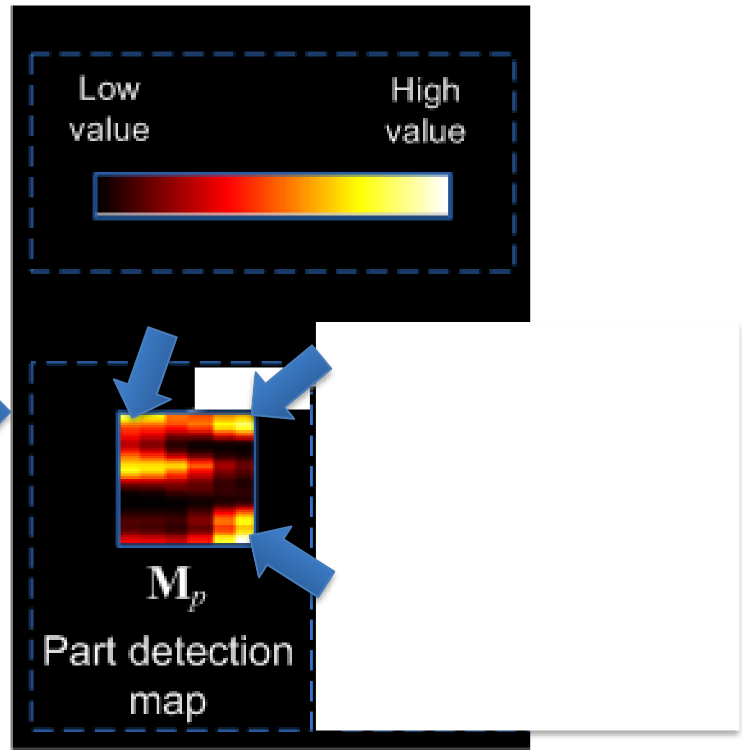


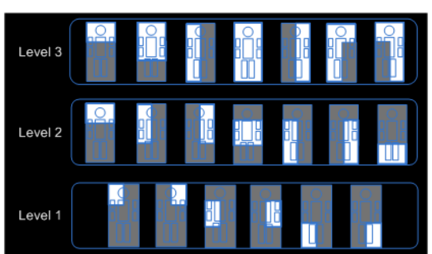
Learned filtered at the second convolutional layer

Deformation Layer

- Infer the location of object parts

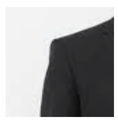


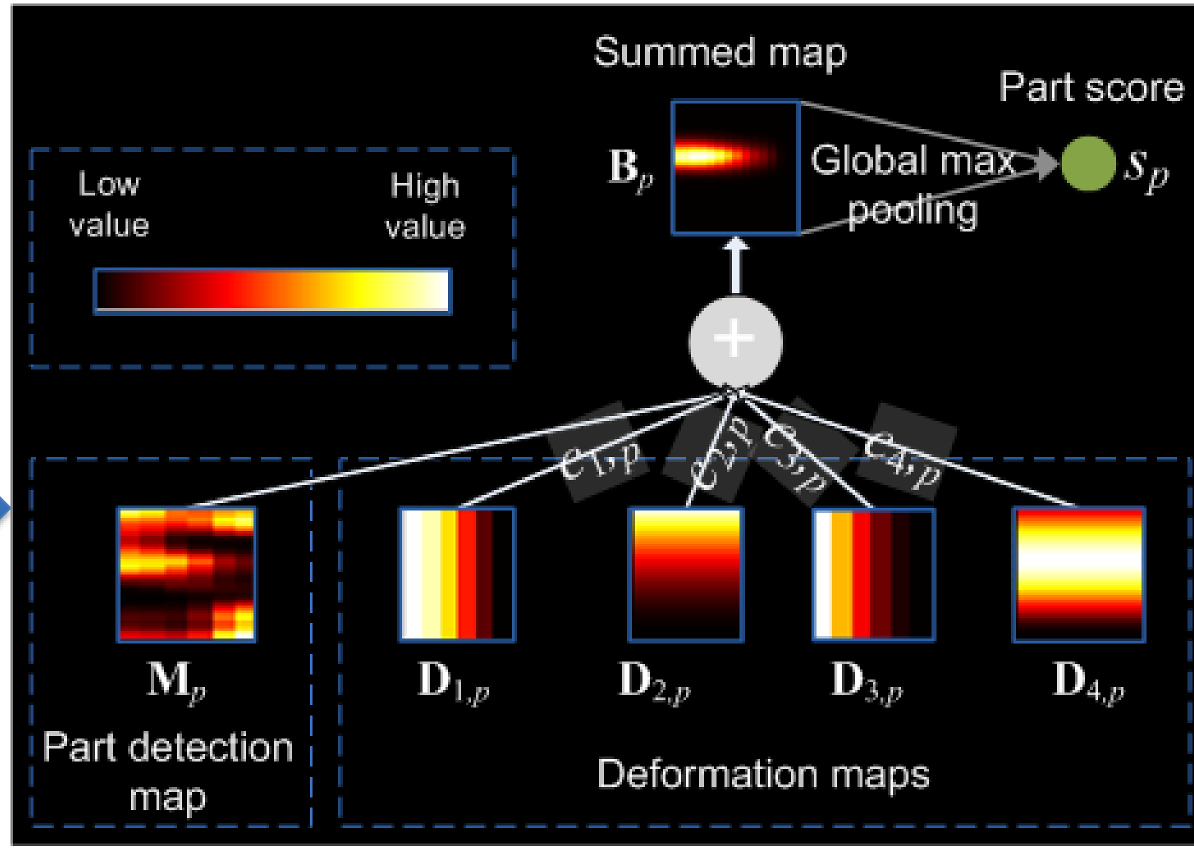


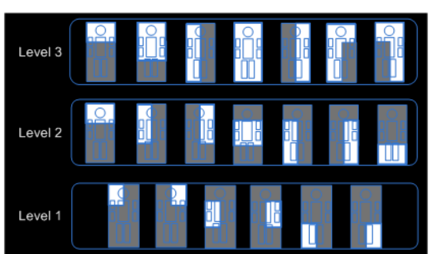


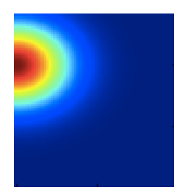
Deformation Layer

- Infer the location of object parts

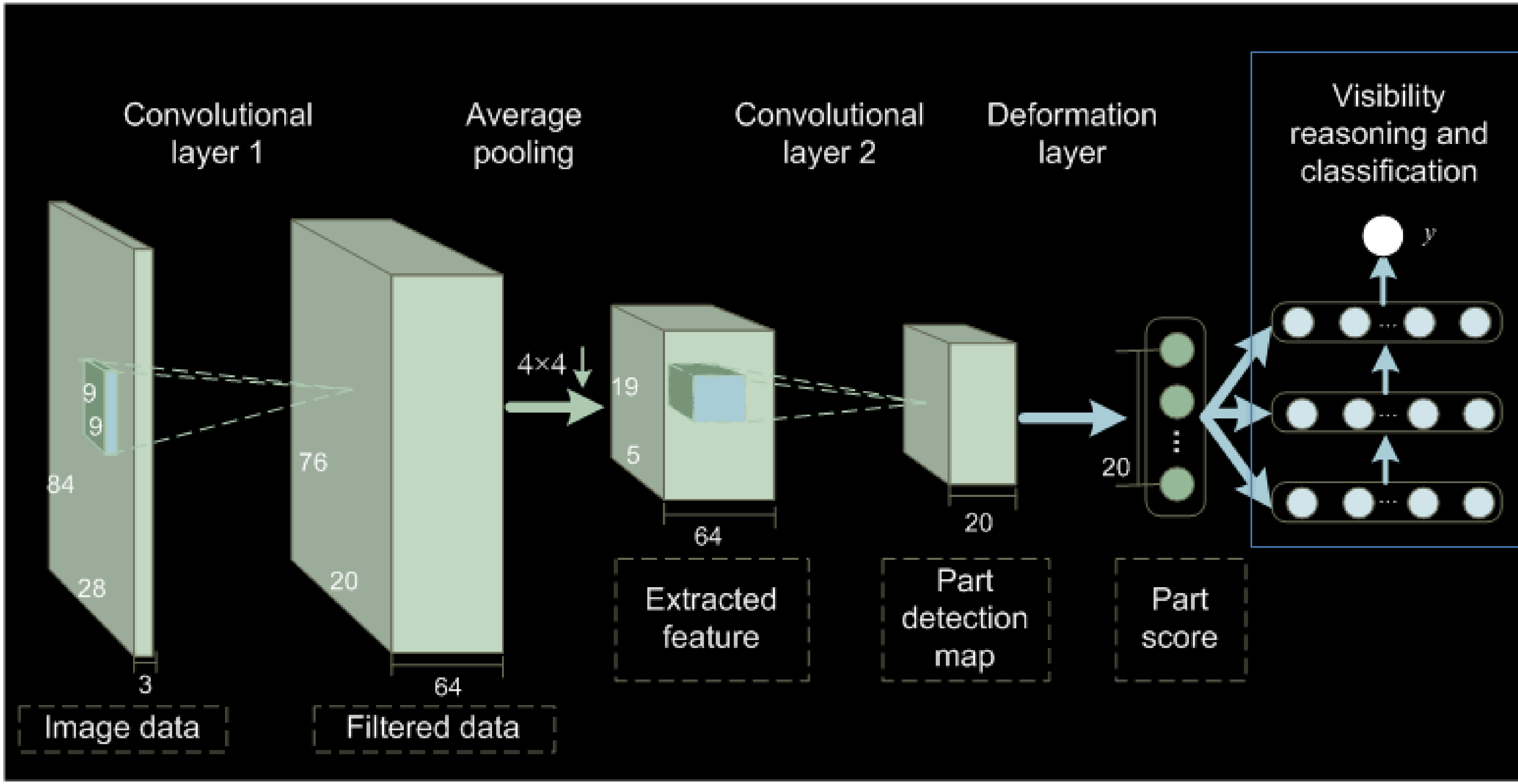








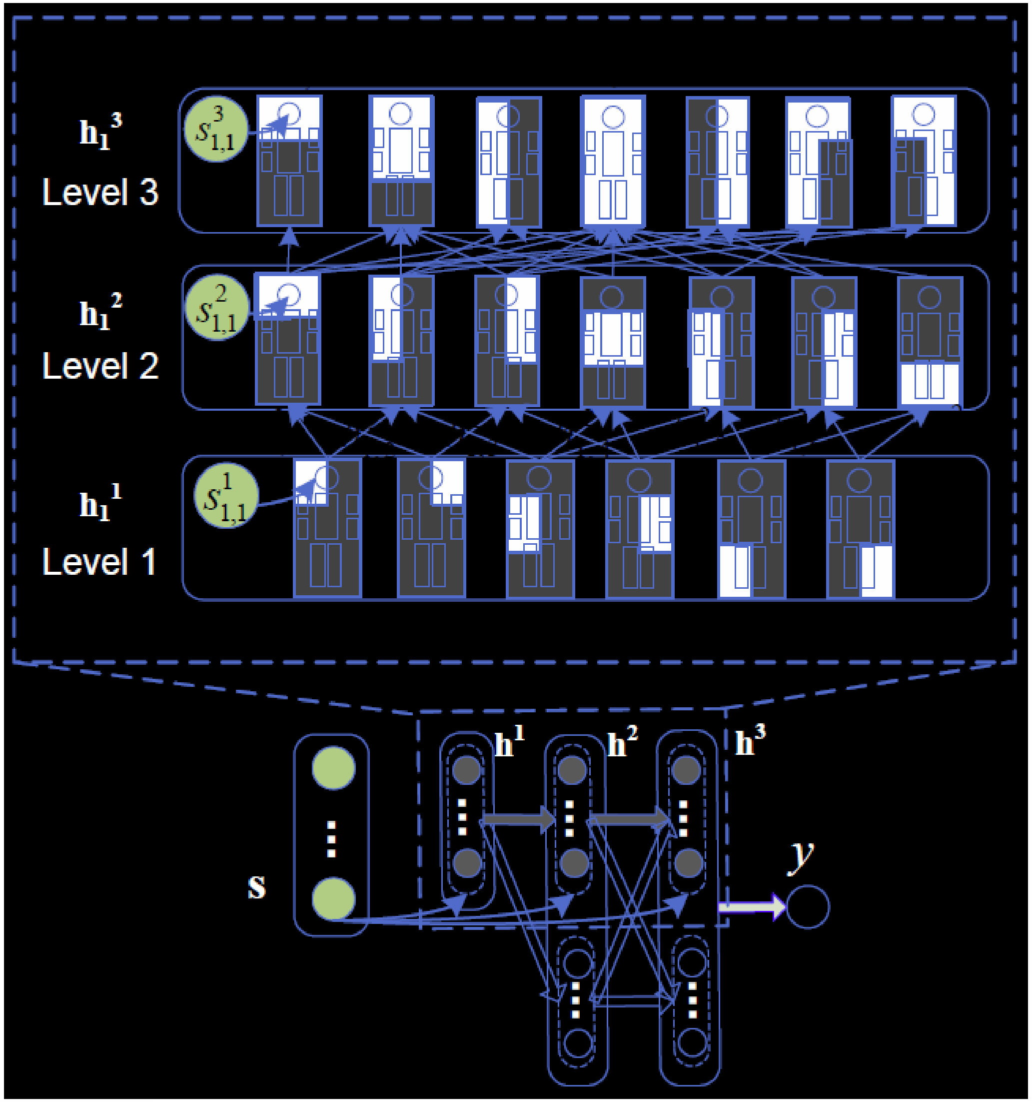
Our Joint Deep Learning Model



W. Ouyang and X. Wang, "Joint Deep Learning for Pedestrian Detection," Proc. ICCV, 2013.

Visibility Reasoning with Deep Belief Net

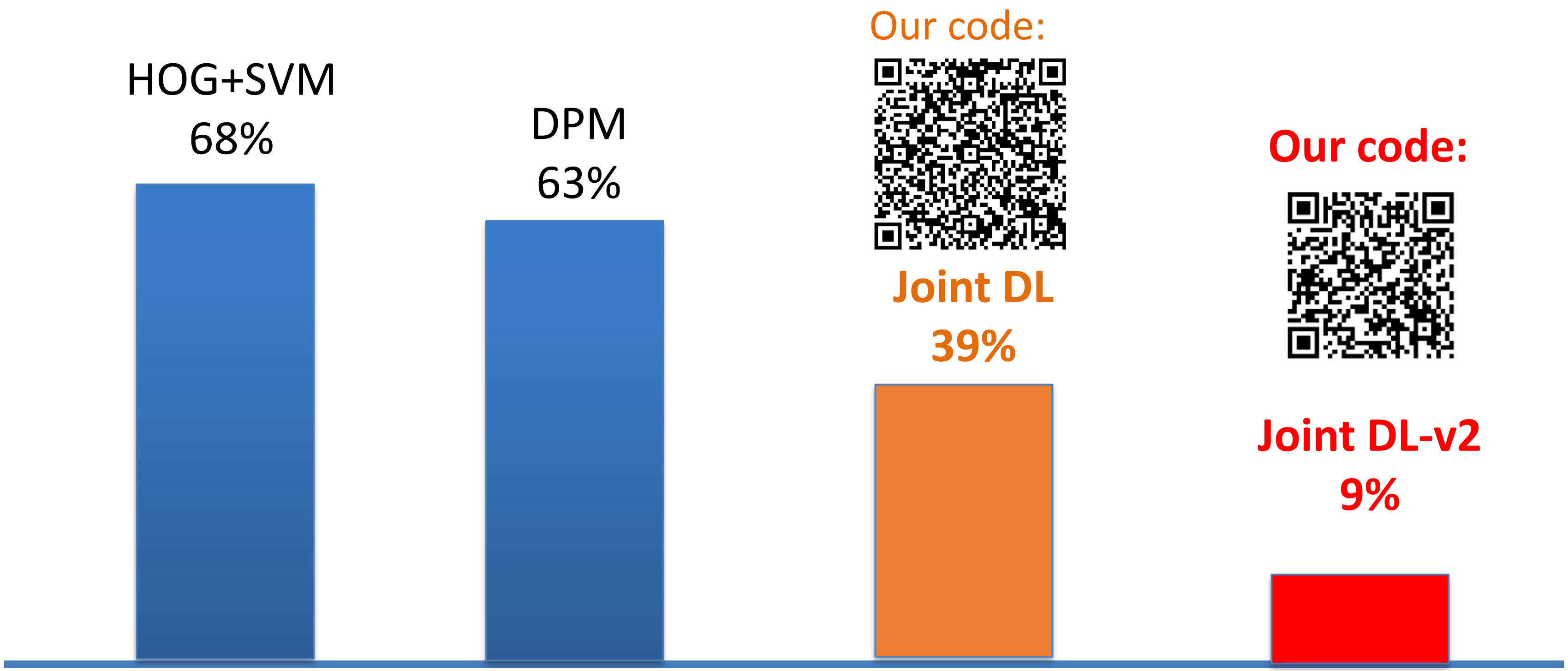
- Infer the visibility of object parts



$$\tilde{h}_j^{l+1} = \sigma(\tilde{\mathbf{h}}^{l\mathrm{T}} \mathbf{w}_{*,j}^{l} + c_j^{l+1} + g_j^{l+1} s_j^{l+1})$$

Correlates with part detection score

Pedestrian Detection on Caltech (average miss detection rates)



W. Ouyang and X. Wang, “Joint Deep Learning for Pedestrian Detection,” ICCV 2013.

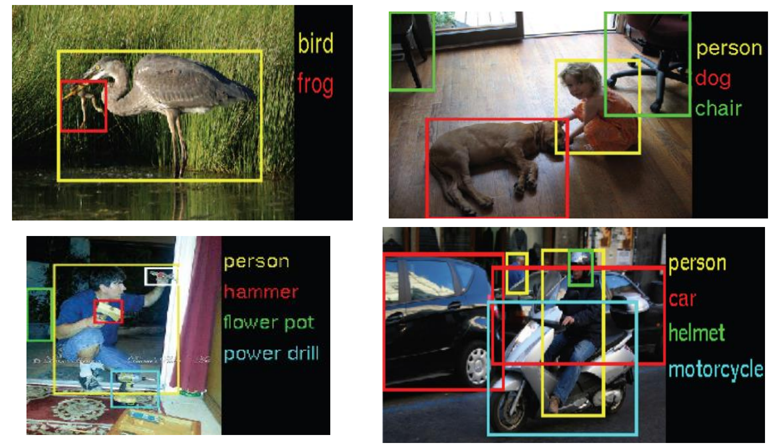
W. Ouyang et. al, “Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection,” TPAMI, accepted.

Generalize from single pedestrian to multiple pedestrians

Single pedestrian

Deformation

Visibility



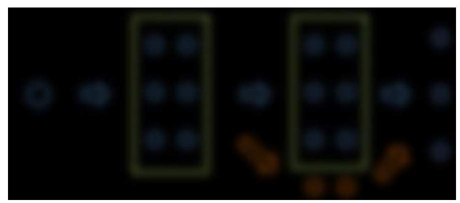
Generic Object detection
TPAMI’17 (most popular)

Multiple pedestrians
IJCV’16

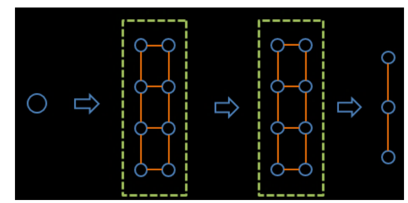
Outline

Introduction

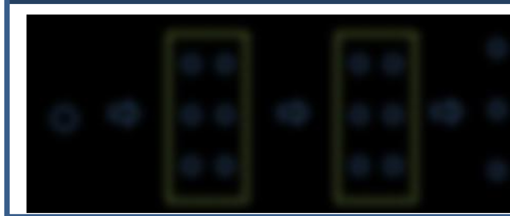
Structured deep learning



Structured features



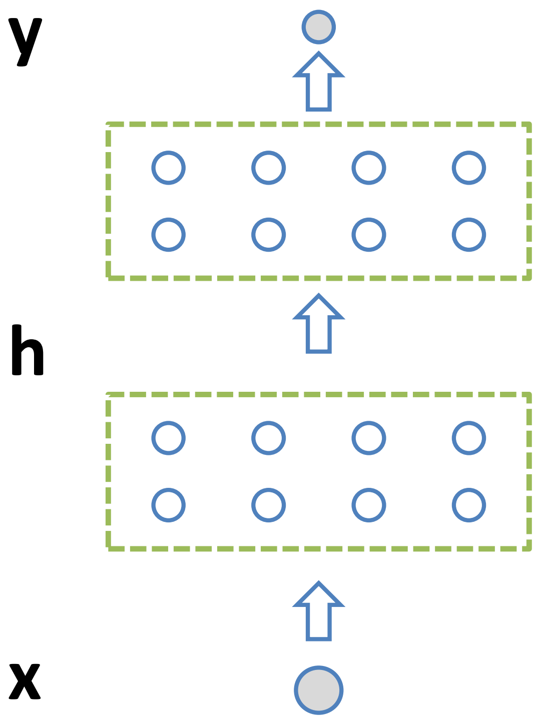
Back-bone model design



Conclusion

Structure in neurons

- Conventional neural networks
  - Neurons in the same layer have no connection
  - Neurons in adjacent layers are fully connected, at least within a local region



Structure exists in brain

Outline

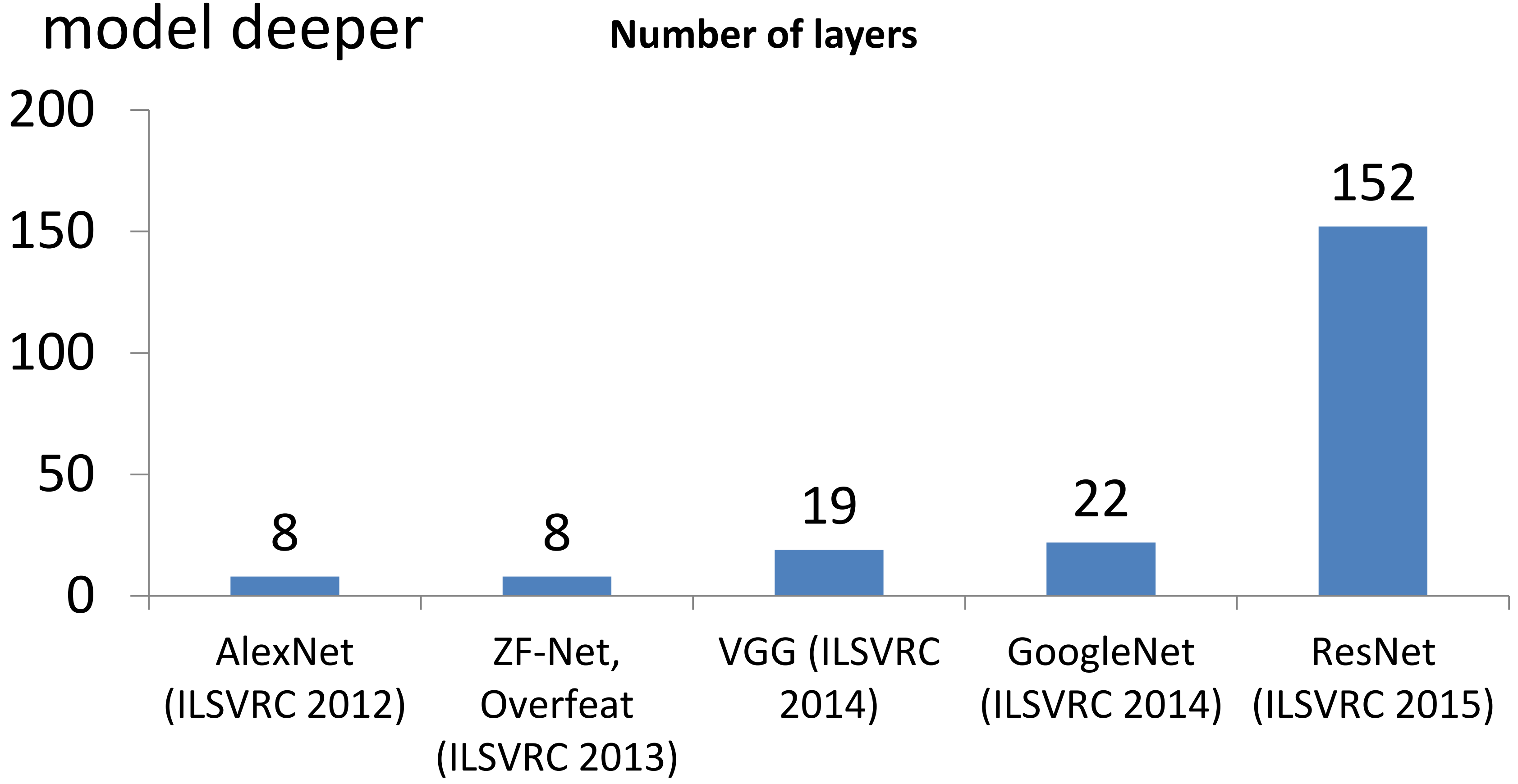
- Structure of features
  - GBD-Net for Object detection (ECCV16)
  - Structured feature learning for pose estimation (CVPR16)
  - CRF-CNN for pose estimation (NIPS 16)
  - Attention-Gated CRFs for Contour Prediction (NIPS17)
  - Scene Graph Generation from Objects, Phrases and Region Captions (ICCV17)

Child | Tooth brush | Woman | Tooth brush

Object detection

Message from past ImageNet Challenge

- Design a good learning strategy (VGG, BN) or a good branching structure (Inception, ResNet) to make the model deeper



Message from past ImageNet Challenge

- Design a good learning strategy (VGG, BN) or a good branching structure (Inception, ResNet) to make the model deeper

Is deeper the only way to go?

