Modeling deep structures for using high performance images

Wanli Ouyang （欧阳万里）

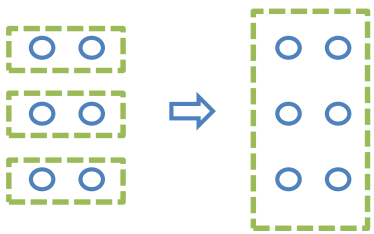
The University of Sydney

Outline

Introduction

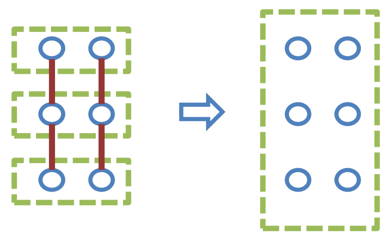
Effectively using high performance images

Feature fusion



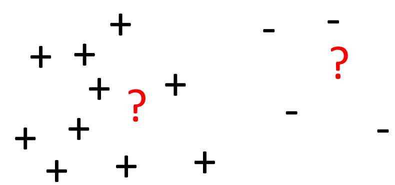
Pedestrian detection (CVPR’17)

Structured features



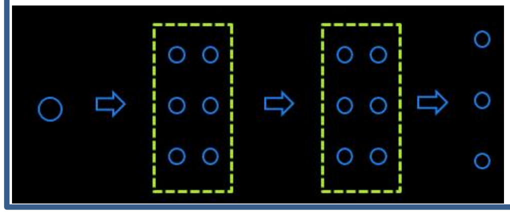
Scene parsing and depth estimation(CVPR18)

Structured samples



3D human pose estimation (CVPR’18)

Back-bone model design



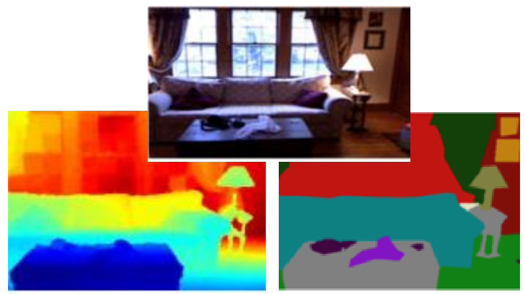
Conclusion

Outline

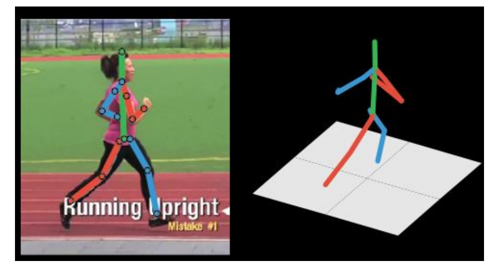
Introduction

Effectively using high performance images

Pedestrian detection (CVPR’17)



Scene parsing and depth estimation(CVPR18)

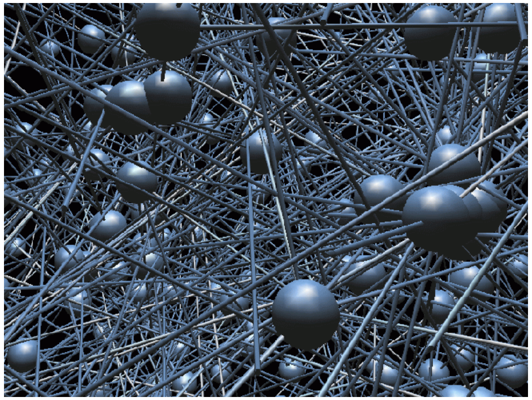


3D human pose estimation (CVPR’18)

Back-bone model design

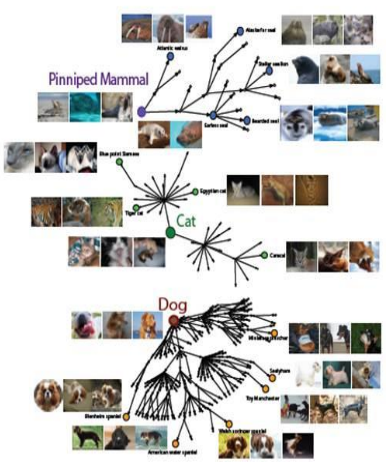
Image/video classification (CVPR’18, NIPS’18)