What can our vision researchers’ observation help?

What can our vision researchers' observation help?

GBD-Net

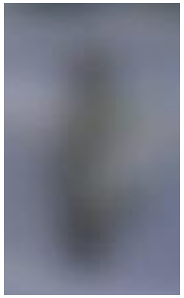
What can our vision researchers' observation help?

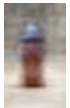
GBD-Net

Context

Context

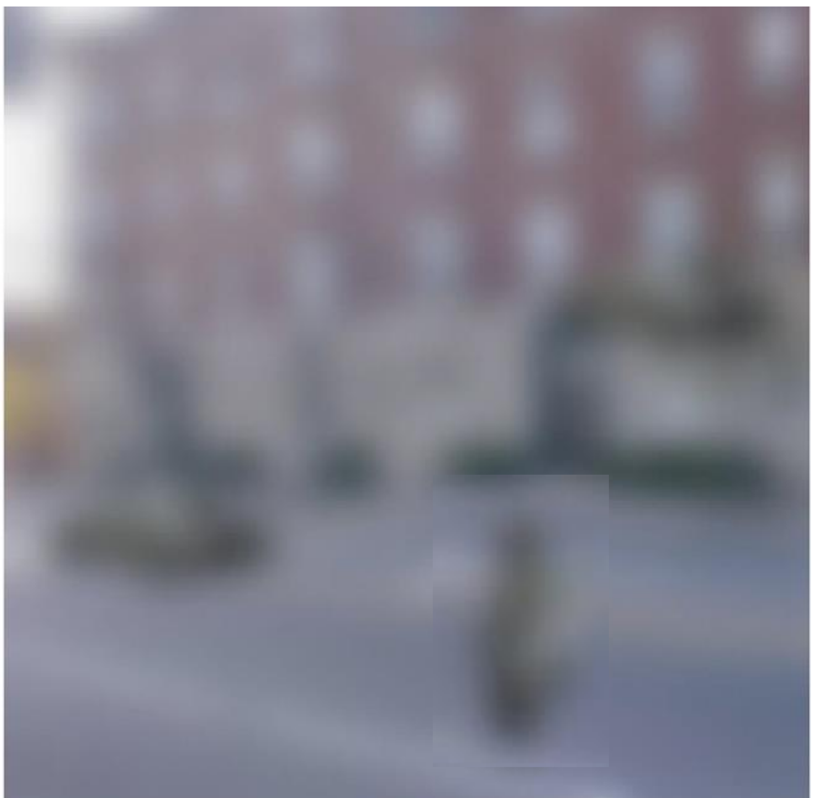
- Visual context helps to identify objects





Context

- Visual context helps to identify objects



TRENDS in Cognitive Sciences

TRENDS in Cognitive Sciences

Motivation

- With the deep model, what can we do for context?

Motivation

- With the deep model, what can we do for context?
- Learning relationship among features of different resolutions and contextual regions.

Motivation

- With the deep model, what can we do for context?
- Learning relationship among features of different resolutions and contextual regions.

Rabbit ear

Rabbit head

Motivation

- With the deep model, what can we do for context?
- Learning relationship among features of different resolutions and contextual regions.
  - Features of different contextual regions validate each other

Rabbit ear

Rabbit head

Motivation

- With the deep model, what can we do for context?
- Learning relationship among features of different resolutions and contextual regions.
  - Features of different contextual regions validate each other

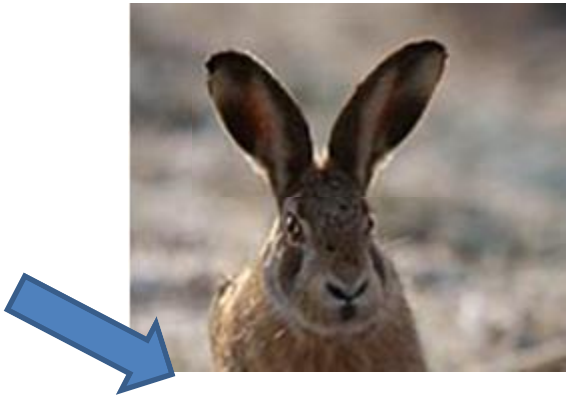
Rabbit ear

Rabbit head

Motivation

- With the deep model, what can we do for context?
- Learning relationship among features of different resolutions and contextual regions.
  - Features of different contextual regions validate each other
  - Not always true

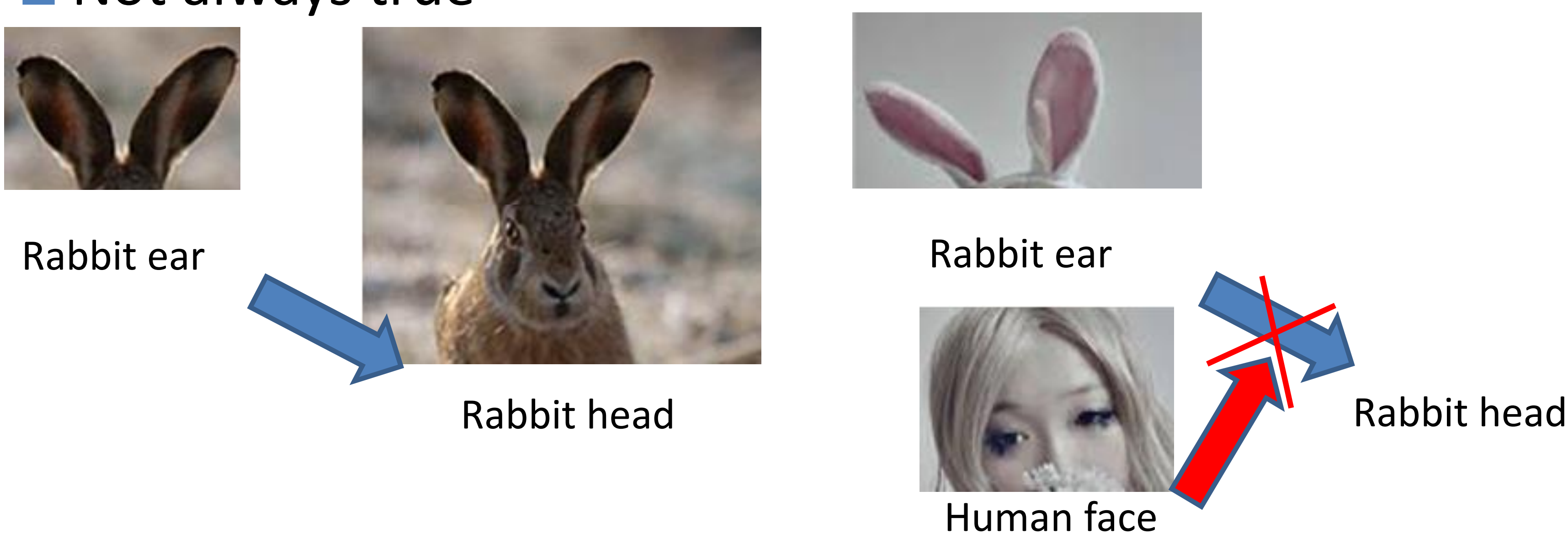
Rabbit ear



Rabbit head

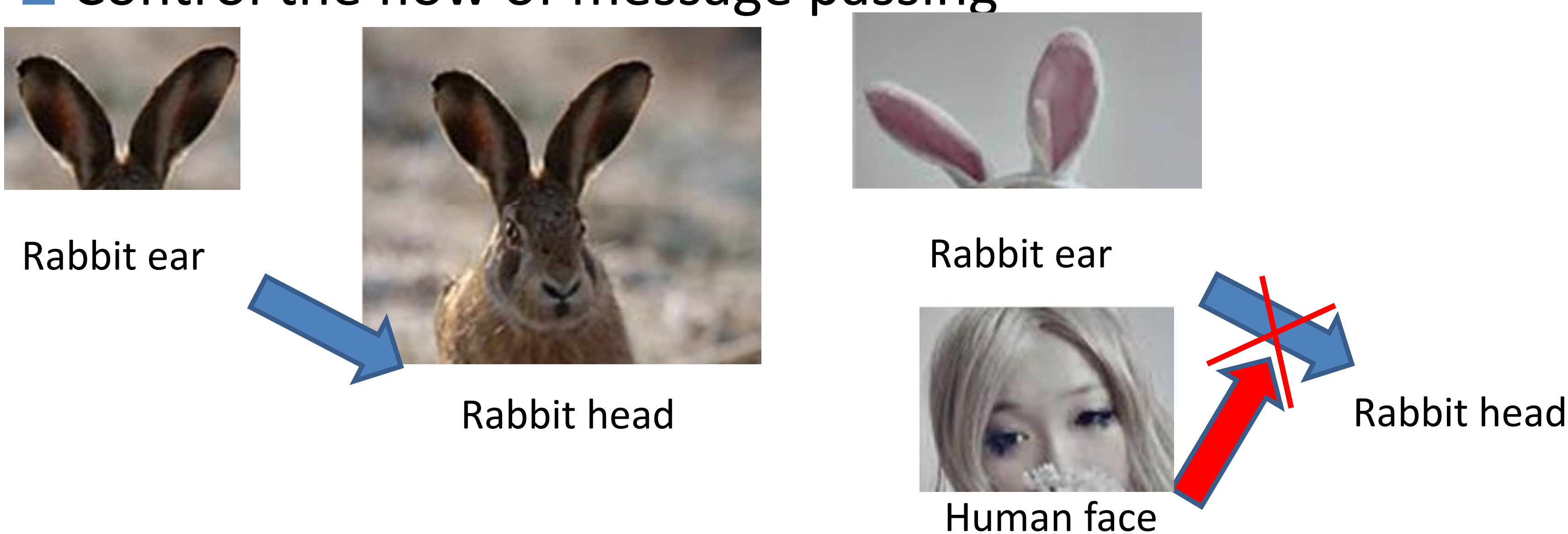
Motivation

- With the deep model, what can we do for context?
- Learning relationship among features of different resolutions and contextual regions.
  - Features of different contextual regions validate each other
  - Not always true

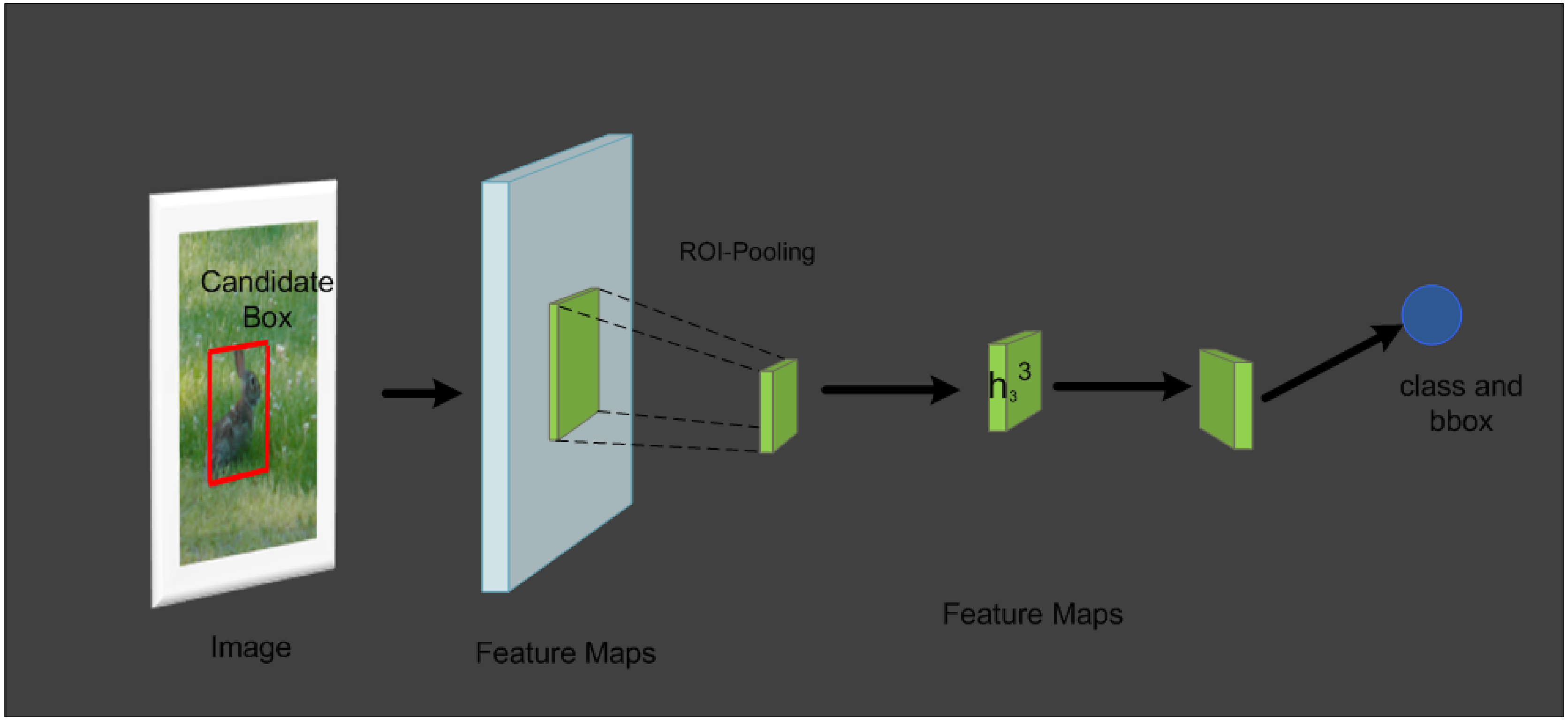


Motivation

- With the deep model, what can we do for context?
- Learning relationship among features of different resolutions and contextual regions.
  - Features of different contextual regions validate each other
  - Control the flow of message passing

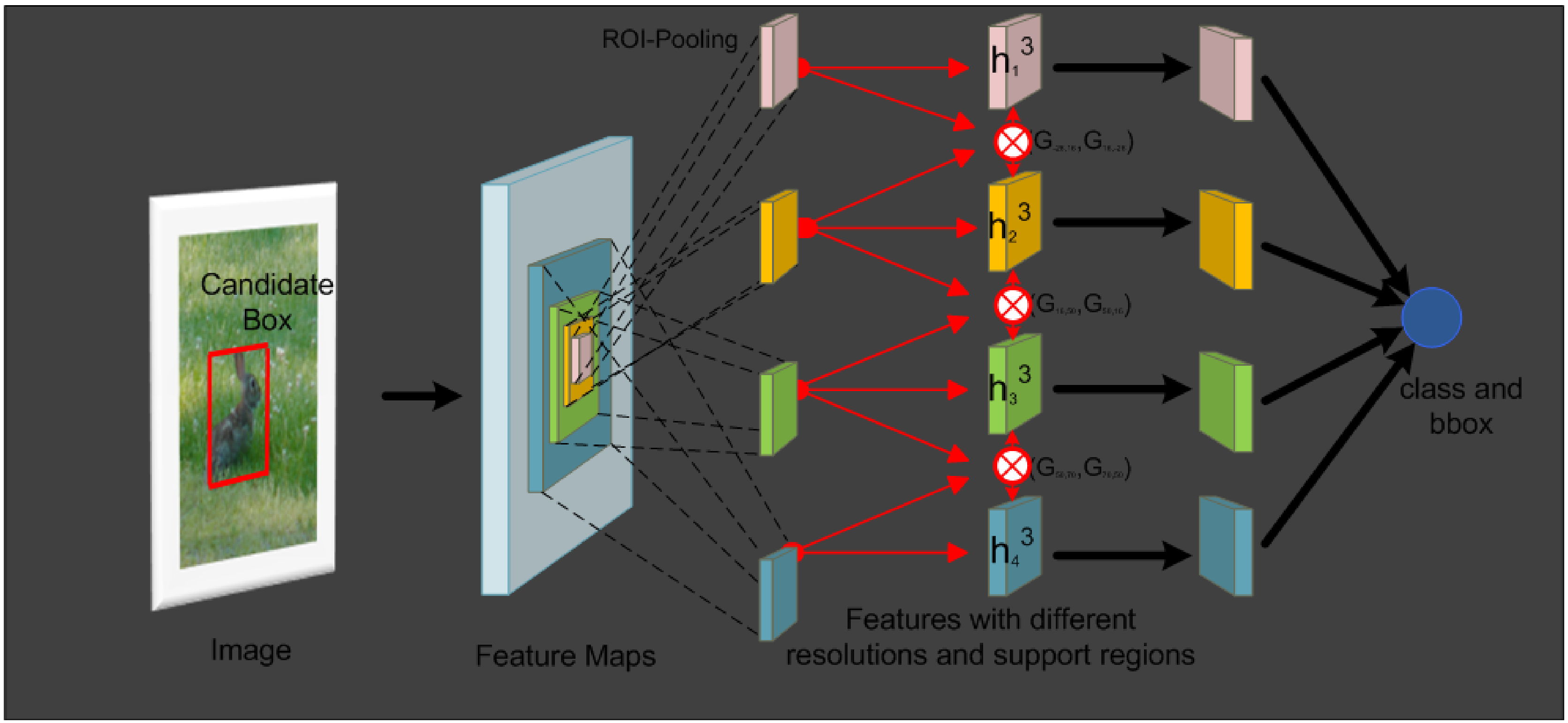


Fast R-CNN

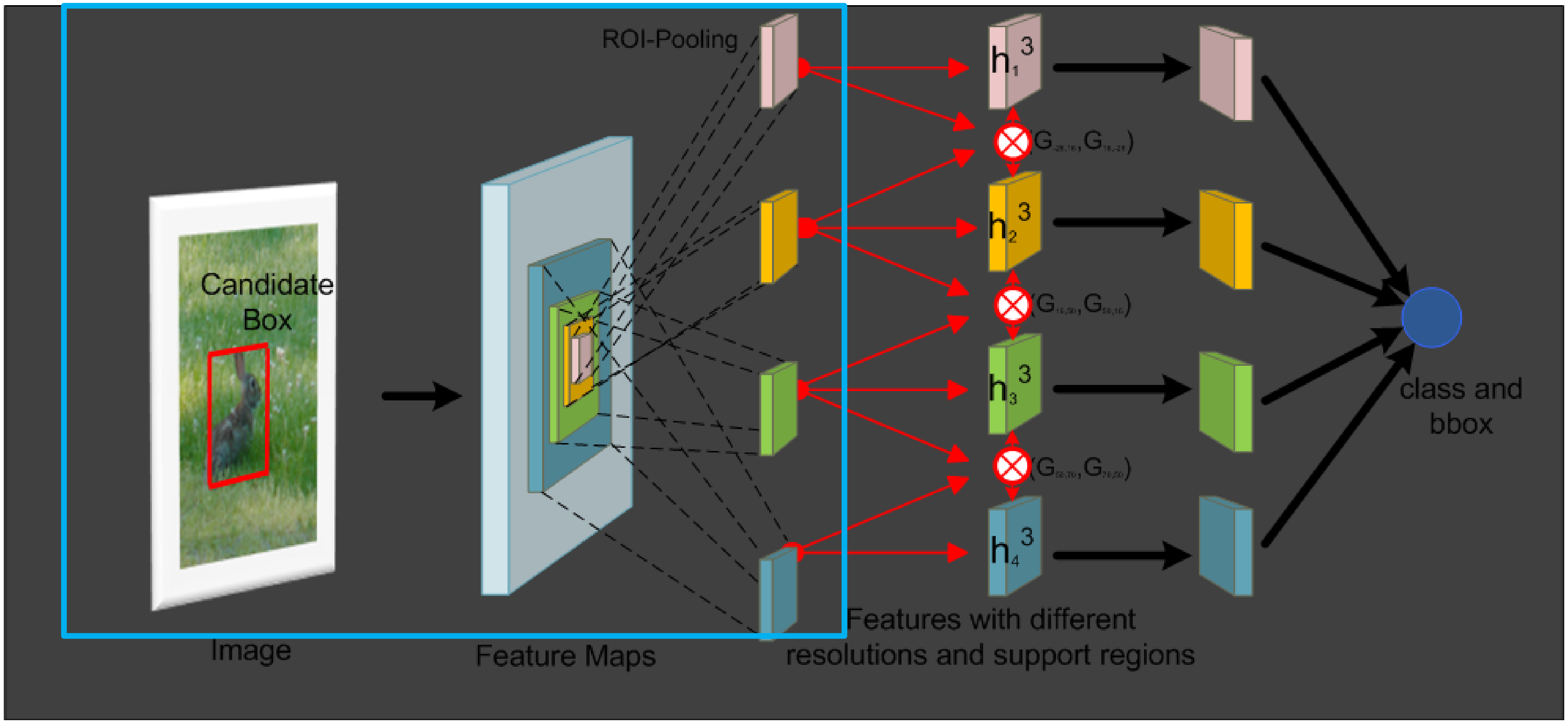


Candidate
Box
ROI-Pooling
h3 3
class and
bbox
Image
Feature Maps
Feature Maps

Gated bi-directional CNN (GBD-Net)

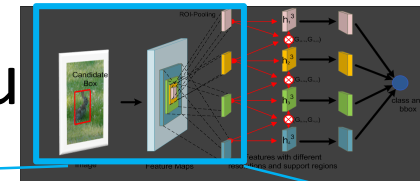


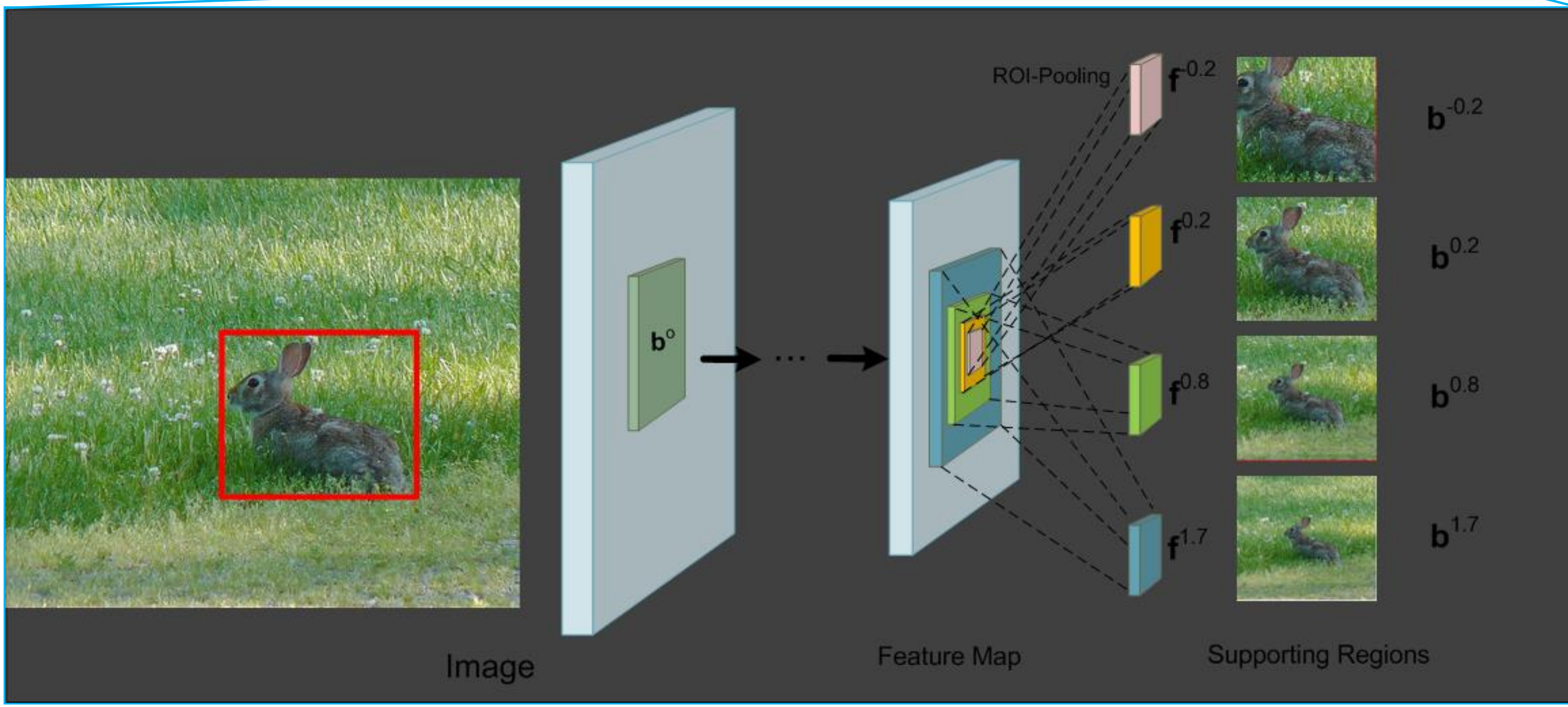
Gated bi-directional CNN (GBD-Net)