Conclusion



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



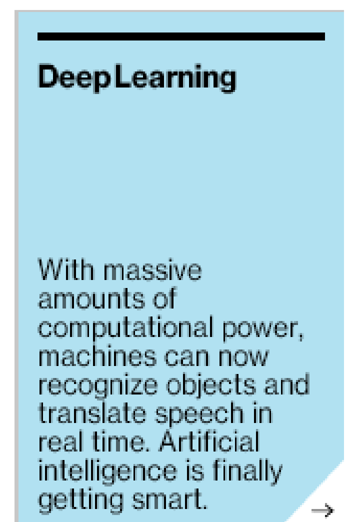
Deep learning

REVOLUTIONARY

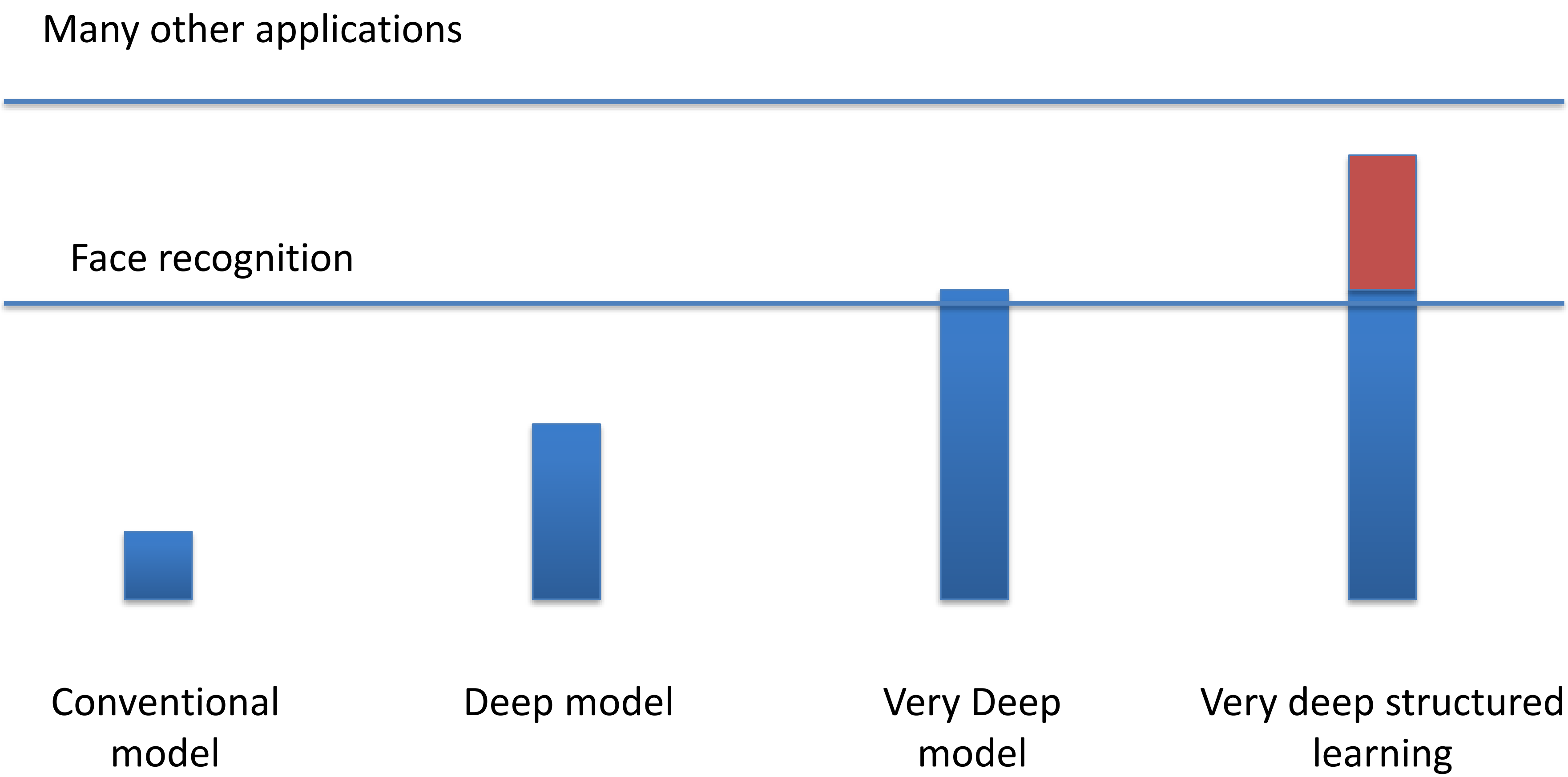
Web-scale visual search, self-driving cars, surveillance, multimedia
…