Gated bi-directional CNN (GBD-Net)

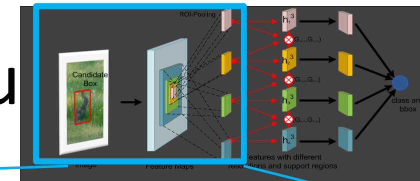
- Features of different context and resolu

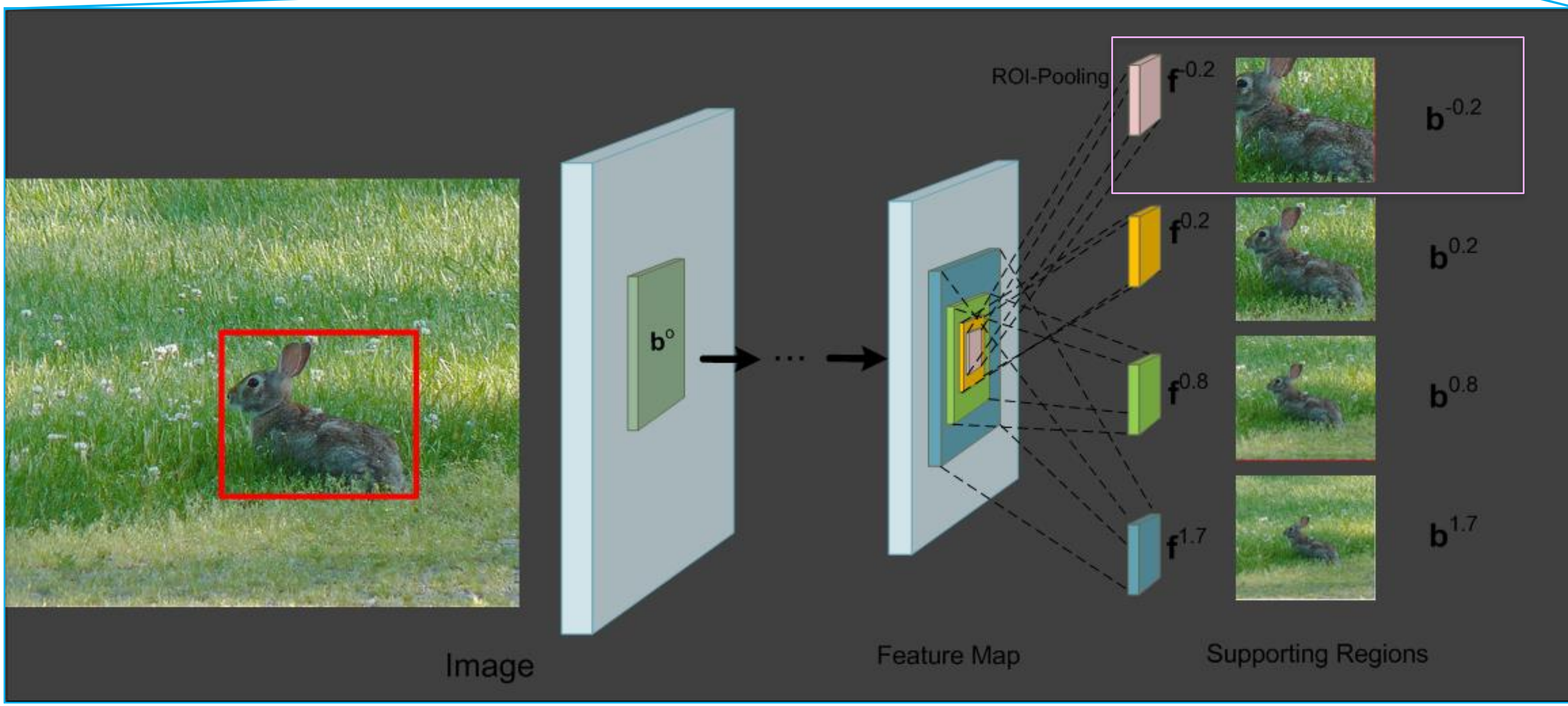




Gated bi-directional CNN (GBD-Net)

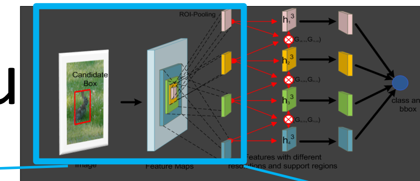
- Features of different context and resolu

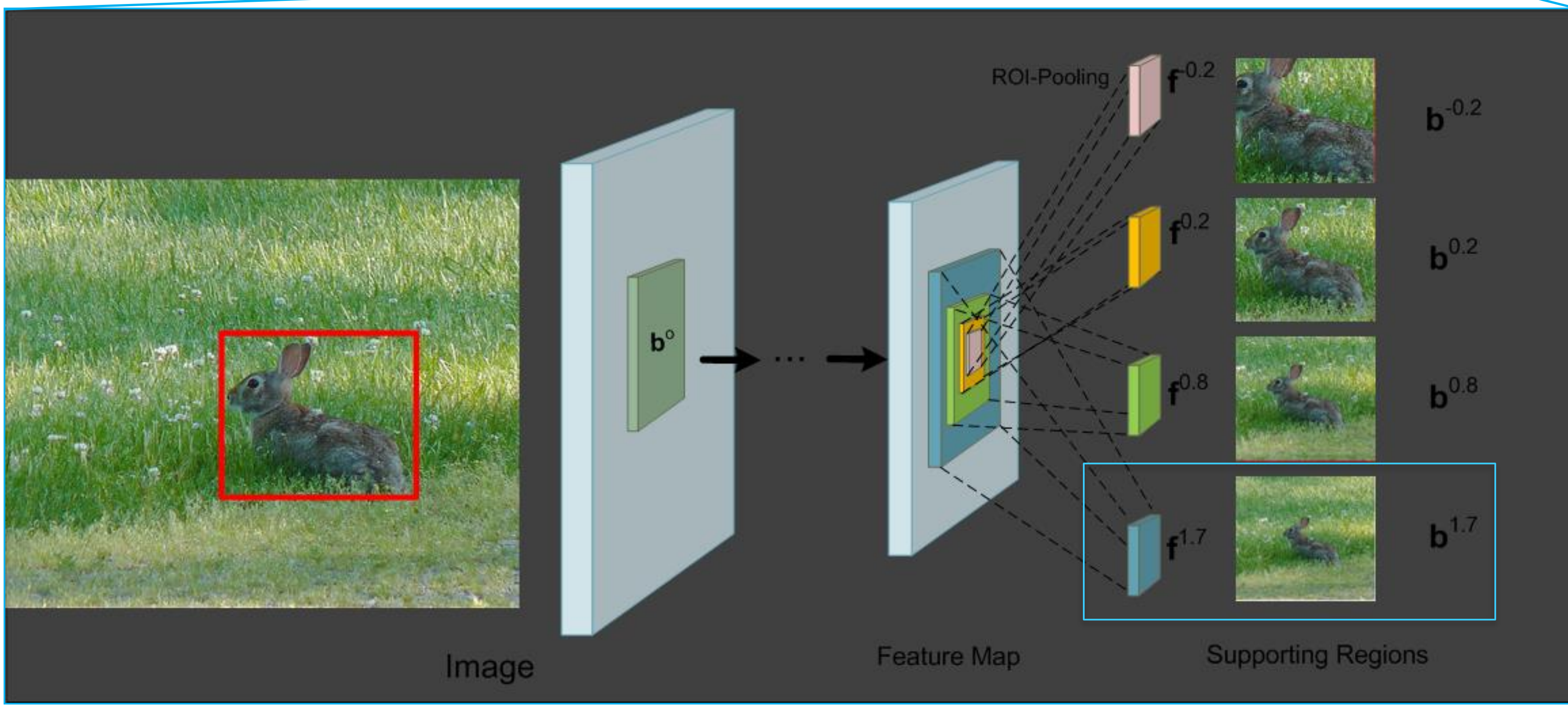




Gated bi-directional CNN (GBD-Net)

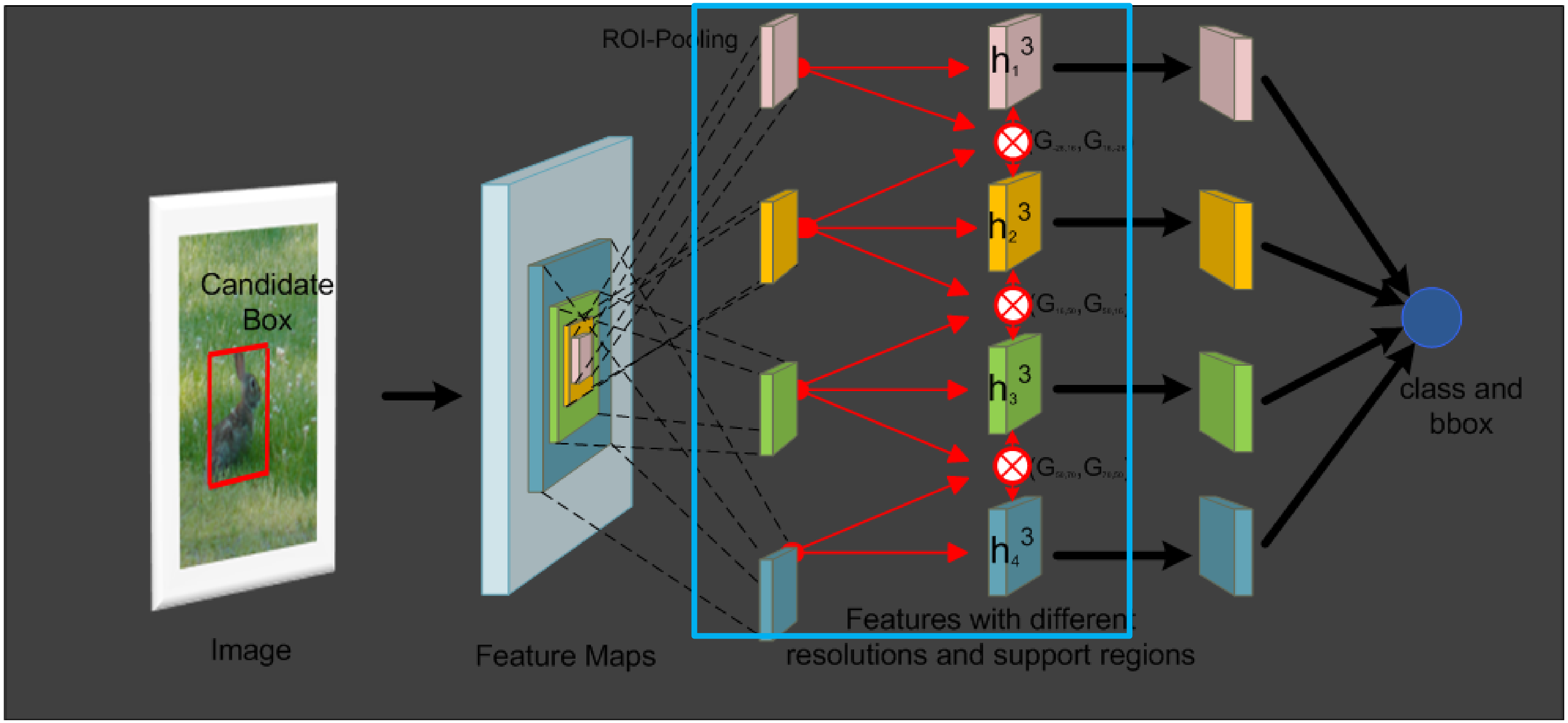
- Features of different context and resolu



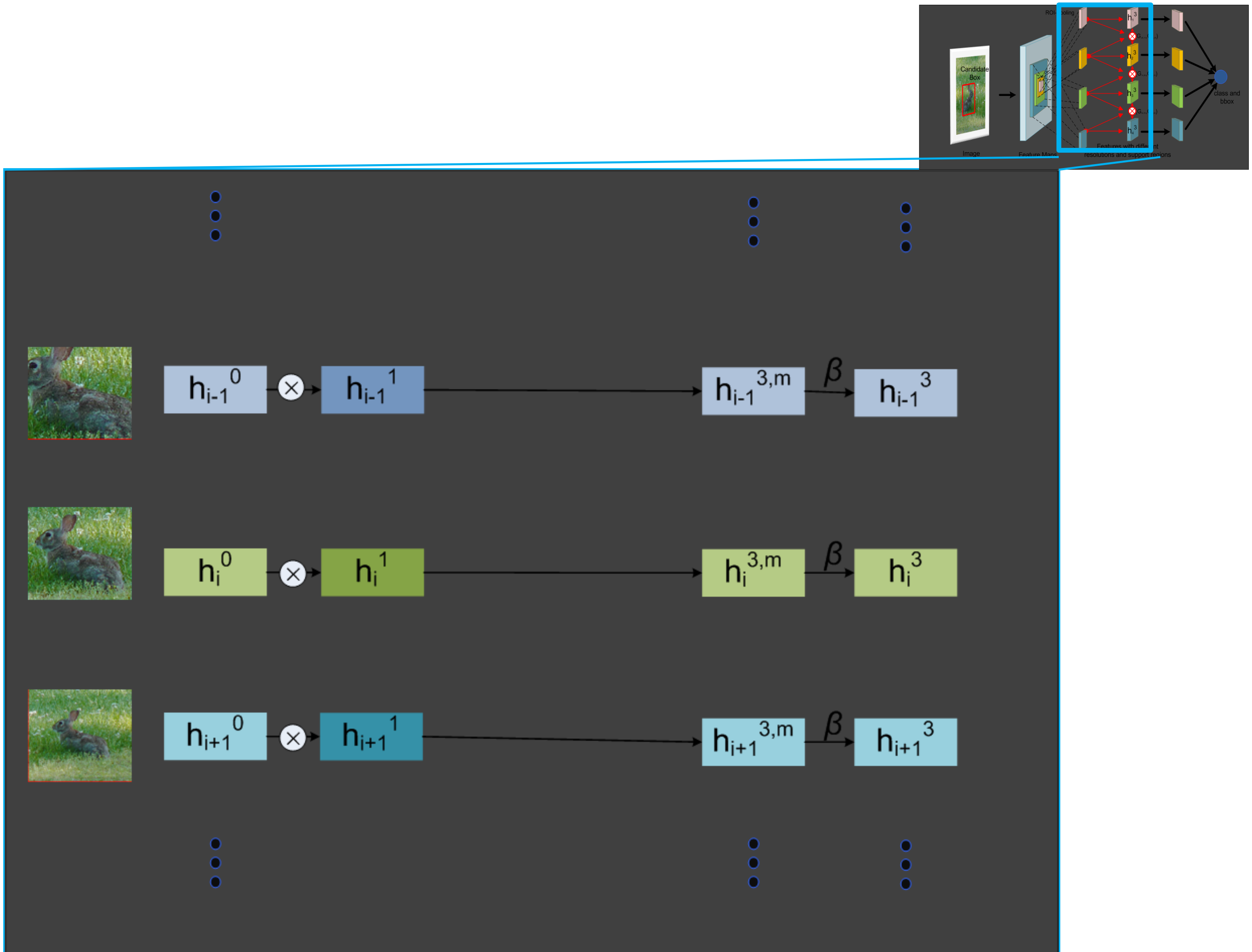


Gated bi-directional CNN (GBD-Net)

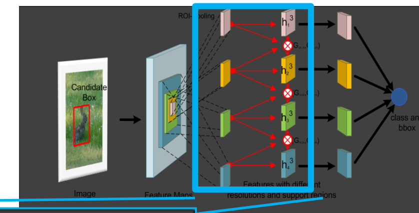
- Passing messages among these features

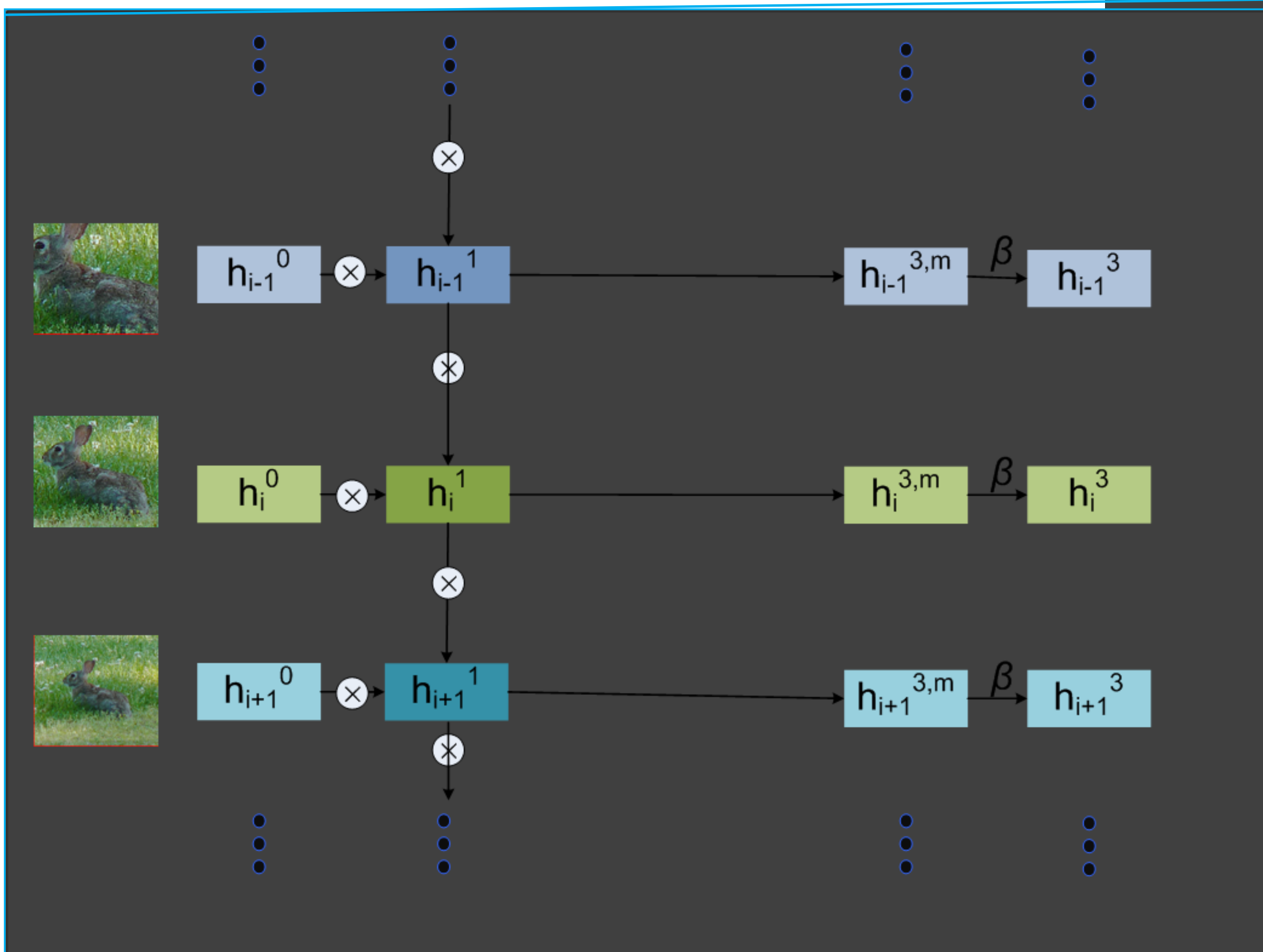


Independent features

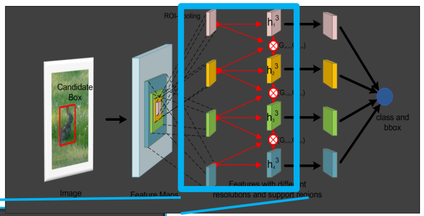


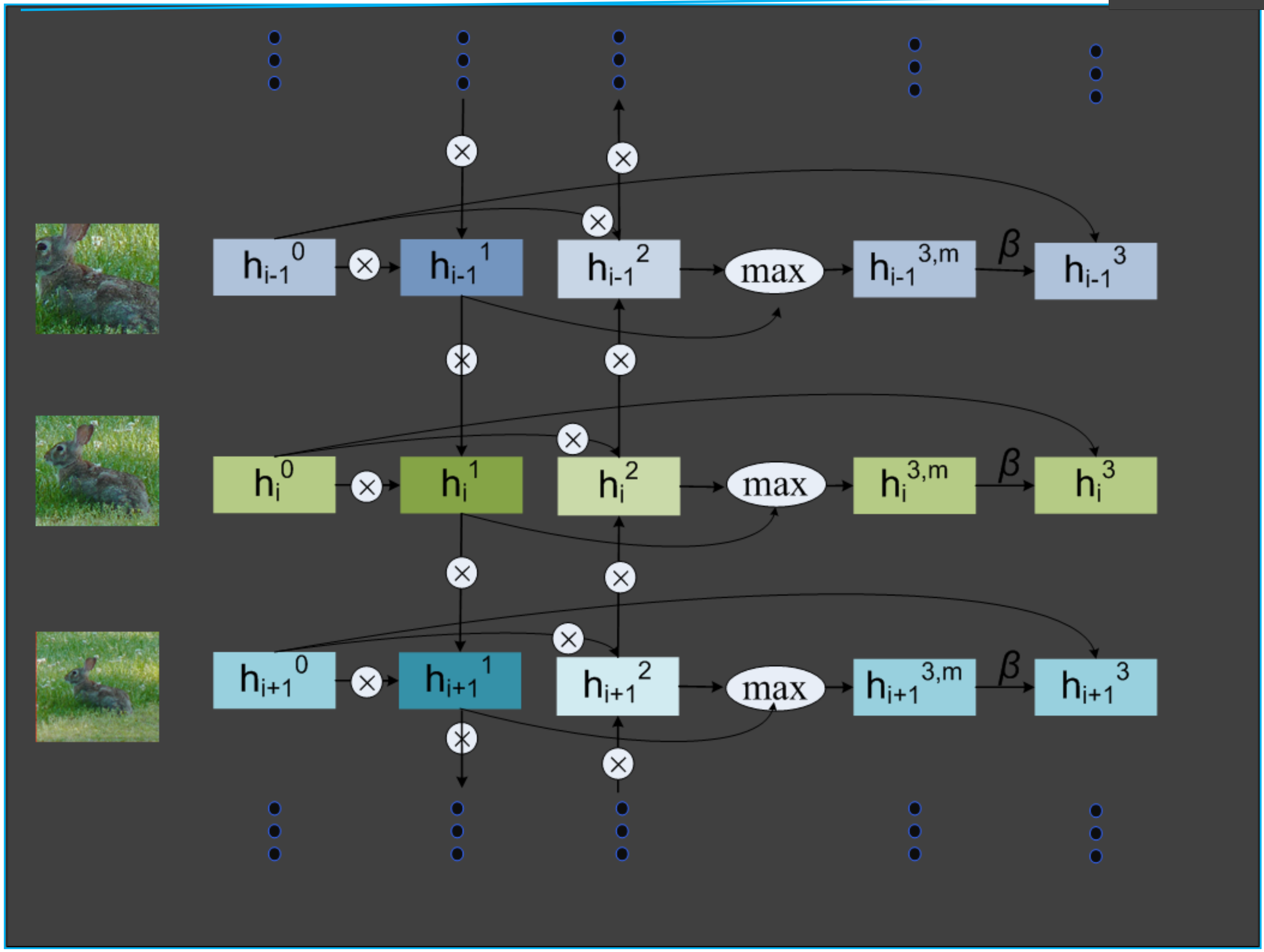
Passing message in one direction





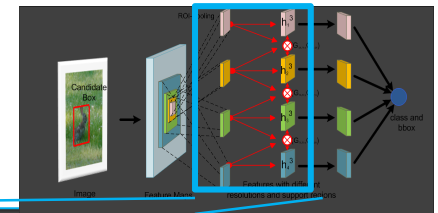
Passing message in two directions

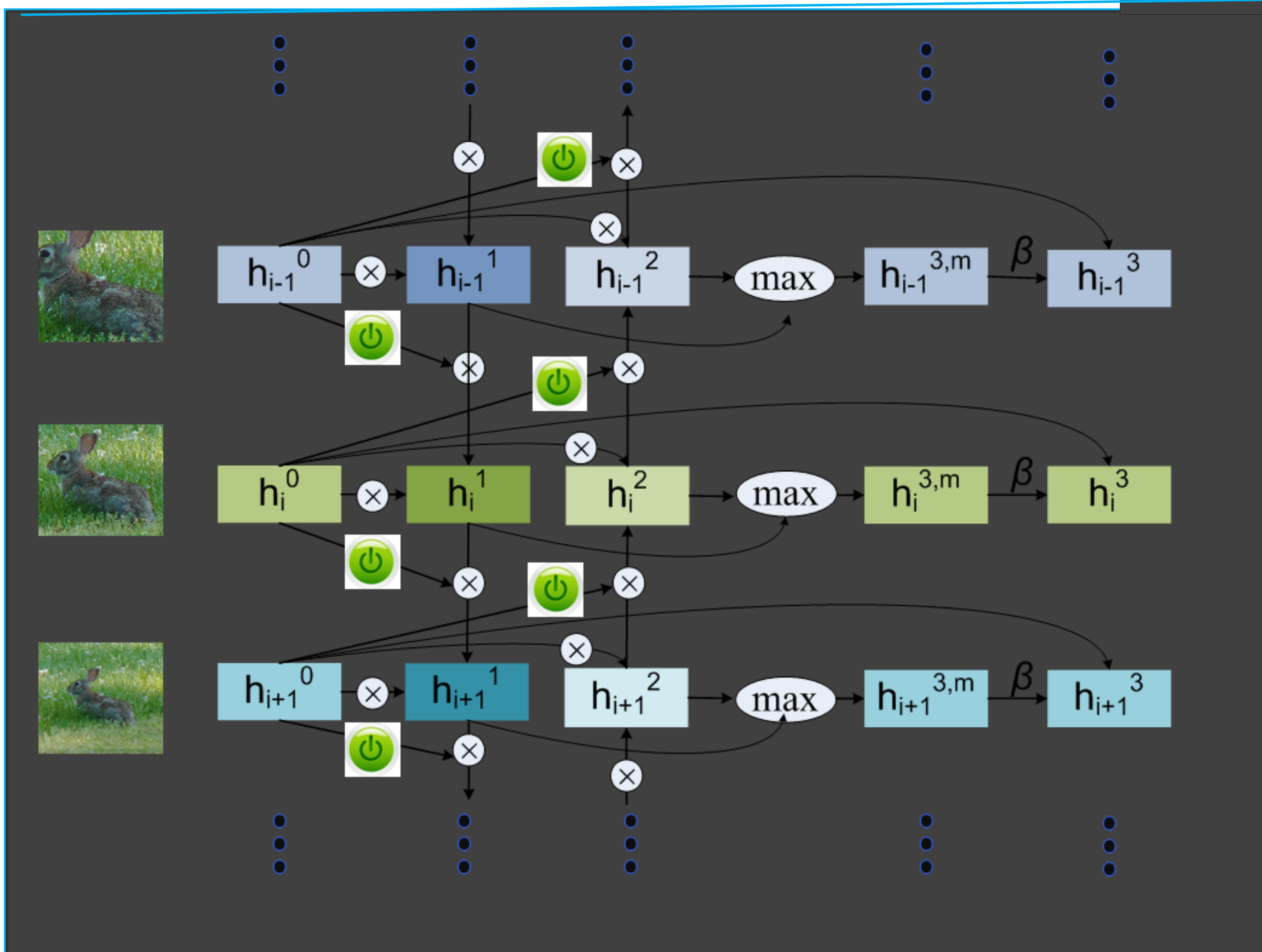




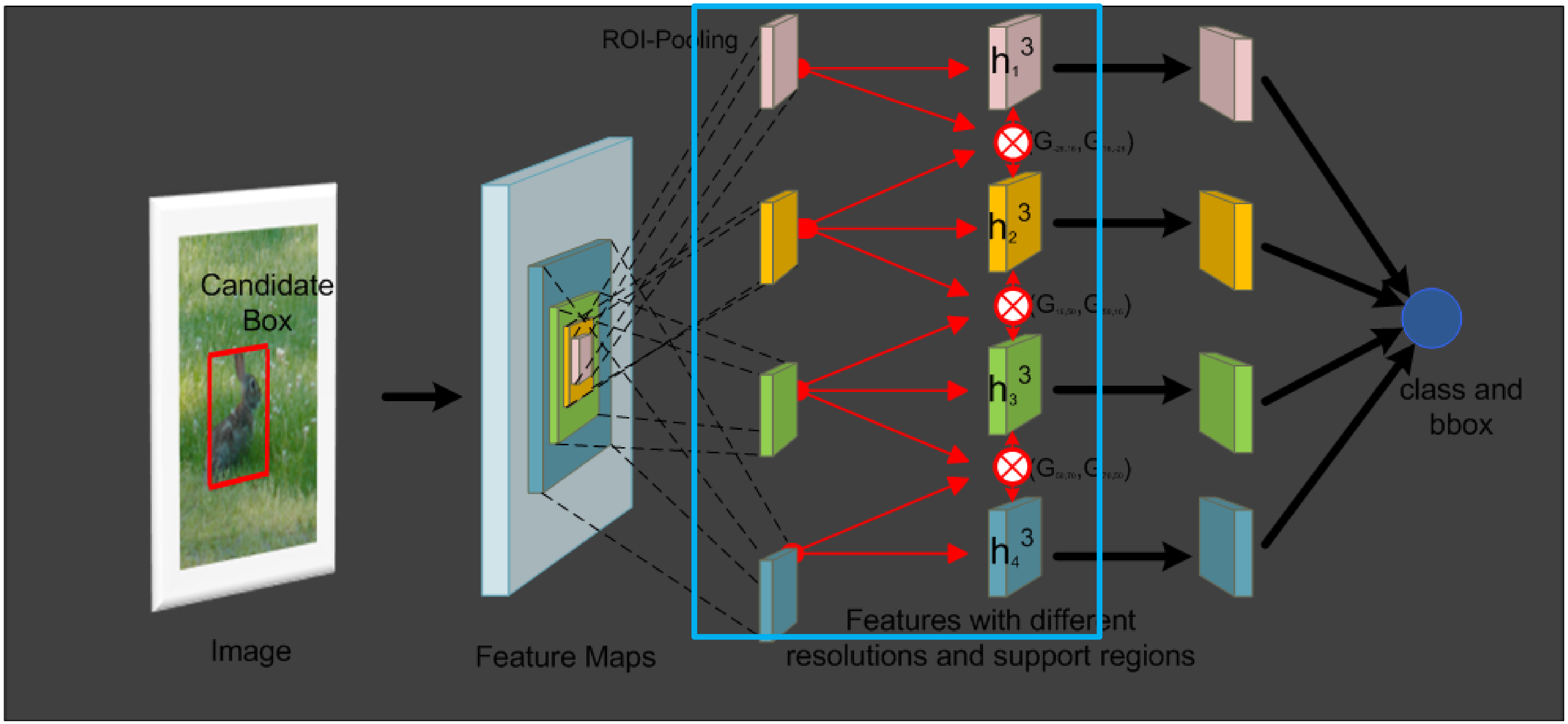
Passing message with gates

- +3.7% mAP on BN-Inception





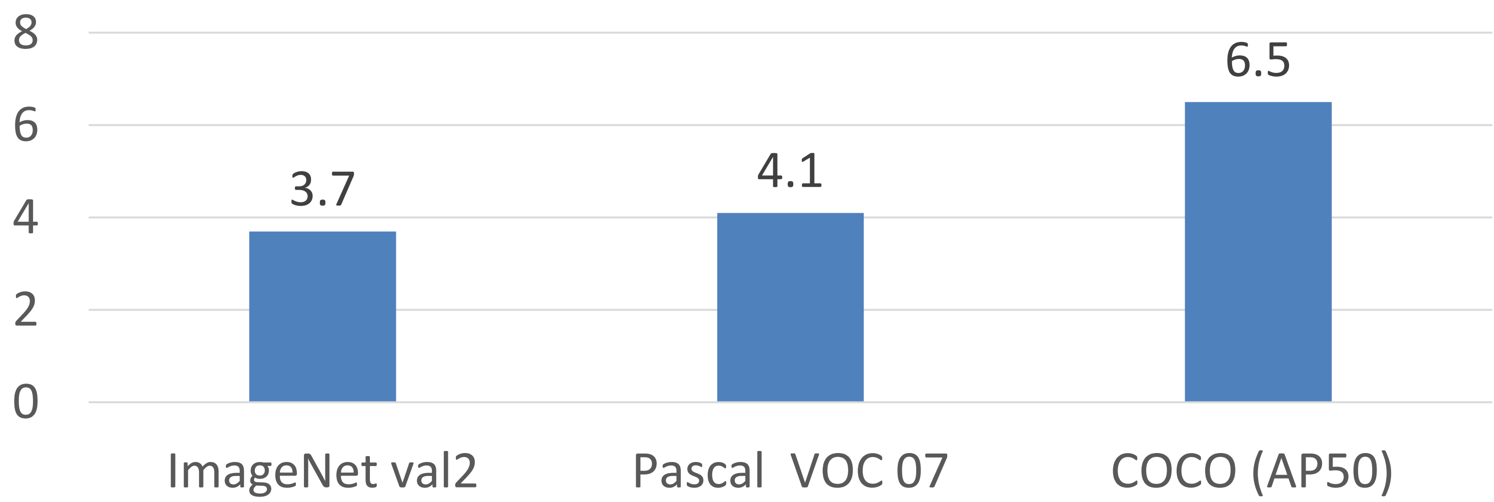
Gated bi-directional CNN (GBD-Net)



Improvement from GBD-net

BN-net (BN-Inception) as the baseline

DataSet	ImageNet val2	Pascal VOC 07	COCO (AP^{50})
Without GBD	48.4	**73.1**	**39.3**
+ GBD	52.1	**77.2**	**45.8**



Brief summary

- Features matter
- Observations from vision researchers also matter
- Use deep model as a tool to model the relationship among features
- Gated bi-directional network (GBD-Net)
  - Pass messages among features from different contextual regions

A pretrained deep model with 269 layers is also provided

Code: https://github.com/craftGBD/craftGBD

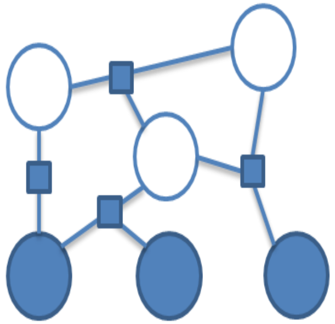
Zeng et al. "Crafting GBD-Net for Object Detection," TPAMI, accepted.

Motivation

- Debate
  – Lack of "general theory"
- Solution
  – Probabilistic model, conditional random field, is used as the theory

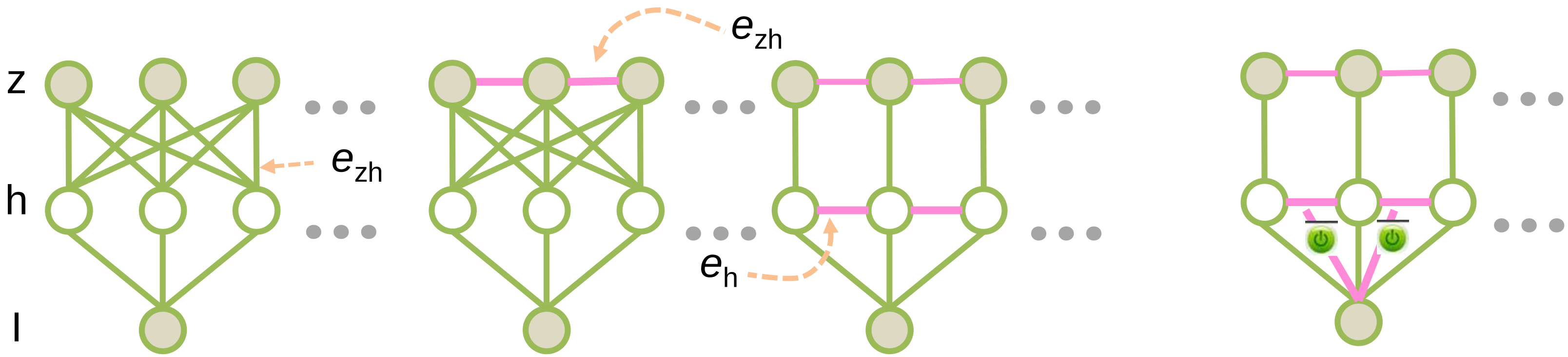
Conditional Random Field

$$p(\mathbf{z}, \boldsymbol{h} | \mathbf{I}, \boldsymbol{\theta}) = \sum_{\mathbf{h}} p(\mathbf{z}, \mathbf{h} | \mathrm{I}, \boldsymbol{\theta})$$



Where,

$$p(\mathbf{z}, \mathbf{h} | \mathrm{I}, \boldsymbol{\theta}) = \frac{e^{-E(\mathbf{z}, \mathbf{h}, \mathrm{I}, \boldsymbol{\theta})}}{\sum_{\mathbf{z}, \mathbf{h}} e^{-E(\mathbf{z}, \mathbf{h}, \mathrm{I}, \boldsymbol{\theta})}}$$



(a) Multi-layer neural network

(b) Structured output space

(c) Structured hidden layer

(d) Attention gated Structured hidden layer

Model (a) $E(\mathbf{z},\mathbf{h},\mathbf{I},\boldsymbol{\theta}) = \sum_k \Phi_h(h_k,\mathrm{I}) + \sum_{(i,k)\in\varepsilon_{zh}} \varphi_{zh}(\mathbf{z}_i, hk)$

Model (b) $E(\mathbf{z},\mathbf{h},\mathbf{I},\boldsymbol{\theta}) = \sum_k \Phi_h(h_k,\mathrm{I}) + \sum_{(i,k)\in\varepsilon_{zh}} \varphi_z(\mathbf{z}_i, h_k) + \sum_{(i,j)\in\varepsilon_z} \varphi_z(\mathbf{z}_i,\mathbf{z}_j)$

Model (c) $E(\mathbf{z},\mathbf{h},\mathbf{I},\boldsymbol{\theta}) = \sum_k \Phi_h(h_k,\mathrm{I}) + \sum_i \varphi_{zh}(\mathbf{z}_i, hi) + \sum_{(i,j)\in\varepsilon_z} \varphi_z(\mathbf{z}_i,\mathbf{z}_j) + \sum_{(k,l)\in\varepsilon_h} \varphi_h(h_k,h_l)$

Model (d) $E(\mathbf{z},\mathbf{h},\mathbf{I},\boldsymbol{\theta}) = \sum_k \Phi_h(h_k,\mathrm{I}) + \sum_i \varphi_{zh}(\mathbf{z}_i, hi) + \sum_{(i,j)\in\varepsilon_z} \varphi_z(\mathbf{z}_i,\mathbf{z}_j) + \sum_{(k,l)\in\varepsilon_h} g_k\varphi_h(h_k,h_l)$

"End-to-End Learning of Deformable Mixture of Parts and Deep Convolutional Neural Networks for Human Pose Estimation", *CVPR* 2016.

"Structured feature learning for pose estimation", CVPR 2016.

"CRF-CNN: Modeling Structured Information in Human Pose Estimation", NIPS, 2016.

Learning Deep Structured Multi-Scale Features using Attention-Gated CRFs for Contour Prediction", NIPS, 2017.

To obtain the estimation of features:

$$p(\boldsymbol{h}|\mathrm{I},\boldsymbol{\theta}) = \prod_i Q(\mathbf{h}_i|\mathrm{I},\boldsymbol{\theta})$$

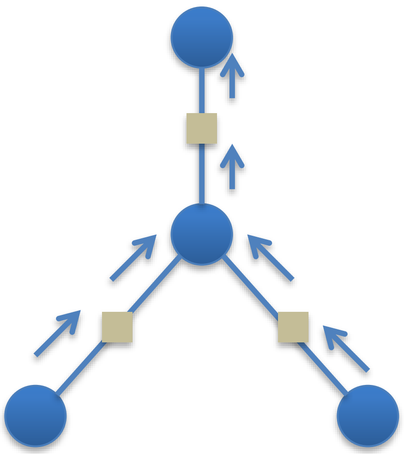
Mean Field Approximation

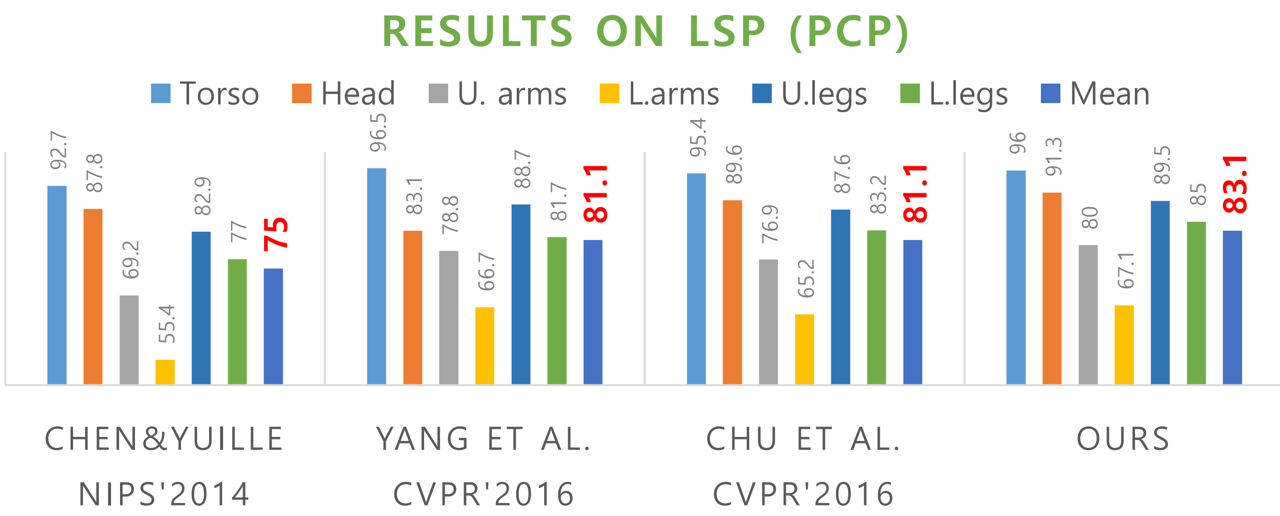
$$Q(\mathbf{h}_i|\mathbf{I},\Theta) = \frac{1}{Z_{h,i}} exp\left\{-\sum_{h_k\in\mathbf{h}_i} \emptyset_h(h_k,\mathbf{I}) - \sum_{\substack{(i,j)\in\boldsymbol{\varepsilon}_h \\ i<j}} \varphi_h(\mathbf{h}_i, Q(\mathbf{h}_j|\mathbf{I},\Theta)\right\}$$

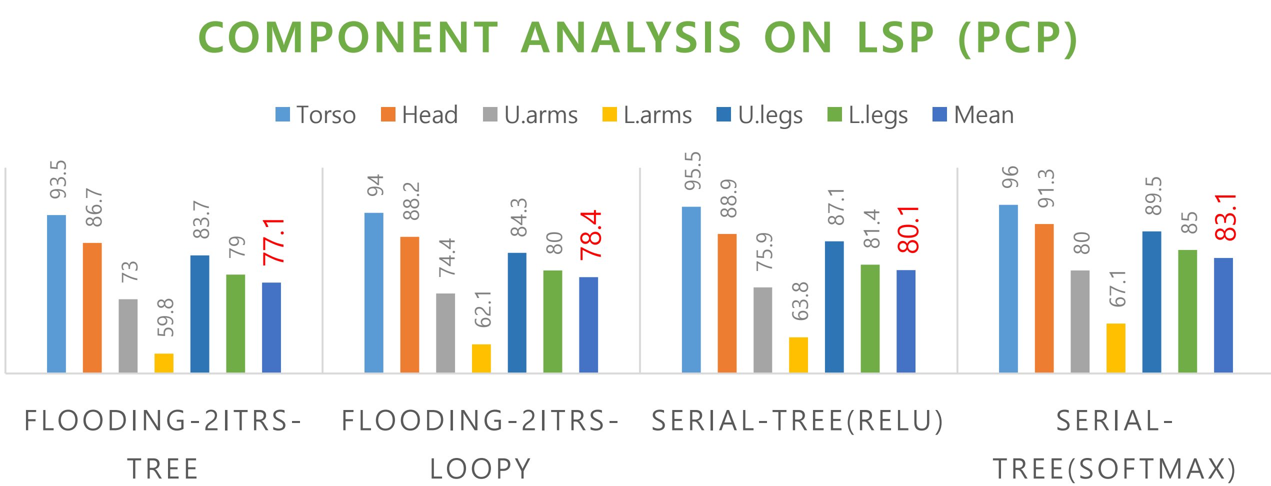
$$Q(\mathbf{h}_i|\mathrm{I},\boldsymbol{\theta}) = \frac{1}{z_{h,i}} e^{\left\{-\sum_{h_k} \Phi_{h(h_k,\mathrm{I})} - \sum_{(i,j)\in\varepsilon_h} \varphi_h(\mathbf{h}_i, Q(hj|\mathrm{I},\boldsymbol{\theta})\right\}}$$

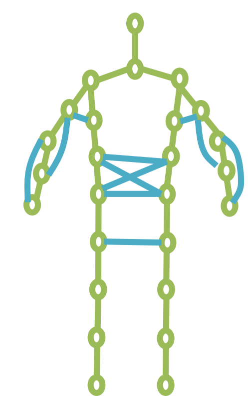
Message passing

- Belief propagation
  - N^2 => $2N$









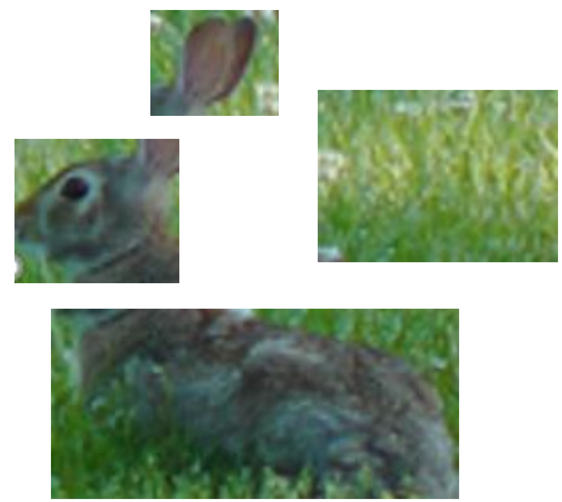
Loopy Structure

Why structured features?