Hold records on most of the computer vision problems

MIT Tech Review
Top 10 Breakthroughs 2013
Ranking No. 1

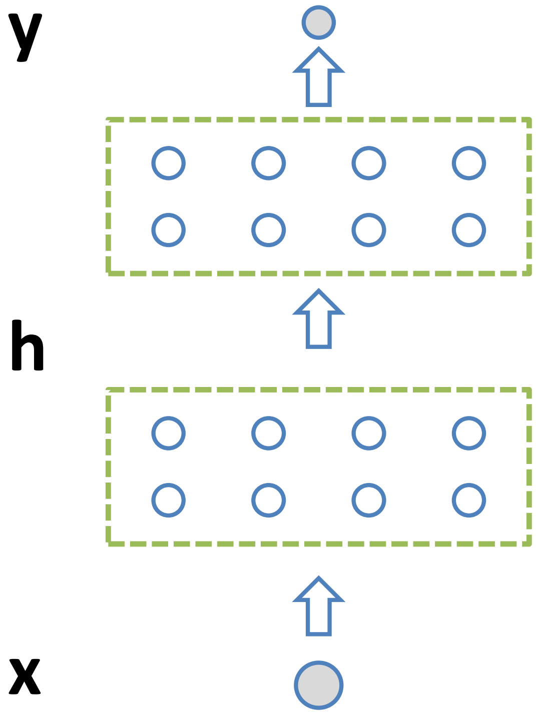


Performance vs practical need



Structure in neurons

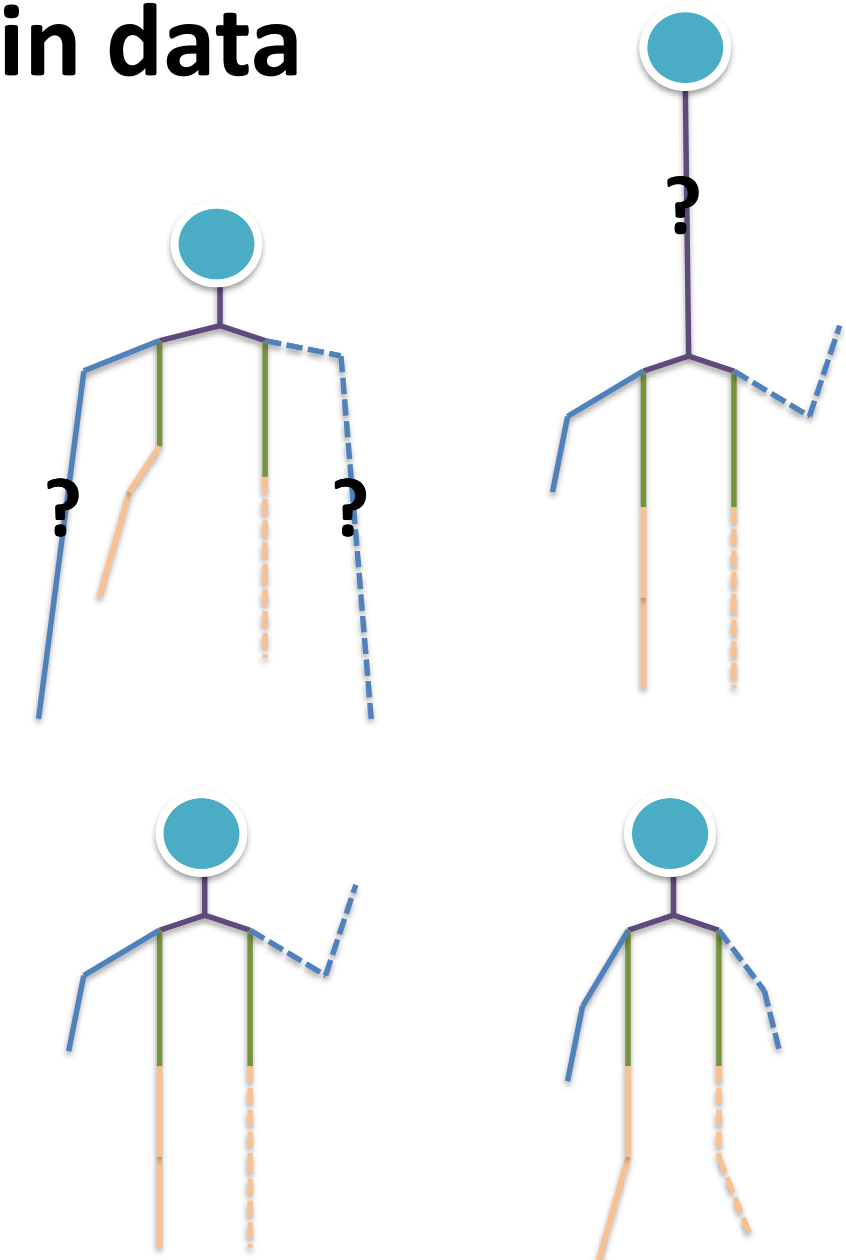
- Conventional neural networks
  - Neurons in the same layer have no connections
  - Neurons in adjacent layers are fully connected, at least within a local region



Structure exists in brain

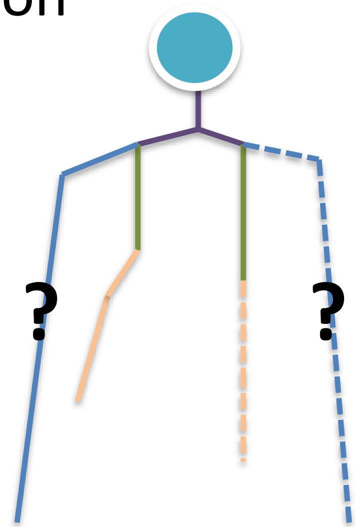
Structure in data

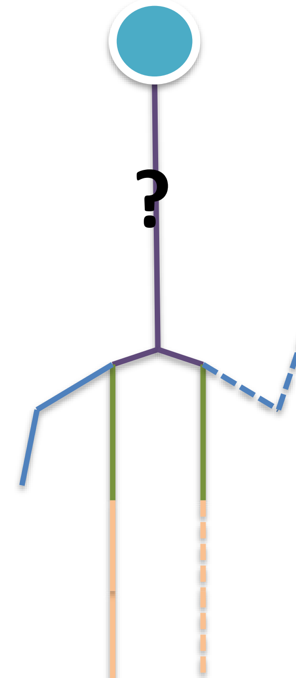
Structure in data

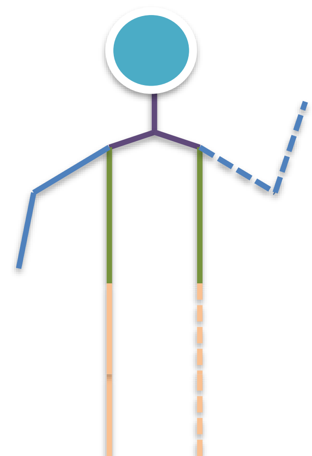


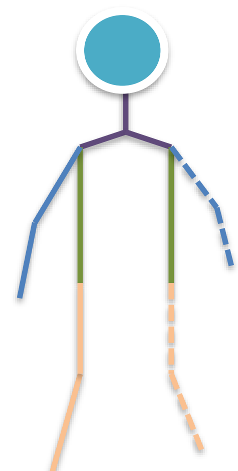
Structure in data

Correlation









Outline

Introduction

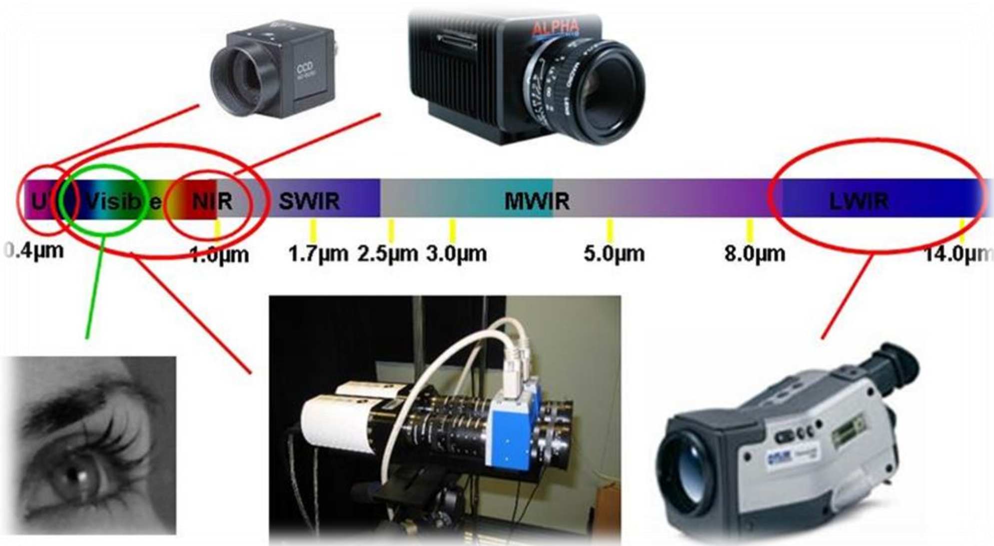
Effectively using high performance images

Back-bone model design

Conclusion

Effectively using high performance imaging data

Training



Multi-modal data

Deployment

RGB only

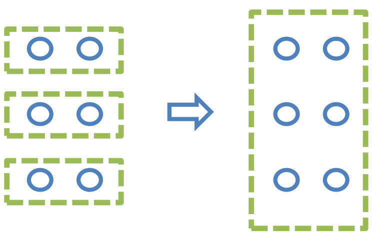
Image from https://research.csiro.au/data61/high-performance-imaging/

Outline

Introduction

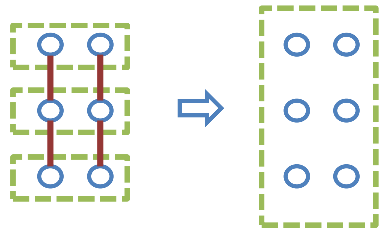
Effectively using high performance images

Feature fusion

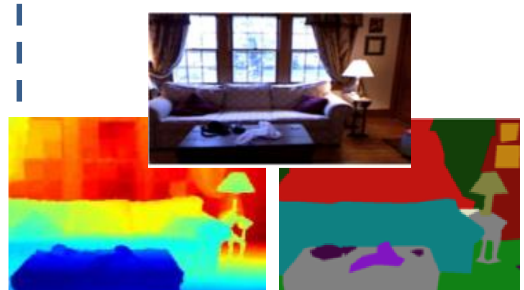


Pedestrian detection (CVPR’17)