- Richer visual information

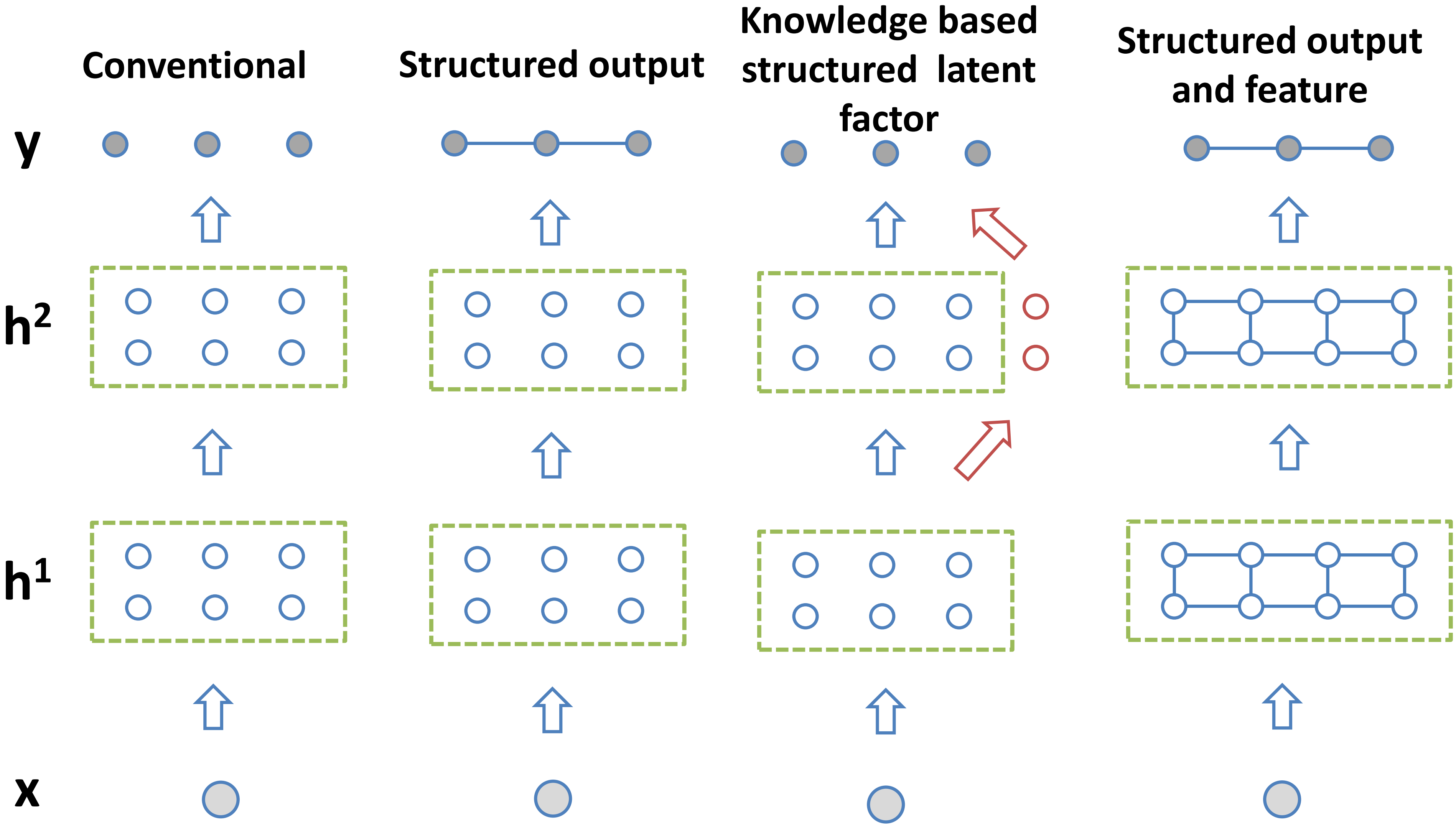
Label: rabbit

Visual feature



Facing left

Model structures among neurons



Is structured learning only effective for object detection?

Application of structured feature learning

- Haze removal (Submitted to CVPR19)
- Depth estimation (TPAMI 18)
- Contour estimation (NIPS 17)
- Detection (TPAMI17, TPAMI18, …)
- Human pose estimation (CVPR16)
- Person re-identification (CVPR18)
- Relationship estimation (ICCV17)
- Image captioning (ICCV17)

Low-level vision

High-level vision

Vision + Language

D. Xu, *et al.*, "Monocular Depth Estimation using Multi-Scale Continuous CRFs as Sequential Deep Networks," *TPAMI* 2018.
W. Ouyang, *et al.*, " Jointly learning deep features, deformable parts, occlusion and classification for pedestrian detection," *TPAMI* 2018.
W. Ouyang, *et. al*. "DeepID-Net: Object Detection with Deformable Part Based Convolutional Neural Networks", *TPAMI* 2017.
X. Chu, **W. Ouyang**, *et. al*. "Structured feature learning for pose estimation". *CVPR* 2016.
Y. Li, **W. Ouyang**, *et. al*. "Scene Graph Generation from Objects, Phrases and Region Captions", *ICCV*, 2017.

Is structured learning only effective for specific vision task?

Outline

Introduction

Structured deep learning

Back-bone model design

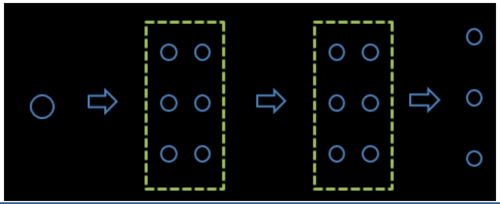
Conclusion

Outline

Introduction

Structured deep learning

Back-bone model design



Conclusion

Back-bone deep model design

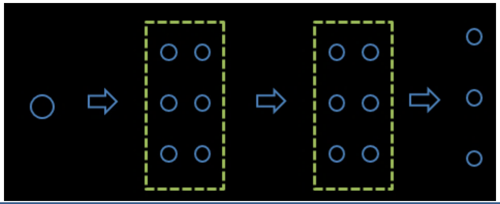
- Basis structure of deep model
  - AlexNet, VGG, GoogleNet, ResNet, DenseNet
  - Validated on large-scale classification tasks such as ImageNet
  - Models pretrained on ImageNet are found to be effective initial model for other tasks

Outline

Introduction

Structured deep learning

Back-bone model design



FishNet (NeurIPS18)

Optical flow guided feature (CVPR18)

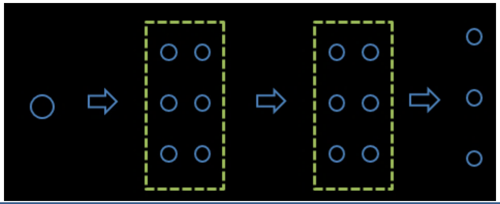
Conclusion

Outline

Introduction

Structured deep learning

Back-bone model design

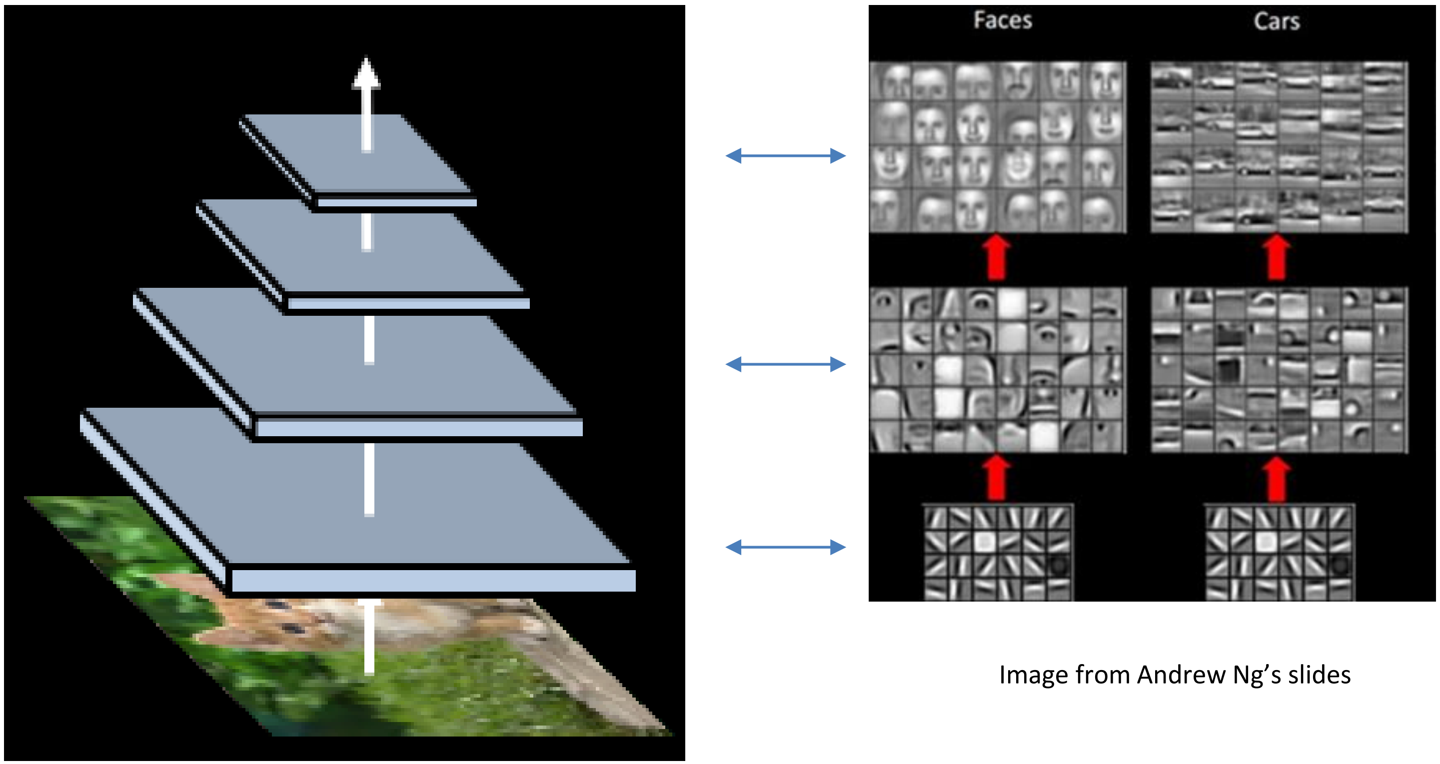


FishNet (NeurIPS18)

Optical flow guided feature (CVPR18)

Conclusion

Low-level and high-level features



High-level

Low-level

Image from Andrew Ng's slides

Current CNN Structures

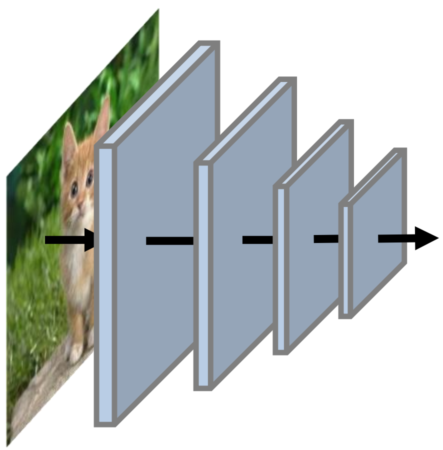
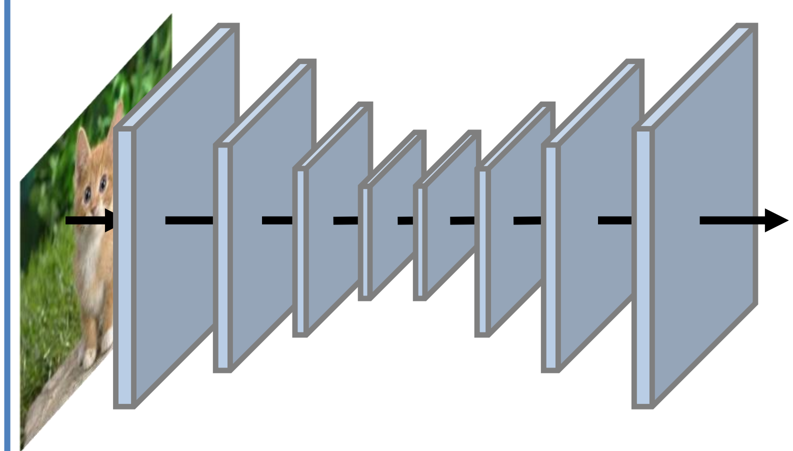


Image Classification:
Summarize high-level semantic information of the whole image.

Called U-Net, Hourglass, or Conv-deconv



Detection/Segmentation:
High-level semantic meaning with high spatial resolution

Architectures designed for tasks of different granularities are

DIVERGING

Unify the advantages of networks for pixel-level, region-level, and image-level tasks

Observation and design

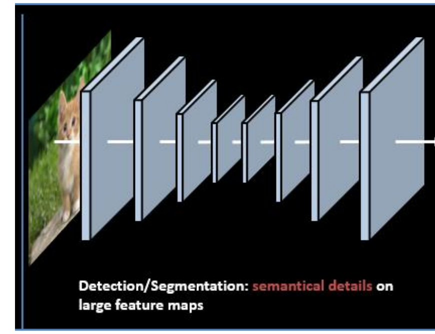
- Our observation

1. Diverged structures for tasks requiring different resolutions.

- Design

1. Unify the advantages of networks for pixel-level, region-level, and image-level tasks.

Hourglass for Classification



Features with high-level semantics and high resolution is good

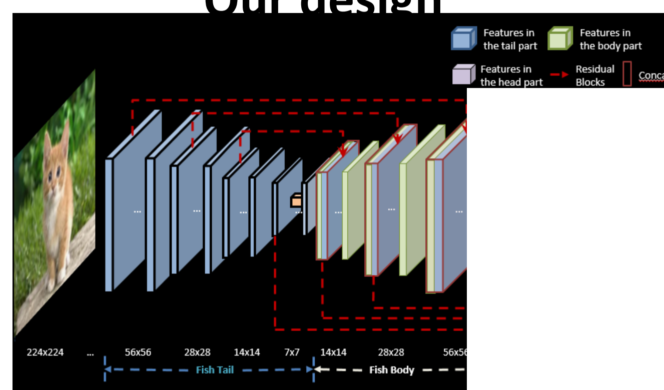
Directly applying hourglass for classification?

Poor performance.

So what is the **problem**?

- **Different tasks require different resolutions of feature**
- **Down sample high-level features with high resolution**

Our design



Observation and design

- Observation

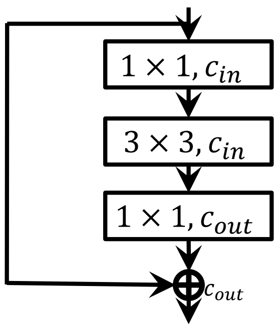
1. Diverged structures for tasks requiring different resolutions.
2. Isolated Conv blocks the direct back-propagation

- Design

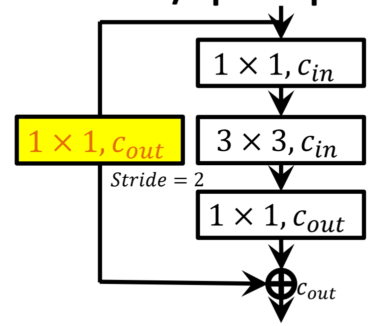
1. Unify the advantages of networks for pixel-level, region-level, and image-level tasks.

Hourglass for Classification

Normal Res-Block



Res-Block for down/up sampling



The 1×1 convolution layer in yellow indicates the Isolated convolution.

- **Hourglass may bring more isolated convolutions than ResNet**

Observation and design

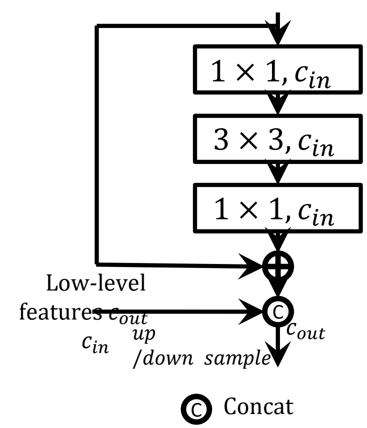
- Observation

1. Diverged structures for tasks requiring different resolutions.
2. Isolated Conv blocks the direct back-propagation

- Design

1. Unify the advantages of networks for pixel-level, region-level, and image-level tasks.
2. Design a network that does not need isolated convolution

Our design



Observation and design

- Observation

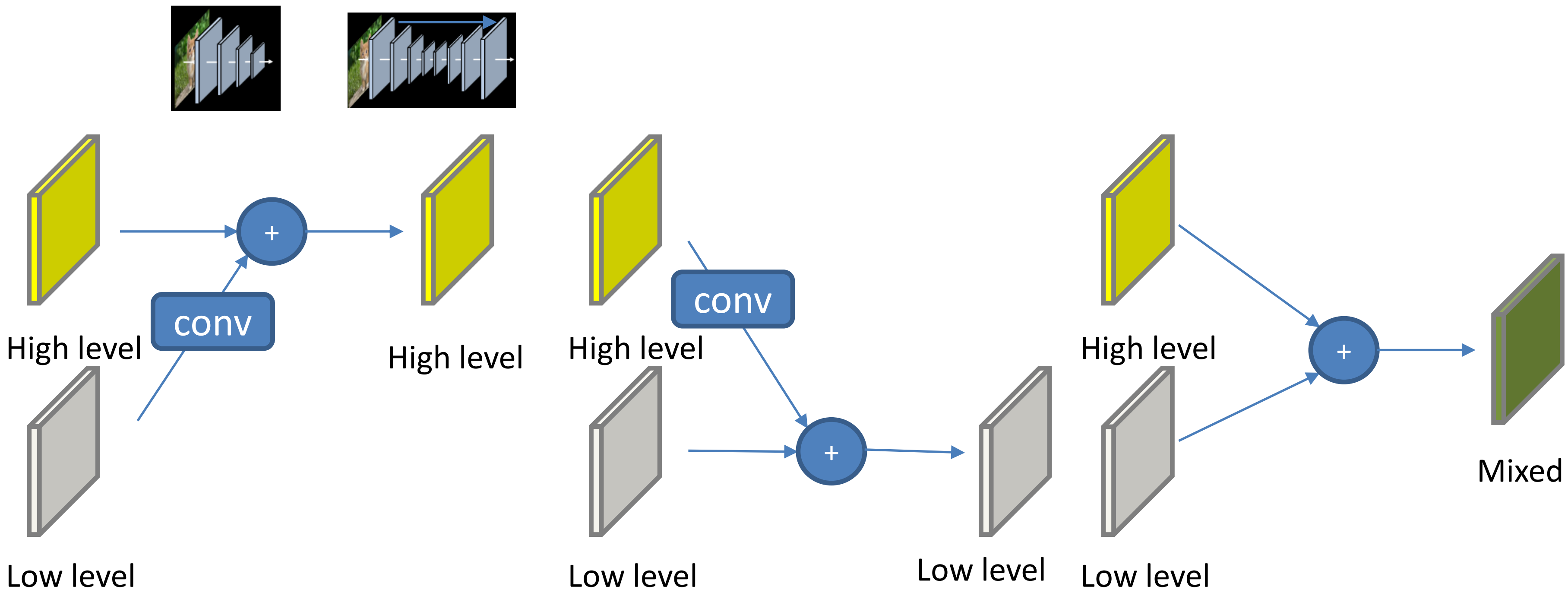
1. Diverged structures for tasks requiring different resolutions.
2. Isolated Conv blocks the direct back-propagation
3. Features with different depths are not fully explored, or mixed but not preserved

- Design

1. Unify the advantages of networks for pixel-level, region-level, and image-level tasks.
2. Design a network that does not need isolated convolution
3. Features from varying depths are **preserved and refined** from each other.

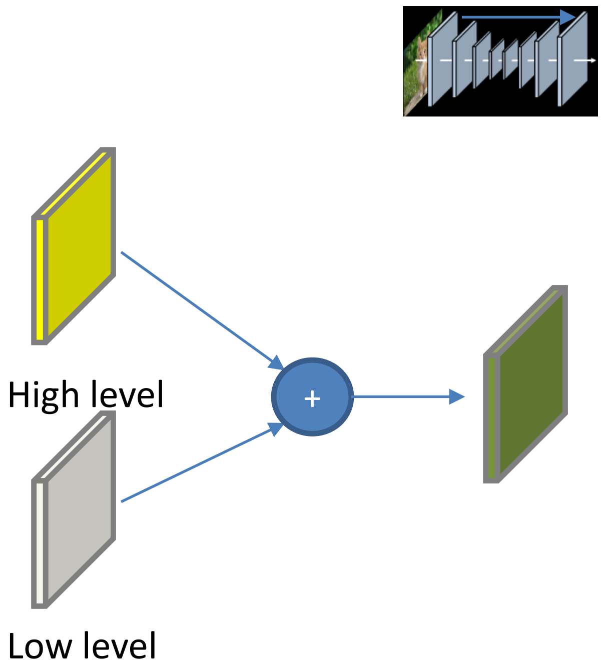
Bharath Hariharan, et al. "Hypercolumns for object segmentation and fine-grained localization." *CVPR'15.*
Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." *ECCV'*16.

Difference between mix and preserve and refine

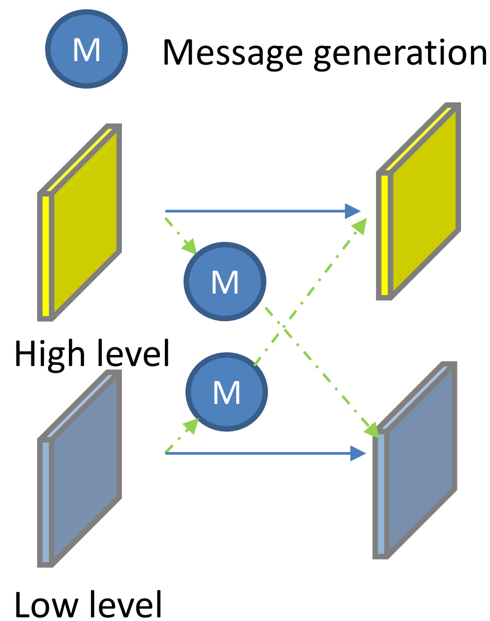


Mixed features

Difference between mix and preserve and refine



Mixed features



Preserve and refine

Observation and design

Solution

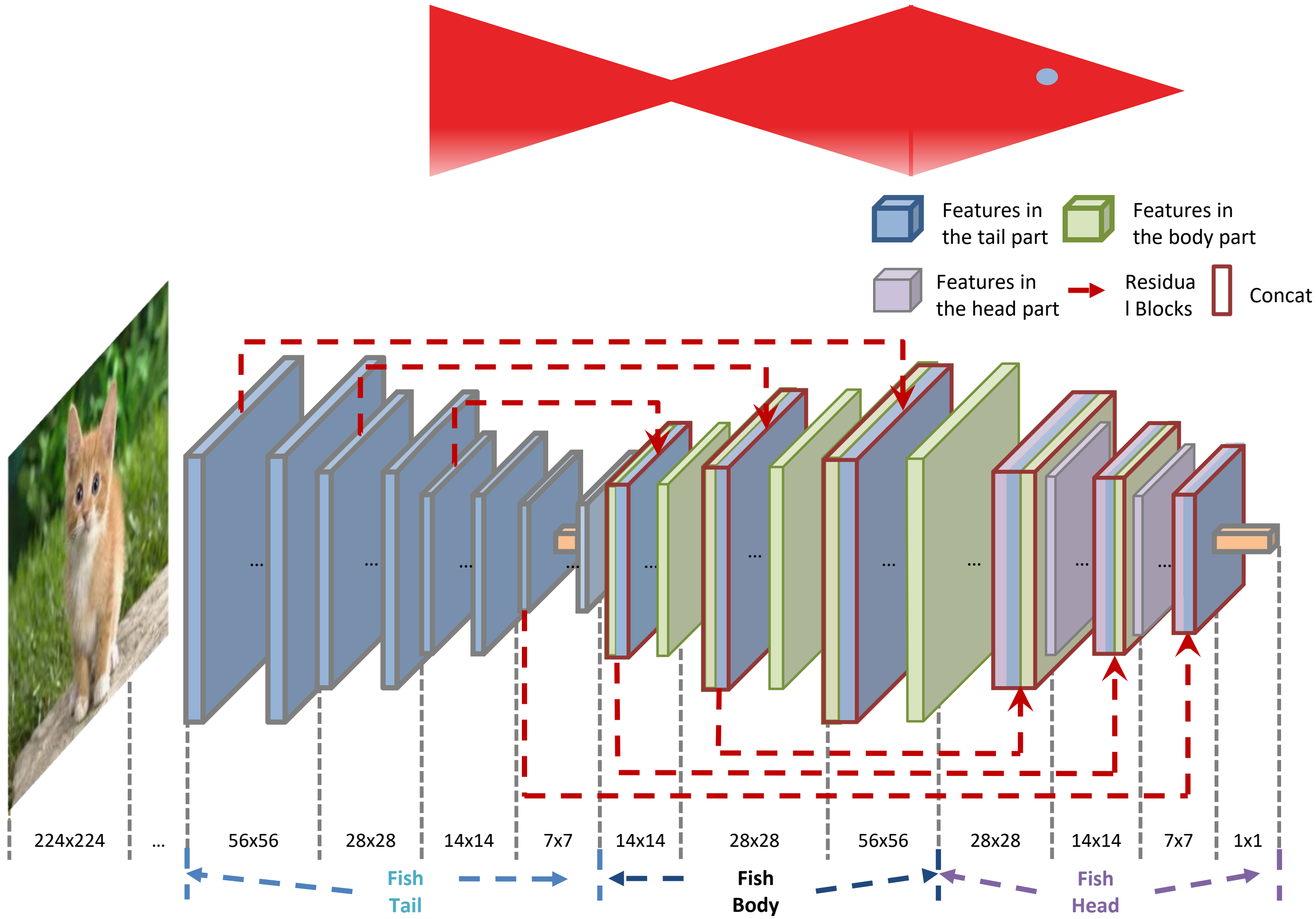
1. Diverged structures for tasks requiring different resolutions.
2. Isolated Conv blocks the direct back-propagation
3. Features with different depths are not fully explored, or mixed but not preserved