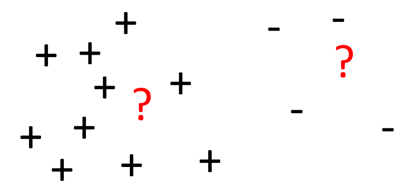
Structured Features



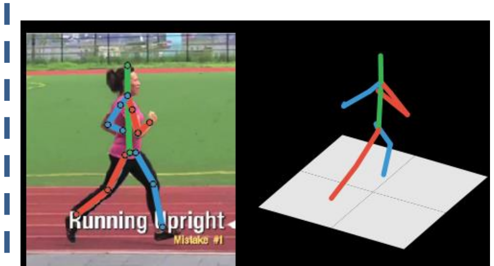
Scene parsing and depth estimation(CVPR18)



Structured Samples



3D human pose estimation (CVPR’18)



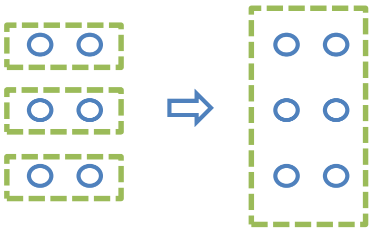
Conclusion

Outline

Introduction

Effectively using high performance images

Feature fusion



Pedestrian detection (CVPR’17)

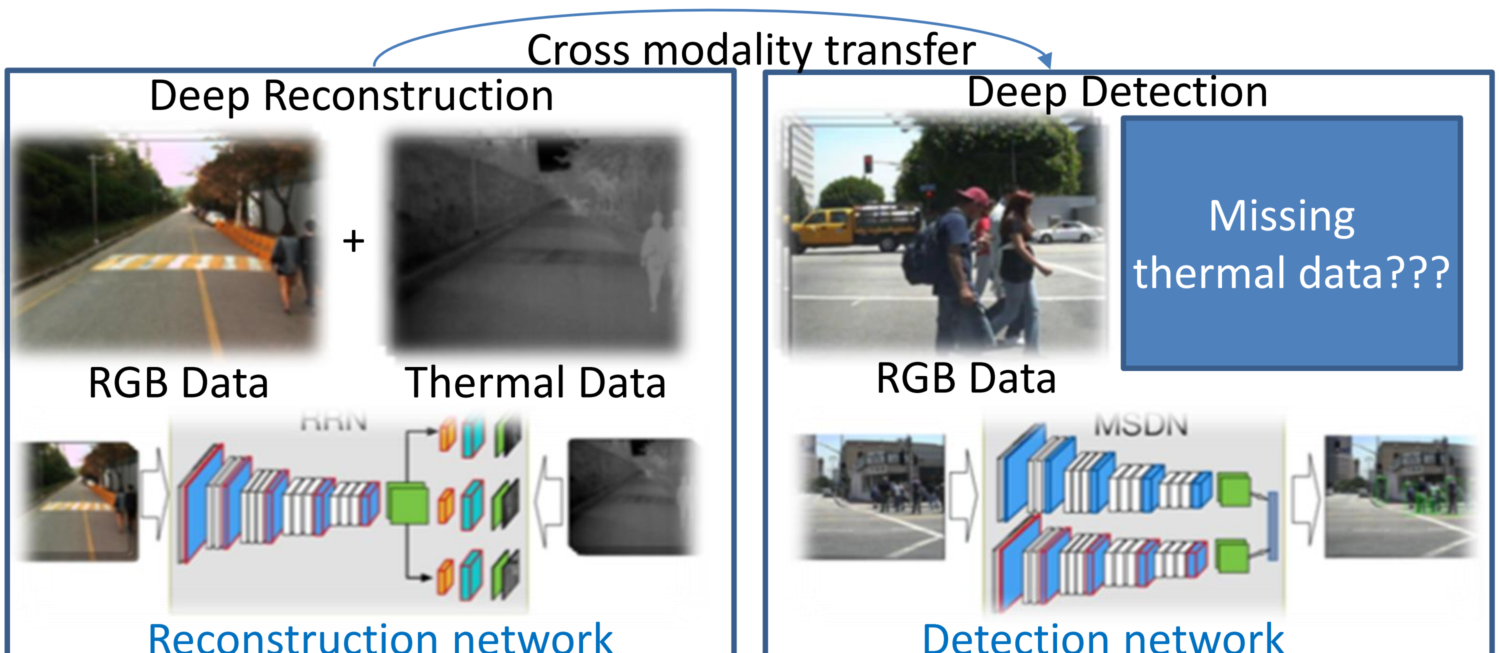
Conclusion

Motivation

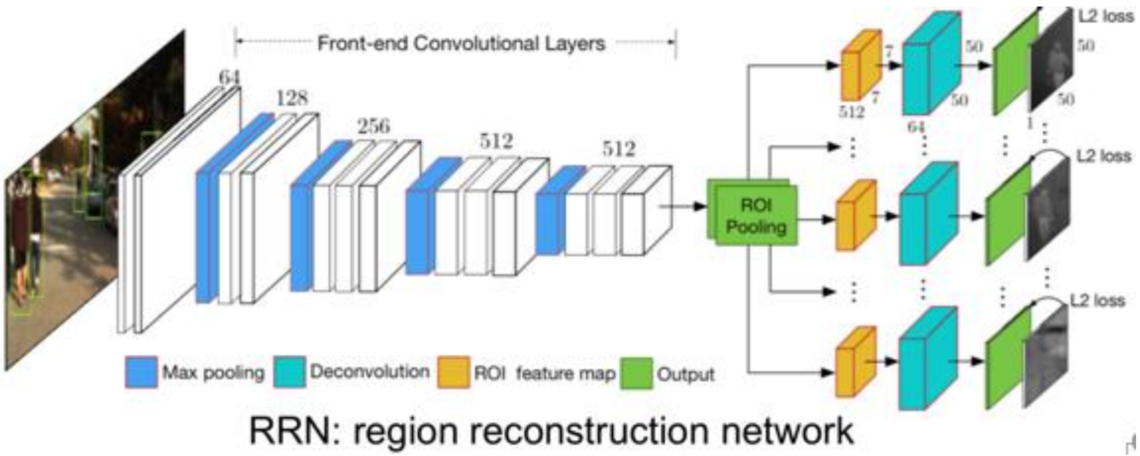
- Challenging open issues in pedestrian detection: illumination variation, shadows, background clutter, and low external light
- Exploiting thermal data in addition to RGB data for learning cross-modal representations