Our observation

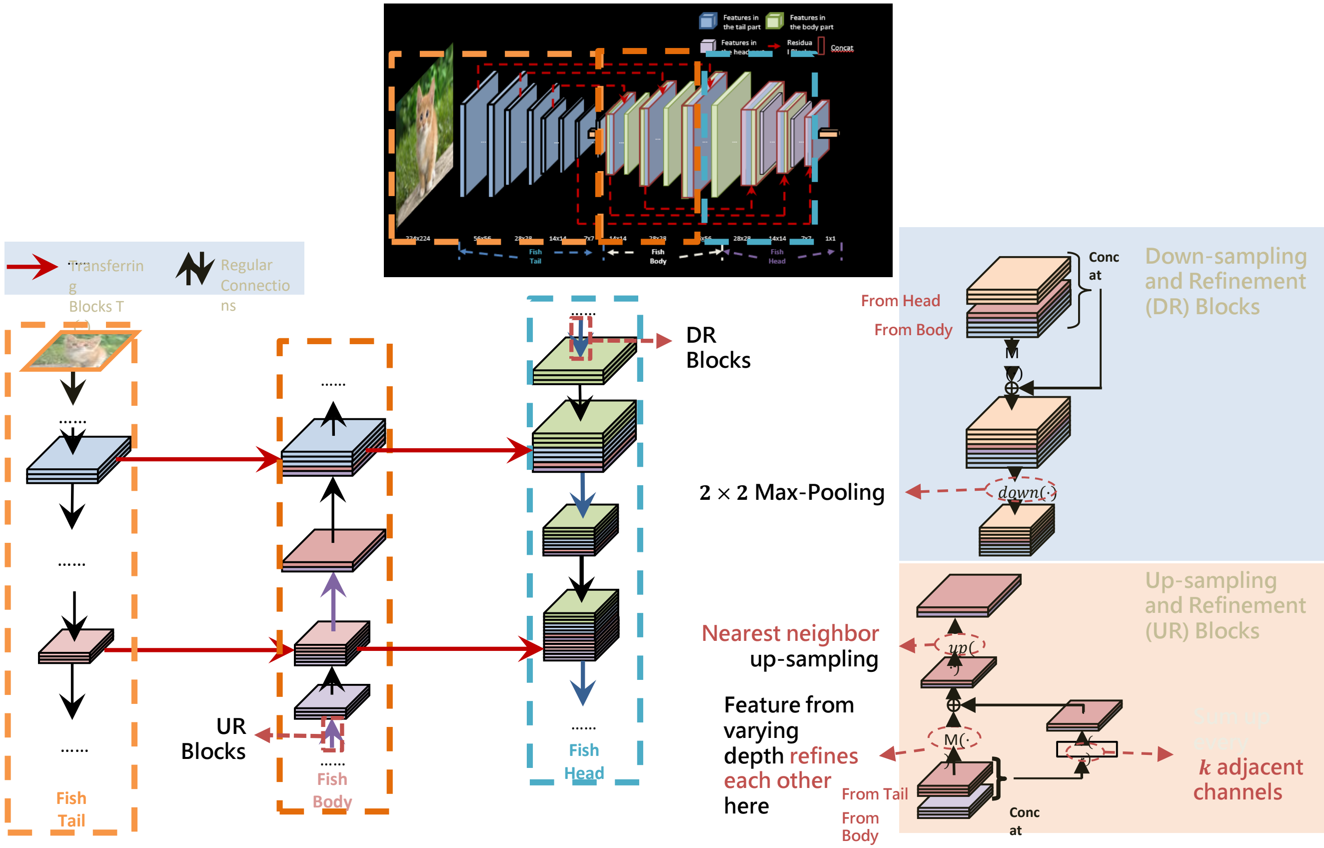
1. Unify the advantages of networks for pixel-level, region-level, and image-level tasks.
2. Design a network that does not need isolated convolution
3. Features from varying depths are **preserved and refined** from each other.

Bharath Hariharan, et al. "Hypercolumns for object segmentation and fine-grained localization." *CVPR'15.*
Newell, Alejandro, Kaiyu Yang, and Jia Deng. "Stacked hourglass networks for human pose estimation." *ECCV'*16.

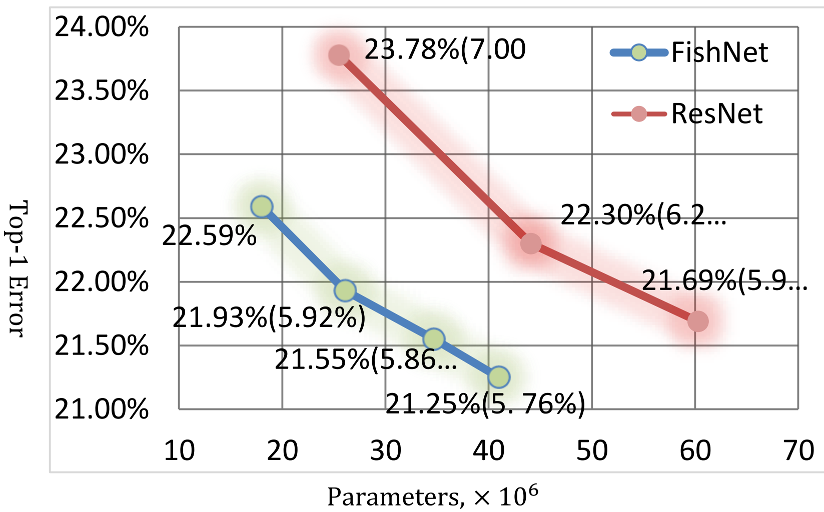
FishNet: Overview

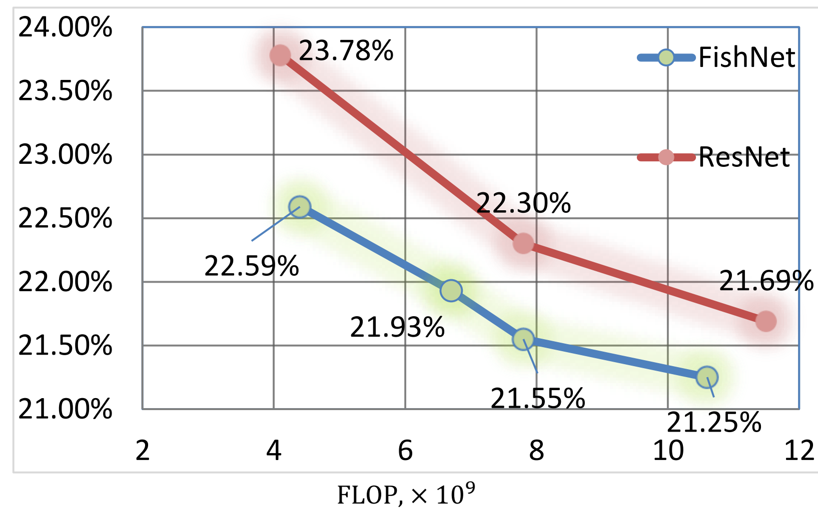


FishNet: Preservation & Refinement



FishNet: Performance-ImageNet

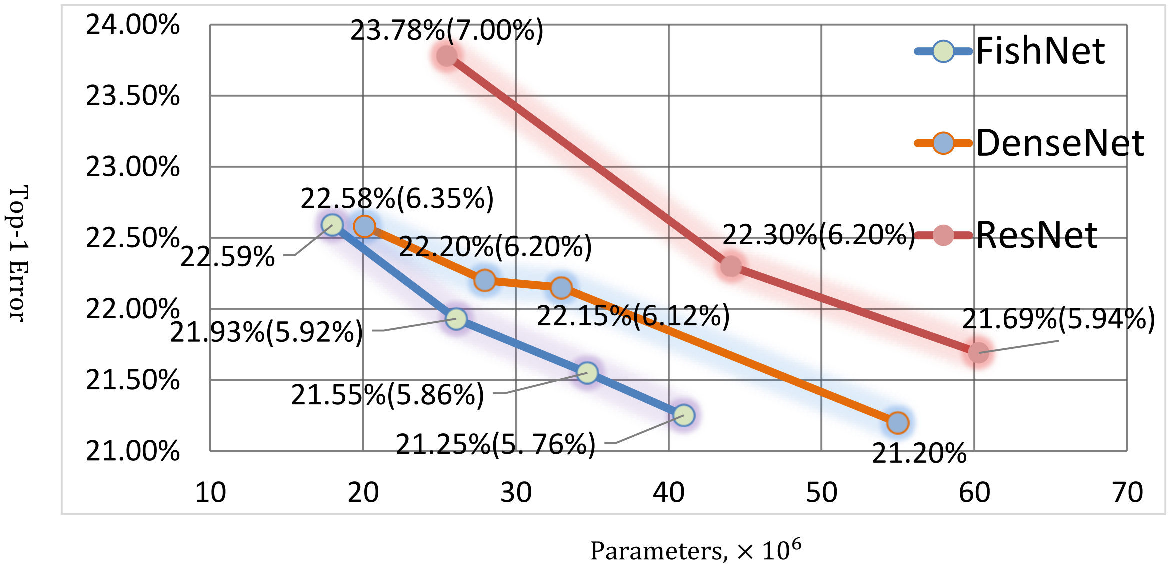




Code

https://github.com/kevin-ssy/FishNet

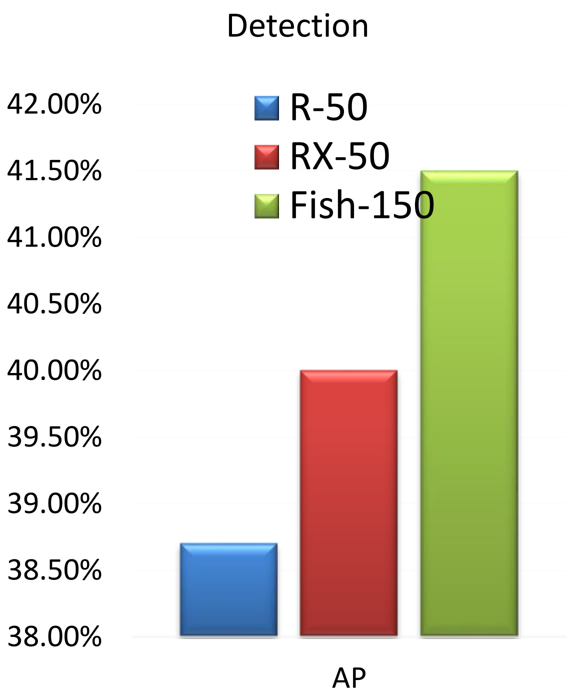
FishNet: Performance-ImageNet

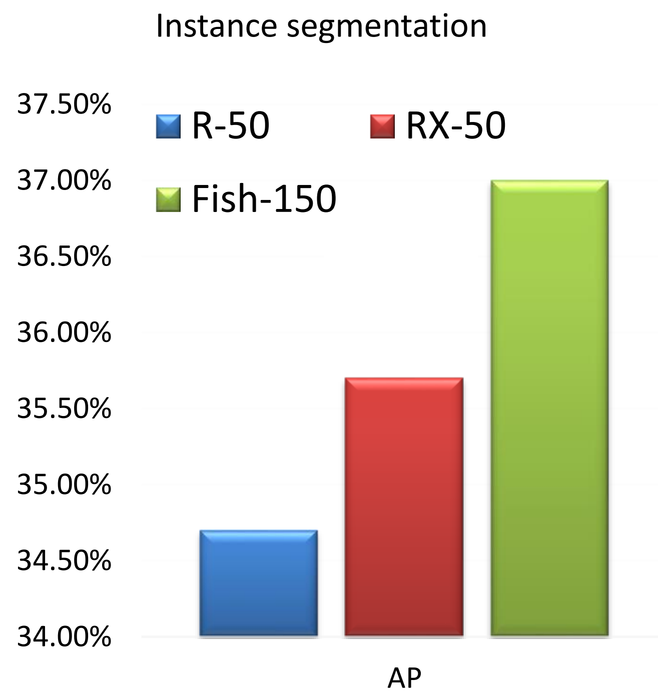


Code

https://github.com/kevin-ssy/FishNet

FishNet: Performance on COCO Detection and Segmentation





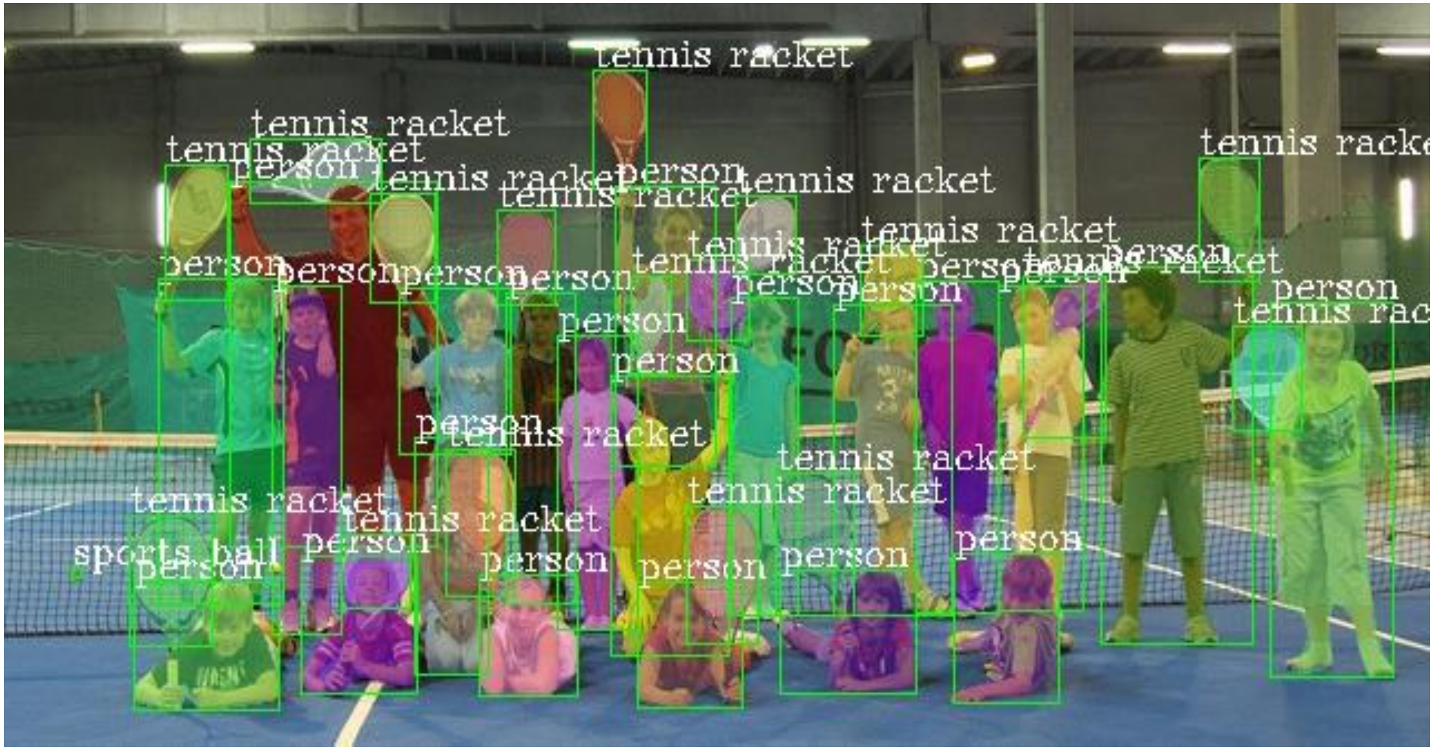
Code

https://github.com/kevin-ssy/FishNet

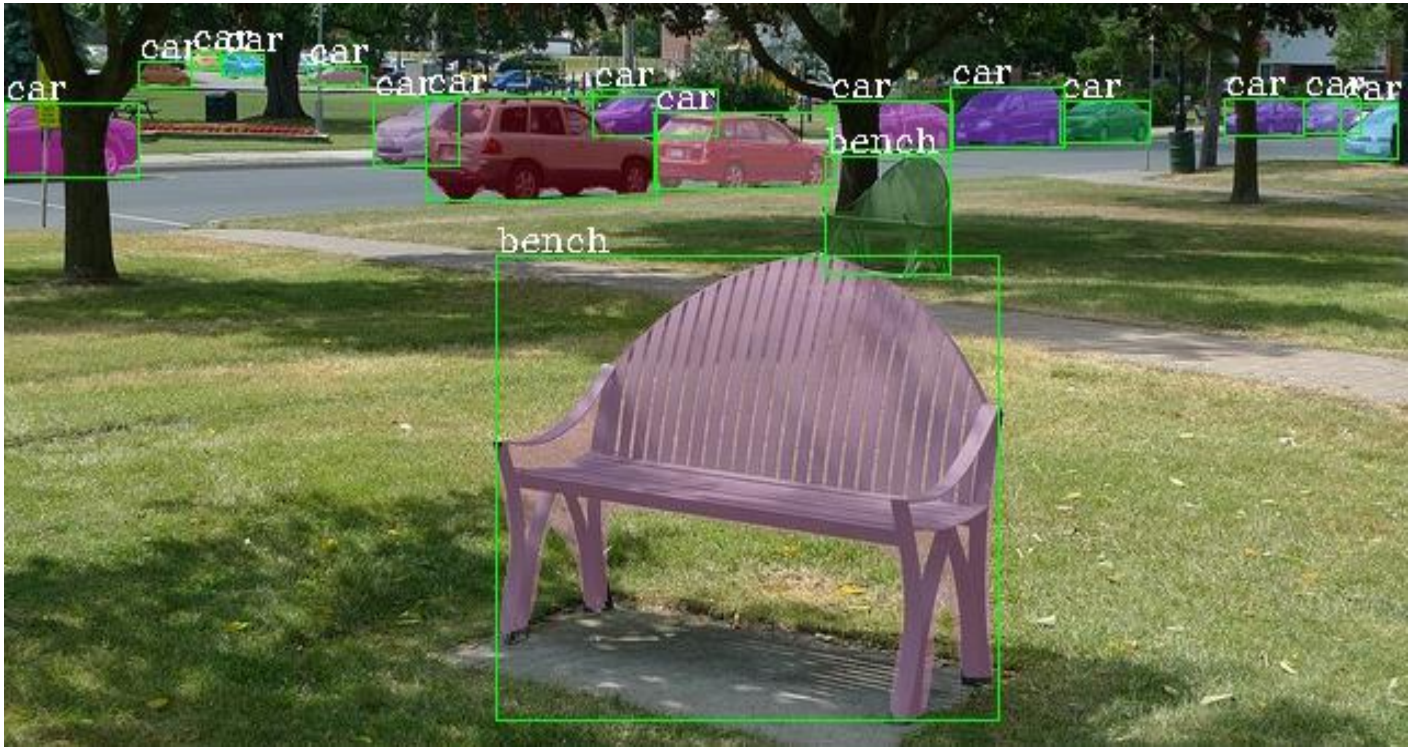
Winning COCO 2018 Instance Segmentation Task

	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}	AR^{1}	AR^{10}	AR^{100}	AR^{S}	AR^{M}	AR^{L}	date
$\boxed{\text{MMDet}}$	0.486	0.730	0.530	0.339	0.520	0.602	0.368	0.593	0.632	0.464	0.665	0.777	2018-08-18
Megvii (Face++)	0.485	0.737	0.532	0.298	0.507	0.641	0.369	0.594	0.630	0.474	0.659	0.767	2018-08-18
FirstShot	0.463	0.681	0.508	0.258	0.483	0.636	0.359	0.580	0.622	0.445	0.655	0.776	2018-08-17

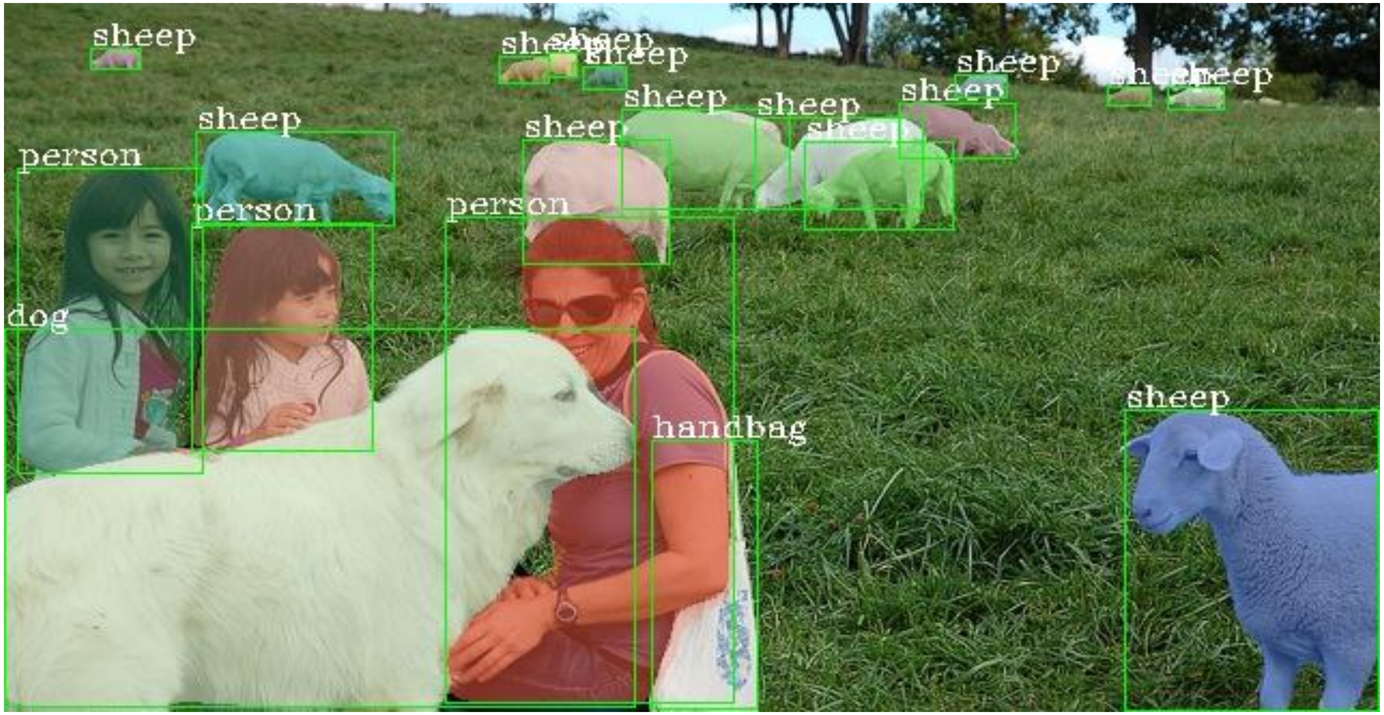
Visualization



Visualization

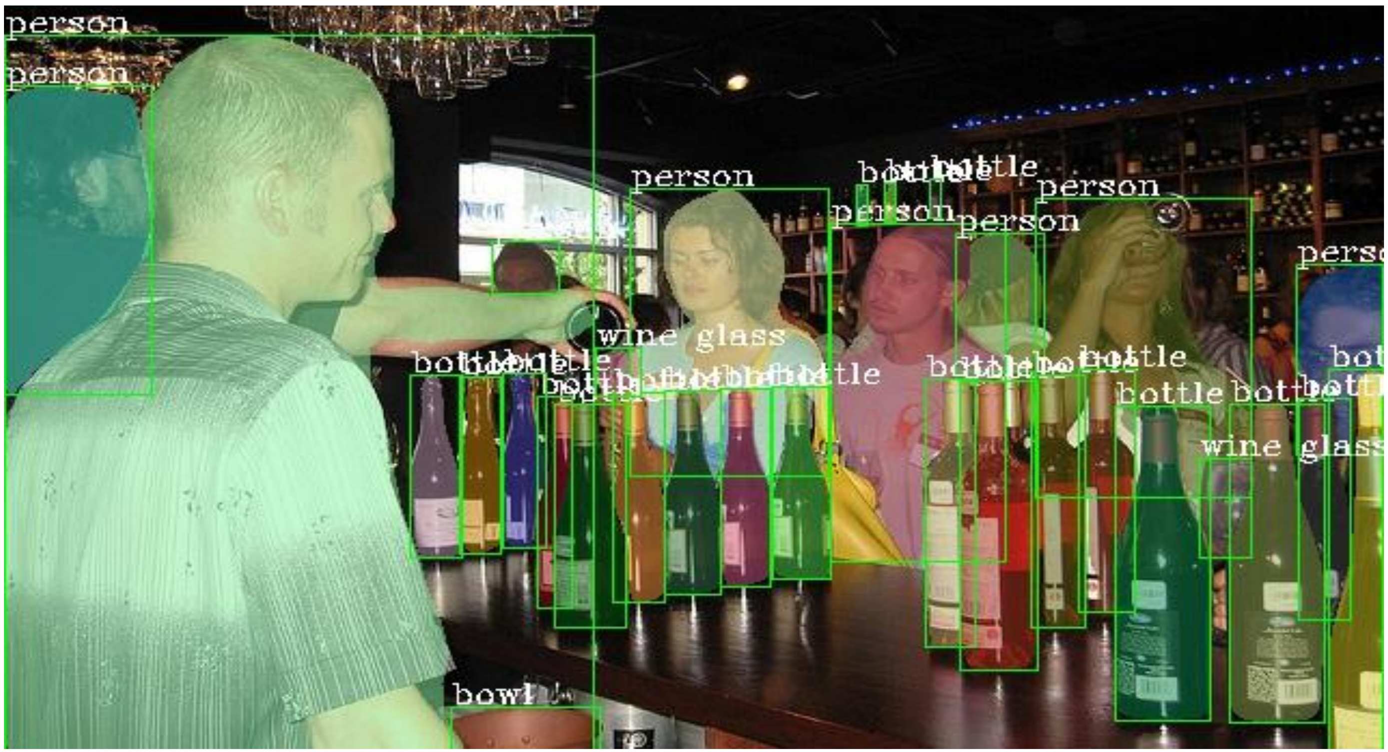


Visualization



Visualization

Visualization



Codebase

- **Comprehensive**
  - ☑ RPN
  - ☑ Mask R-CNN
  - ☑ Cascade R-CNN
  - ☐ More … …
  - ☑ Fast/Faster R-CNN
  - ☑ FPN
  - ☑ RetinaNet

- **High performance**
  - ☑ Better performance
  - ☑ Optimized memory consumption
  - ☑ Faster speed

- **Handy to develop**
  - ☑ Written with PyTorch
  - ☑ Modular design

GitHub: mmdet

FishNet: Advantages

1. Better gradient flow to shallow layers
2. Features
   - contain rich low-level and high-level semantics
   - are preserved and refined from each other

Code

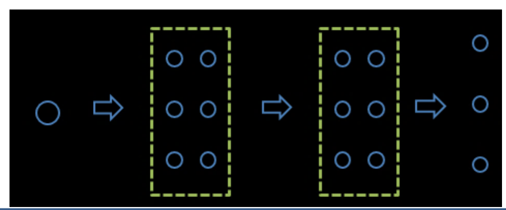
https://github.com/kevin-ssy/FishNet

Outline

Introduction

Structured deep learning

Back-bone model design



FishNet (NeurIPS18)

Optical flow guided feature (CVPR18)

Conclusion

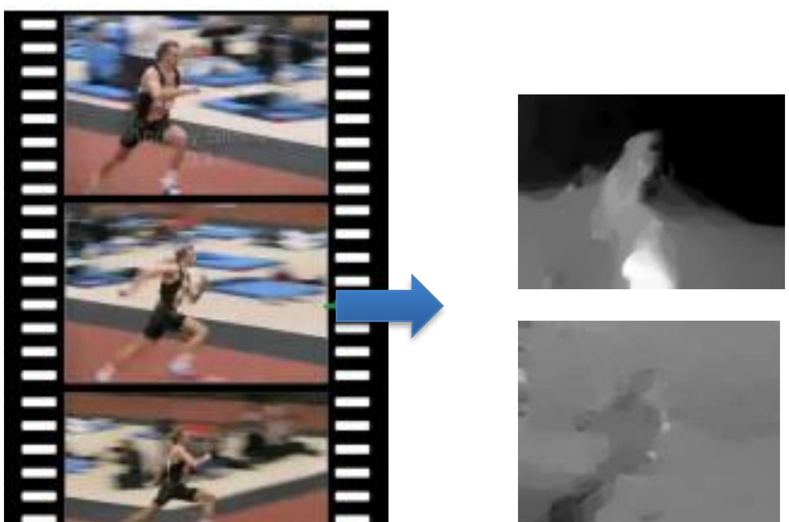
Action Recognition

- Recognize action from videos

Optical flow in Action Recognition

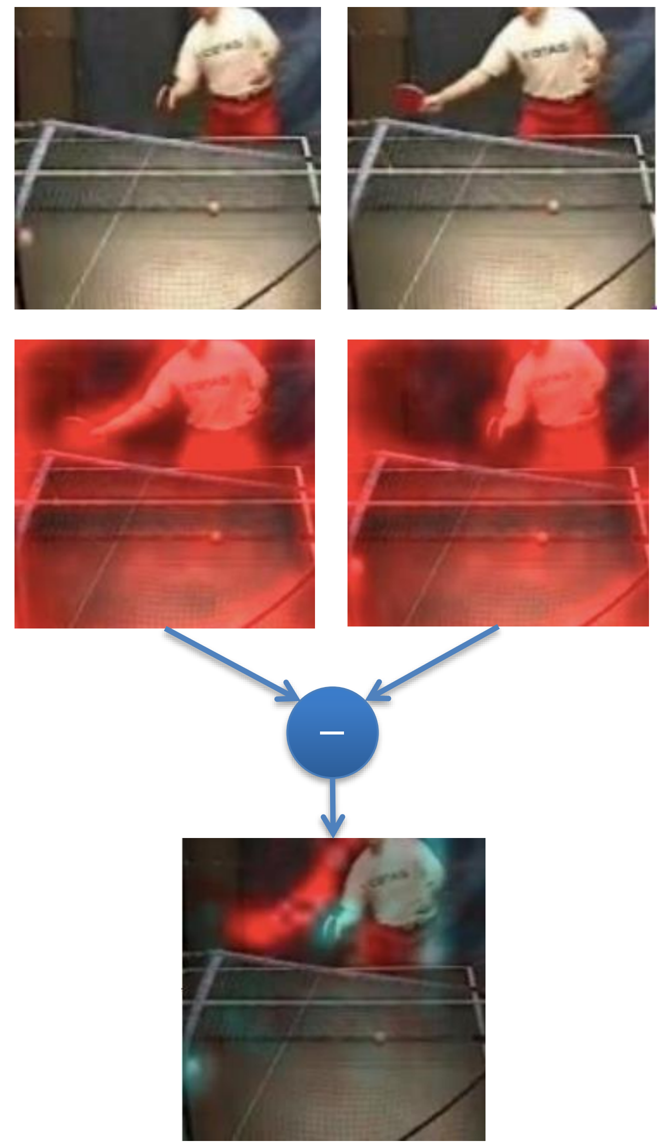
- Motion is the important information
- Optical flow
  - Effective
  - Time consuming

We need a better motion representation



Modality	Acc.
RGB	85.5%
RGB+Optical Flow	94.0%

Optical flow guided feature



Optical flow guided feature

Optical flow:

$$I(x,y,t) = I(x+\Delta x, y+\Delta y, t+\Delta t)$$

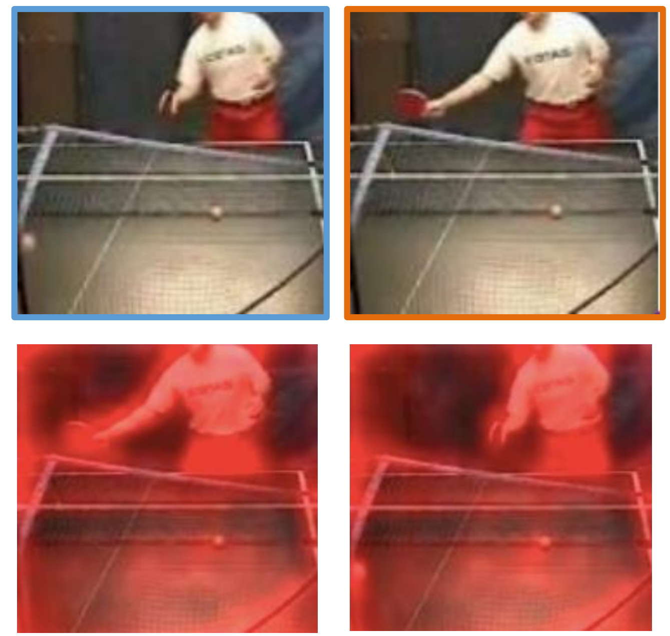
$$\frac{\partial I(x,y,t)}{\partial x} v_x + \frac{\partial I(x,y,t)}{\partial y} v_y + \frac{\partial I(x,y,t)}{\partial t} = 0$$

$\{v_x, v_y\}$ = optical flow

Intuitive Inspiration

Coefficient for optical flow:

$$\{\frac{\partial I(x,y,t)}{\partial x}, \frac{\partial I(x,y,t)}{\partial y}, \frac{\partial I(x,y,t)}{\partial t}\}$$



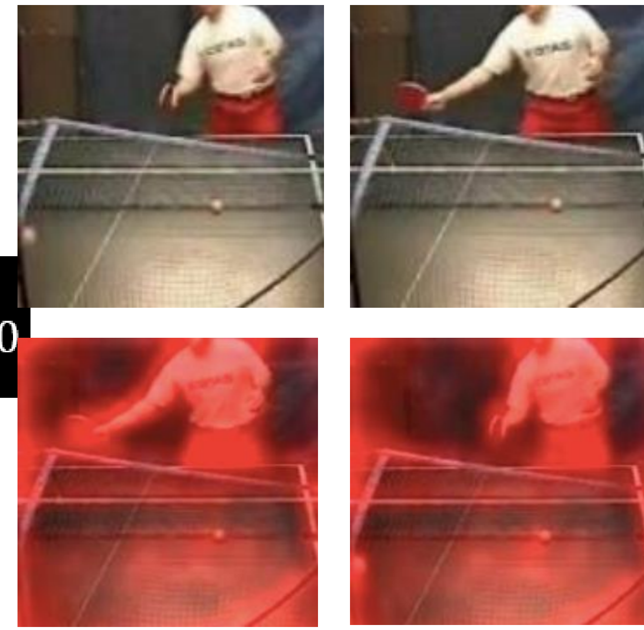
Optical flow guided feature

Feature flow:

$$f(I(x,y,t))=f(I(x+\Delta x,y+\Delta y,t+\Delta t))$$

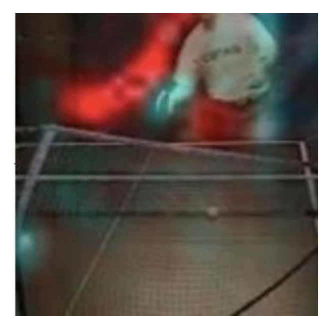
$$\frac{\partial f(I(x,y,t))}{\partial x}\tilde{v}_x+\frac{\partial f(I(x,y,t))}{\partial y}\tilde{v}_y+\frac{\partial f(I(x,y,t))}{\partial t}=0$$

$\{\tilde{v}_x,\tilde{v}_y\}$ = feature flow

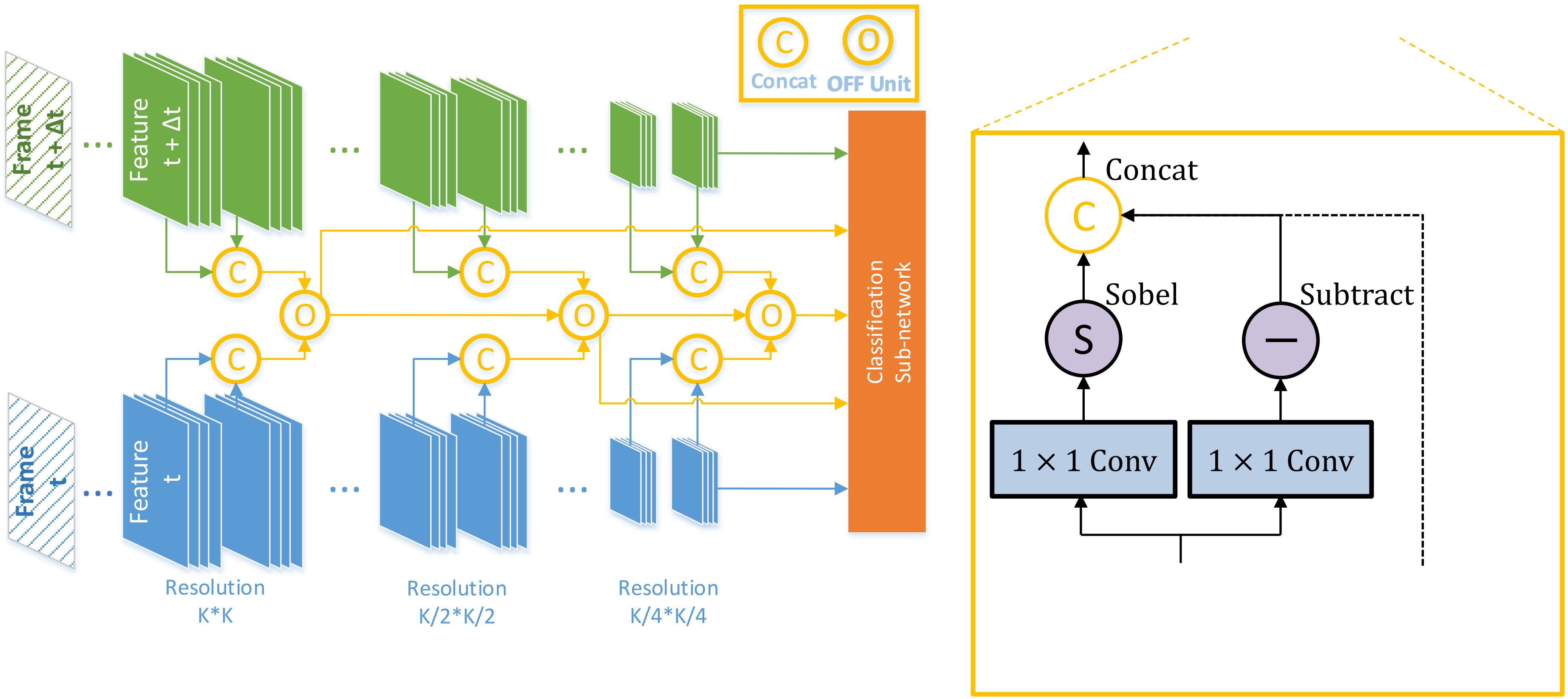


Optical flow guided feature (OFF):

$$\{\frac{\partial f(I(x,y,t);w)}{\partial x},\frac{\partial f(I(x,y,t);w)}{\partial y},\frac{\partial f(I(x,y,t);w)}{\partial t}\}$$

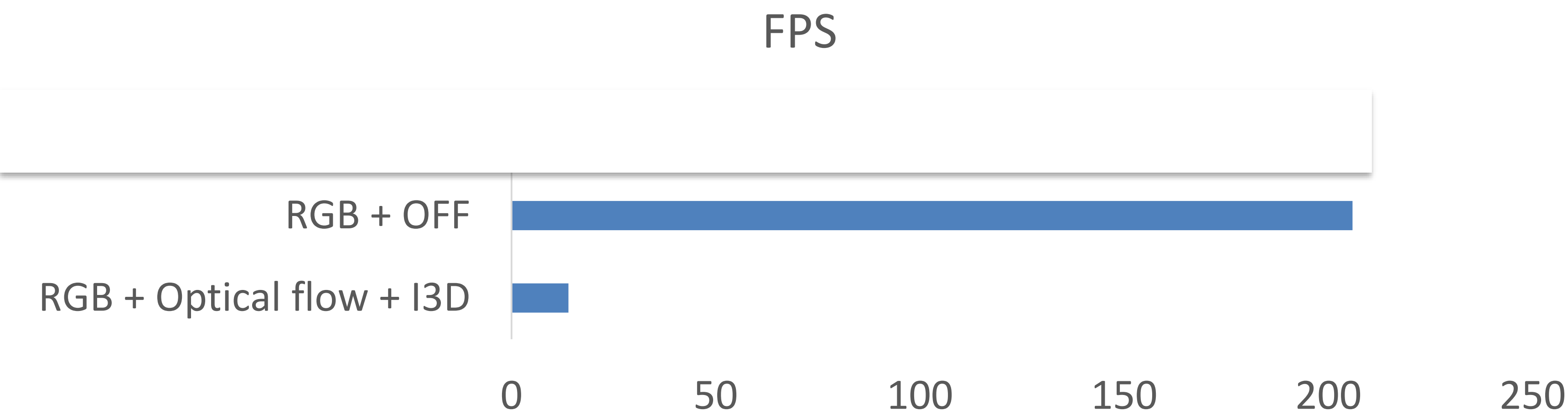


Optical flow guided feature

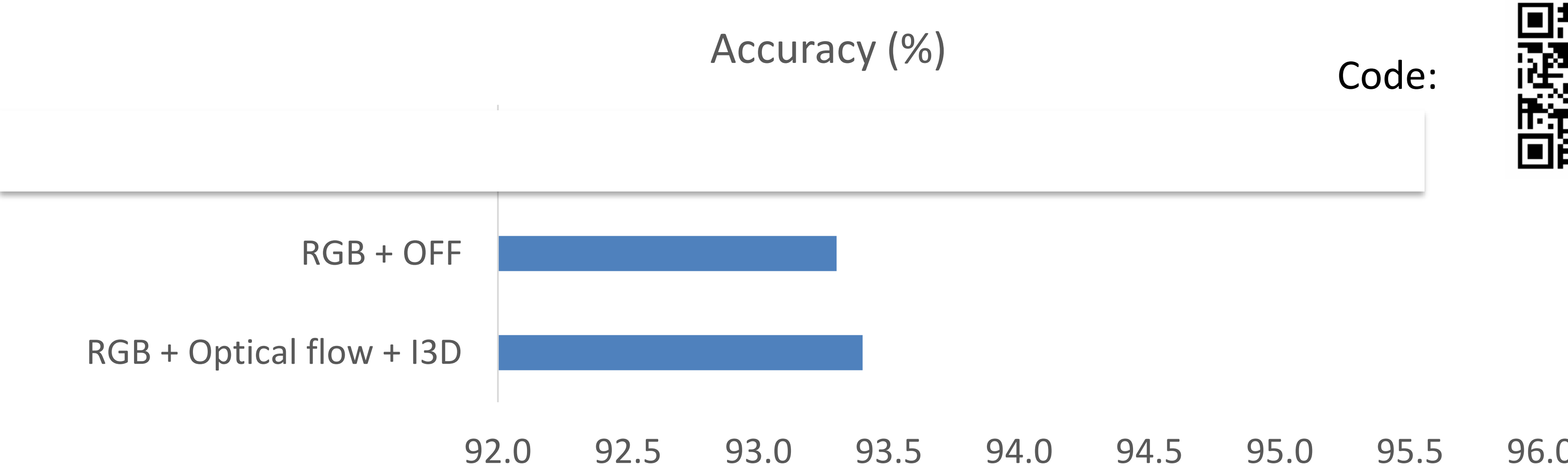


Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, Wei Zhang. "Optical Flow Guided Feature: A Motion Representation for Video Action Recognition", *Proc. CVPR*, 2018.

Optical Flow Guided Feature (OFF): Experimental results



Code:

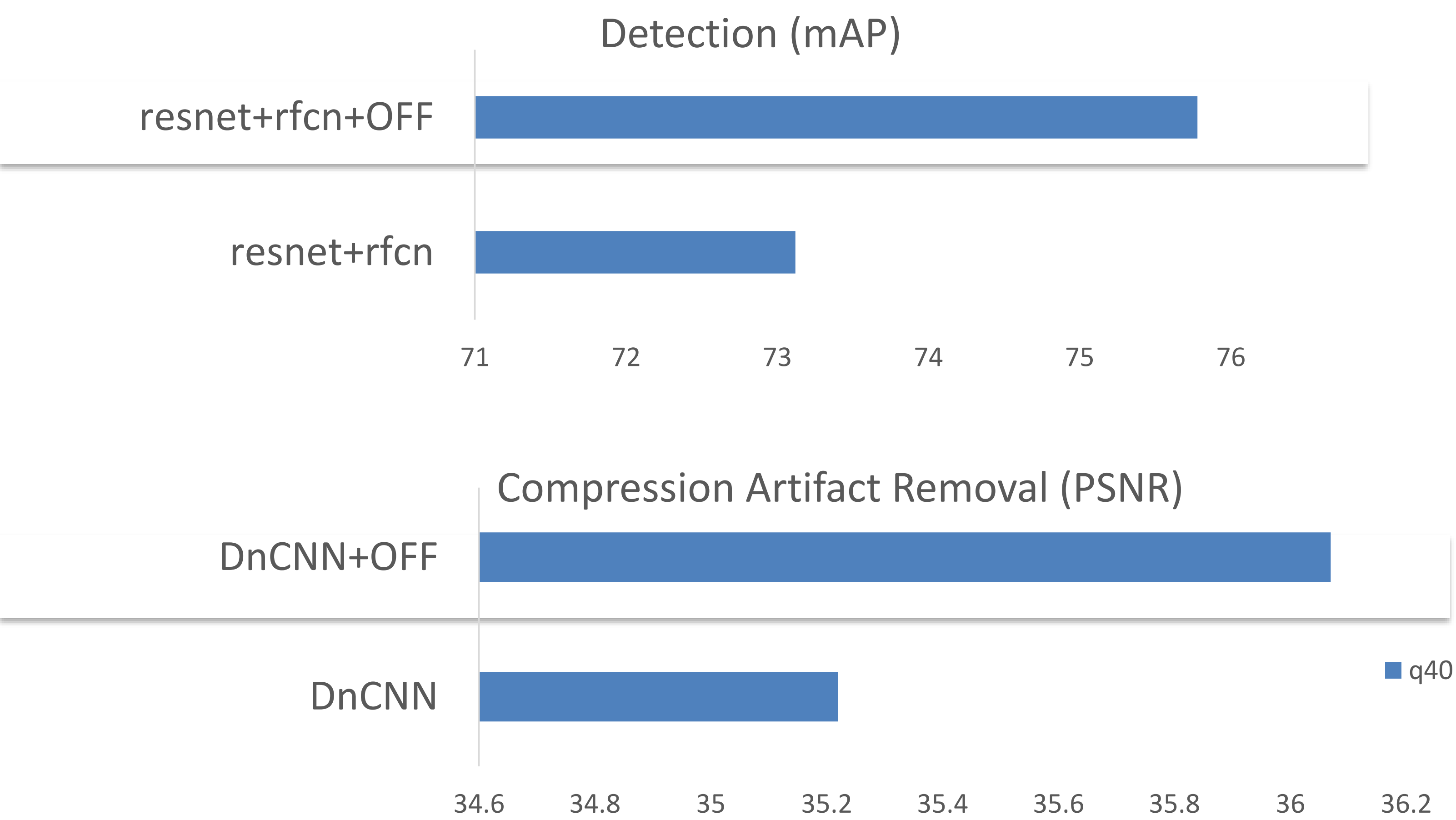


1. OFF with only RGB inputs is **comparable** with the other state-of-the-art methods using optical flow as input.

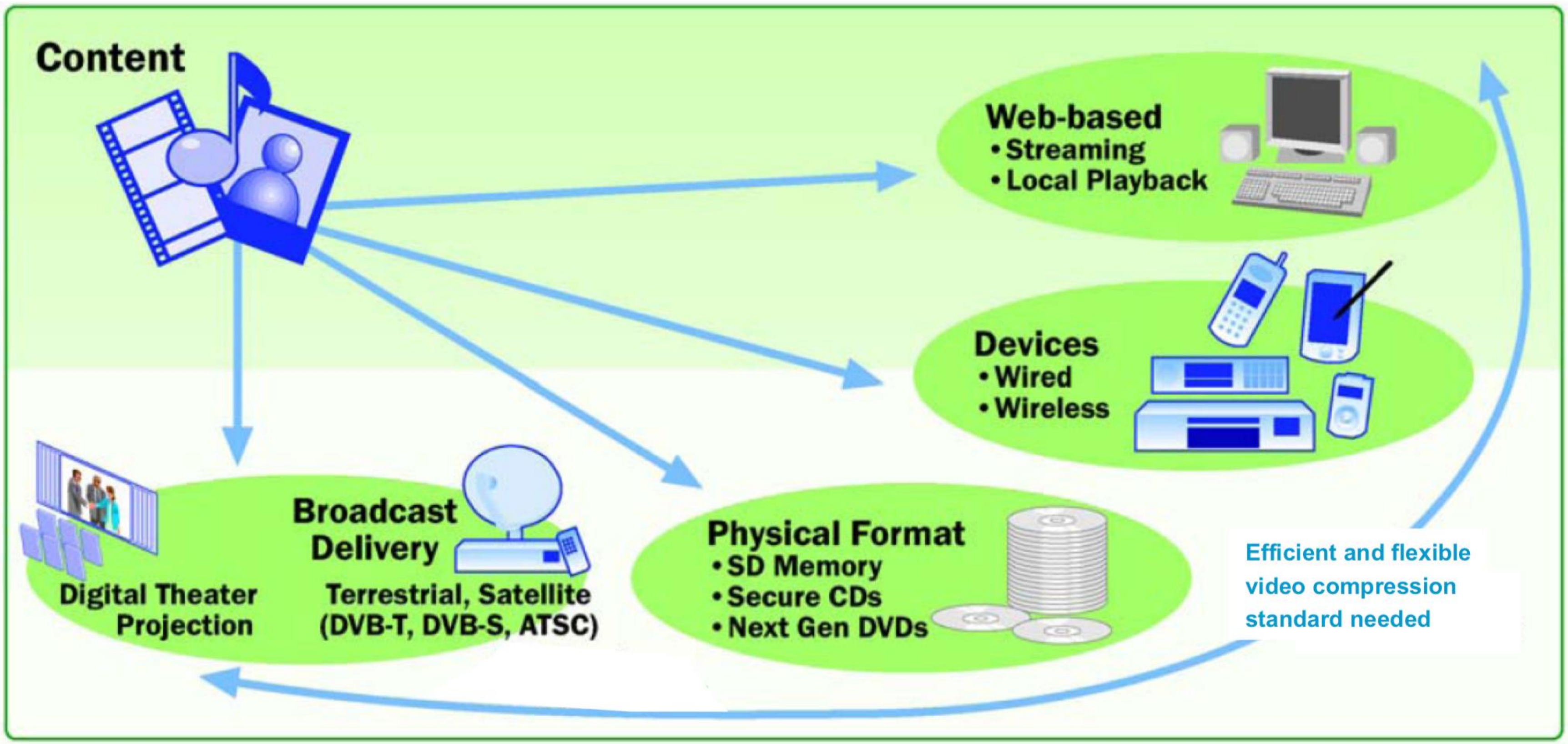
Not only for action recognition

- Also effective for
  - Video object detection
  - Video denoising

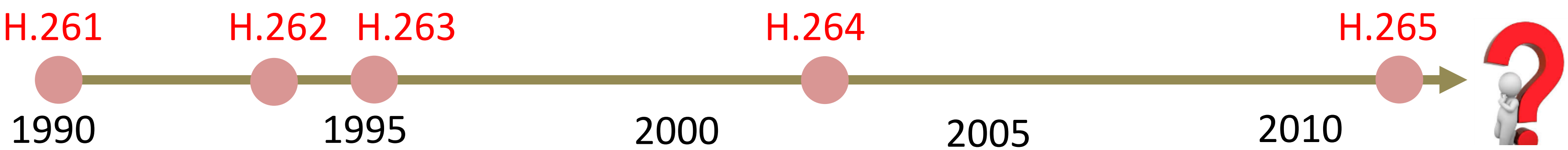
Optical Flow Guided Feature (OFF): Experimental results



1. q40 means quantization factor.



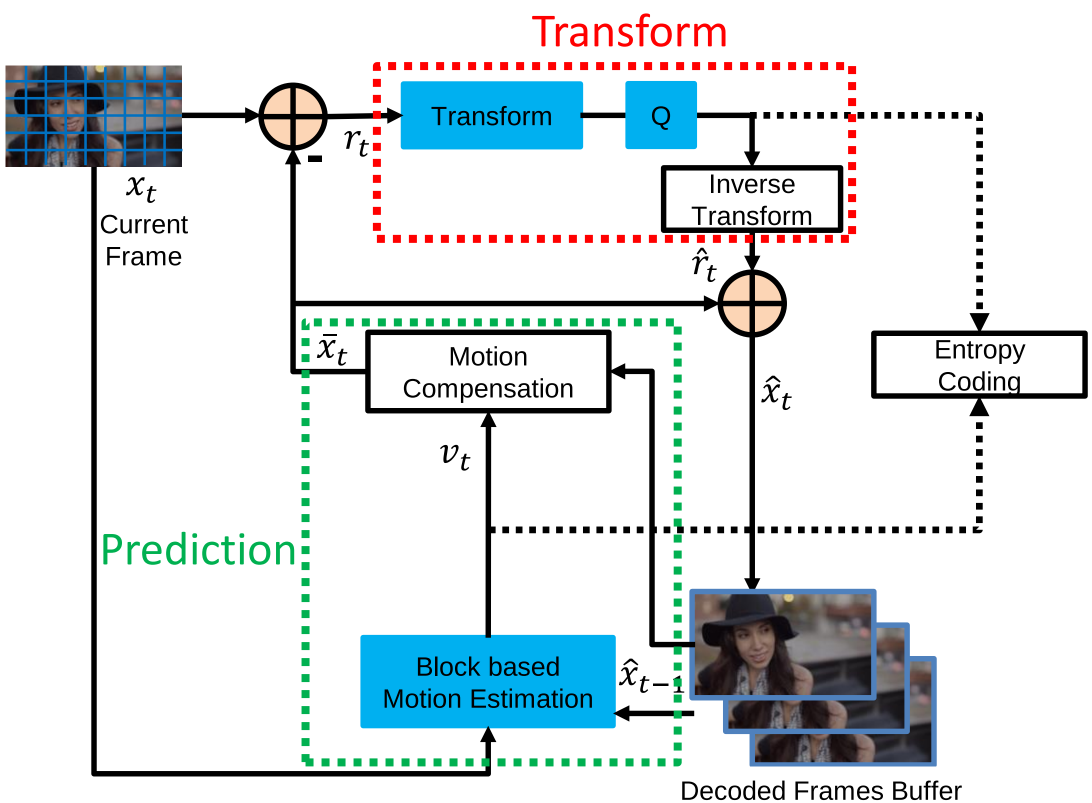
The figure is from Bernd Girod's slides



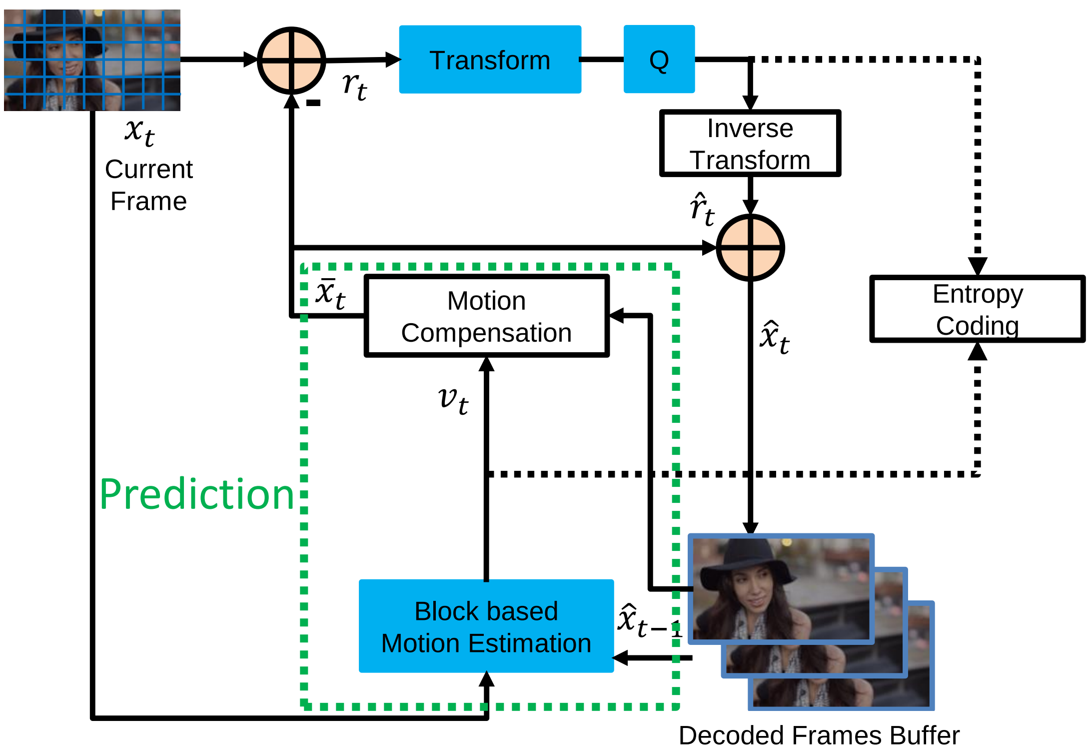
Disadvantages:

- Hand-crafted techniques
- Not friendly for emerging contents
- Not easy to improve the efficiency in the old pipeline

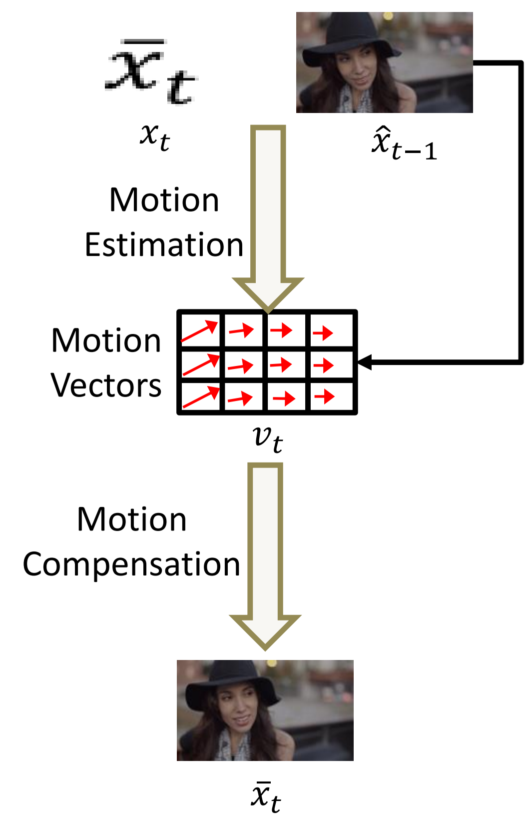
What happens when video compression meets deep learning?



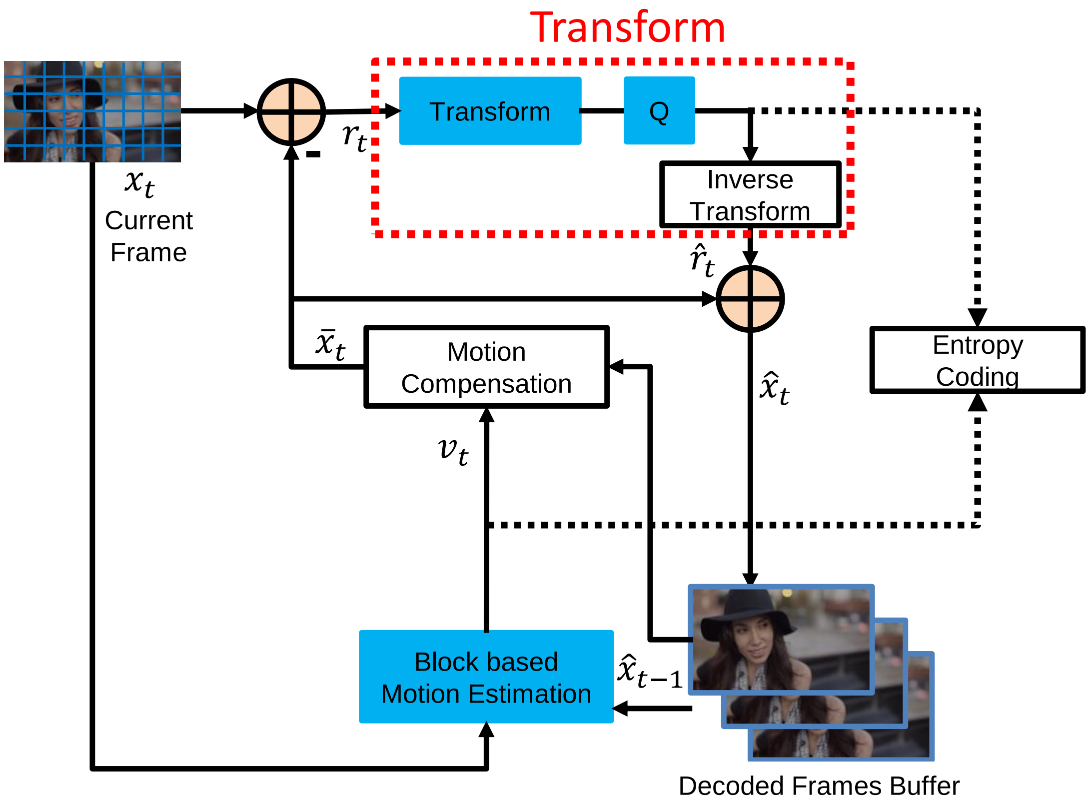
Transform
Transform
Q
r_t
x_t
Current Frame
Inverse Transform
$\hat{r}_t$
$\bar{x}_t$
Motion Compensation
$\hat{x}_t$
Entropy Coding
v_t
Prediction
Block based Motion Estimation
$\hat{x}_{t-1}$
Decoded Frames Buffer



x_t
Current
Frame
r_t
Transform
Q
Inverse
Transform
$\hat{r}_t$
$\hat{x}_t$
Entropy
Coding
$\bar{x}_t$
Motion
Compensation
v_t
Prediction
Block based
Motion Estimation
$\hat{x}_{t-1}$
Decoded Frames Buffer



$\bar{x}_t$
x_t
$\hat{x}_{t-1}$
Motion
Estimation
Motion
Vectors
v_t
Motion
Compensation
$\bar{x}_t$



$$r_t \left\{ \begin{bmatrix} 3 & 7 & 4 & 2 \\ 0 & 3 & 4 & 5 \\ 1 & 10 & 8 & 4 \\ 8 & 0 & 8 & 6 \end{bmatrix} \right\}$$

$$\Downarrow \text{DCT}$$

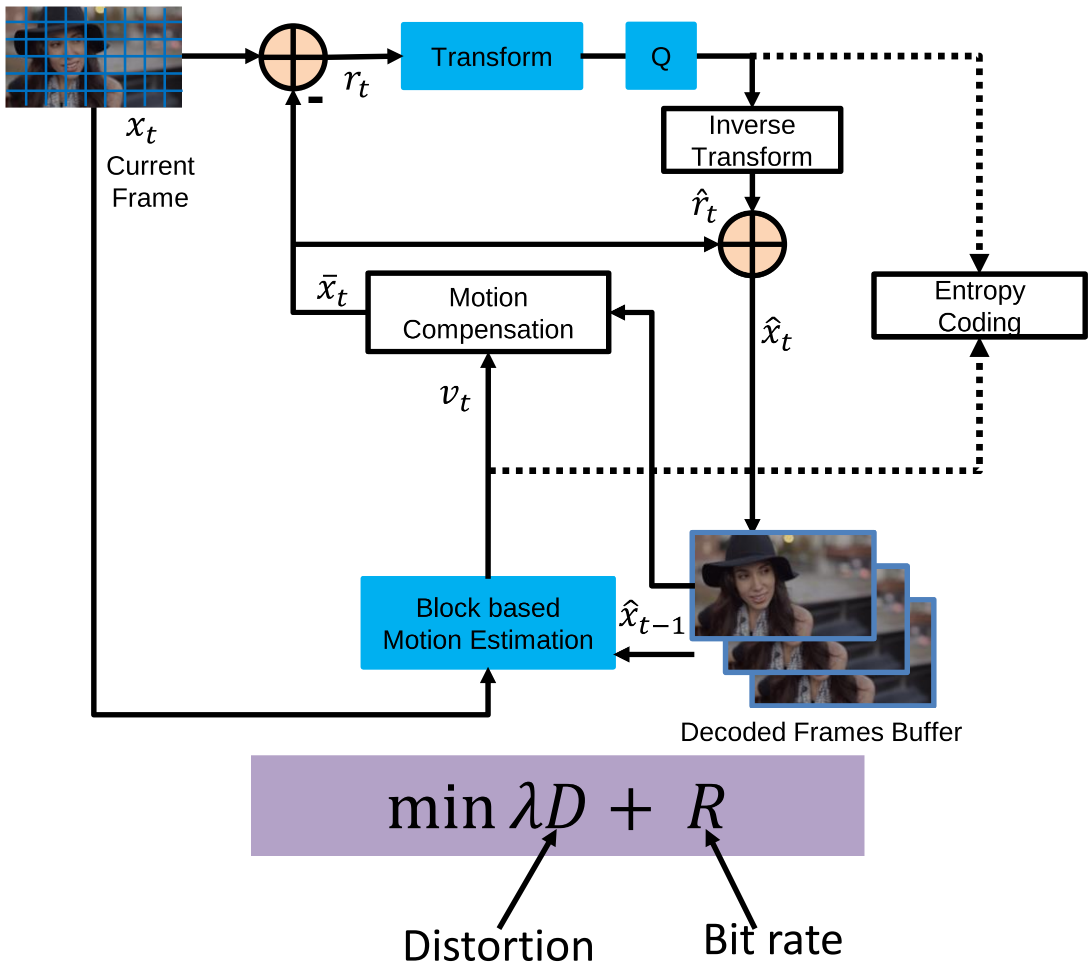
$$\left\{ \begin{bmatrix} 18 & -2 & -4 & 1 \\ -3 & 1 & -2 & -4 \\ 1 & 3 & 4 & 3 \\ 3 & 2 & -5 & -3 \end{bmatrix} \right\}$$

$$\Downarrow \text{Q}$$

$$\left\{ \begin{bmatrix} 16 & -1 & -2 & 0 \\ -2 & 1 & -1 & -2 \\ 1 & 2 & 4 & 2 \\ 2 & 2 & -4 & -2 \end{bmatrix} \right\}$$

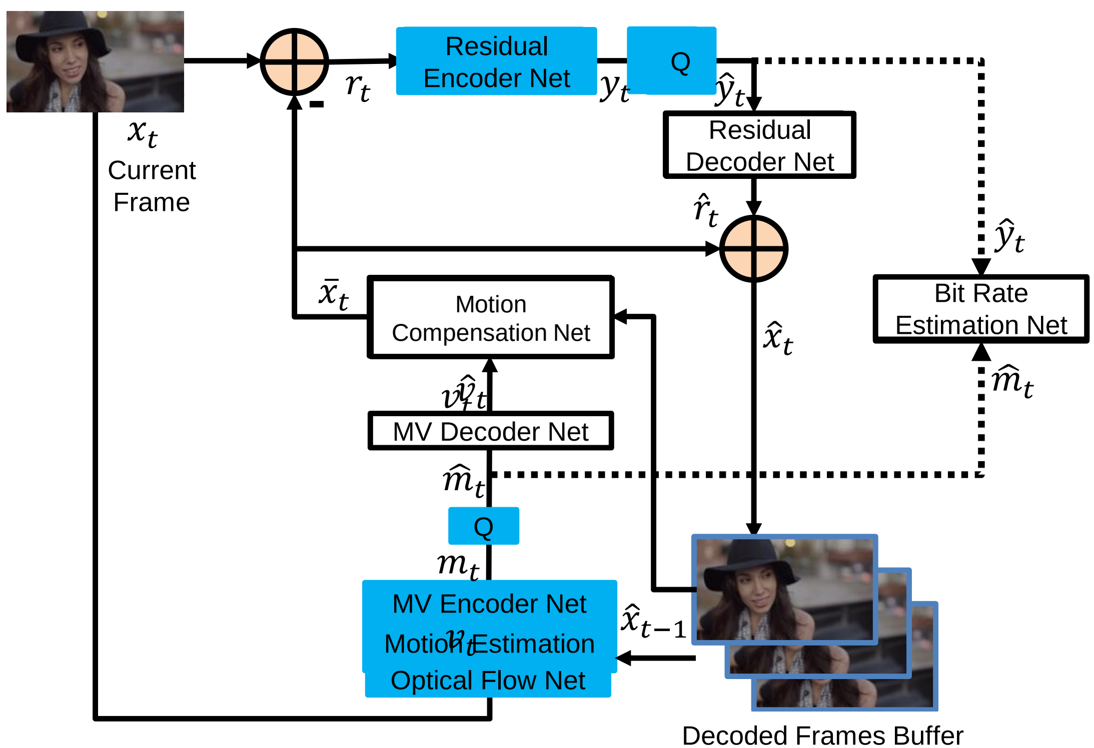
$$\Downarrow \text{IDCT}$$

$$\hat{r}_t \left\{ \begin{bmatrix} 4 & 5 & 3 & 2 \\ 1 & 2 & 4 & 4 \\ 1 & 9 & 6 & 3 \\ 7 & 2 & 6 & 6 \end{bmatrix} \right\}$$



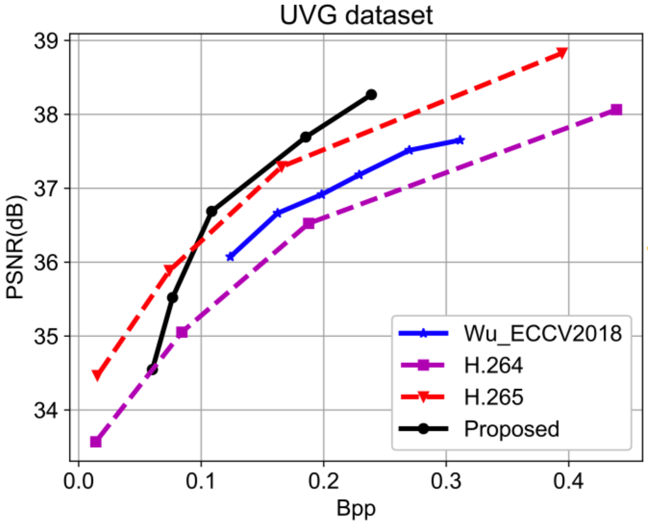
x_t
Current Frame
r_t
Transform
Q
Inverse Transform
$\hat{r}_t$
$\bar{x}_t$
Motion Compensation
$\hat{x}_t$
Entropy Coding
v_t
Block based Motion Estimation
$\hat{x}_{t-1}$
Decoded Frames Buffer
$\min \lambda D + R$
Distortion
Bit rate

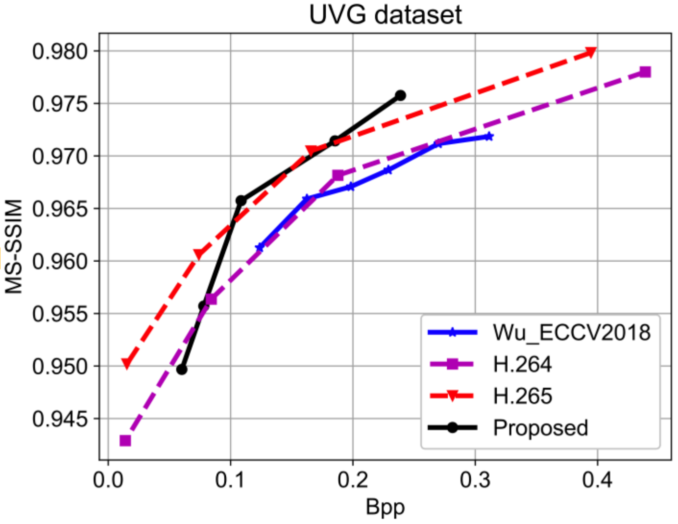
Method



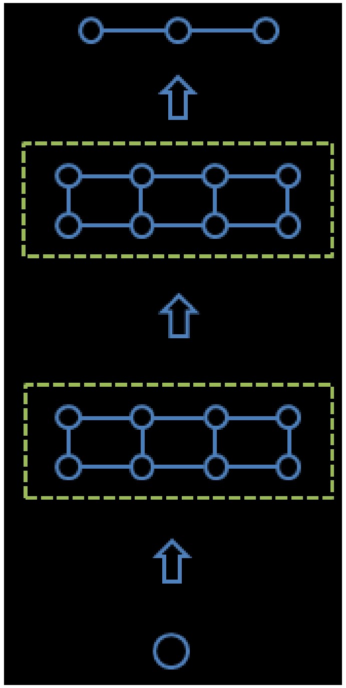
$$\min D + \lambda R$$

Deep Video Compression Model





Take home message



- Structured deep learning is
  - effective
  - for output and features
  - from observation
- End-to-end joint training bridges the gap between structure modeling and feature learning