Motivation

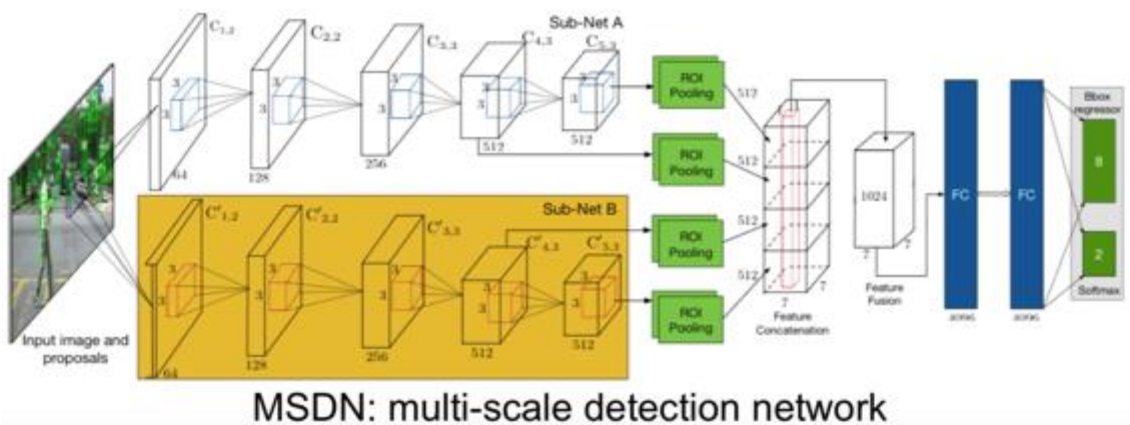
- Challenging open issues in pedestrian detection: illumination variation, shadows, background clutter, and low external light
- Exploiting thermal data in addition to RGB data for learning cross-modal representations
- Can we transfer the learned cross-modal representations?



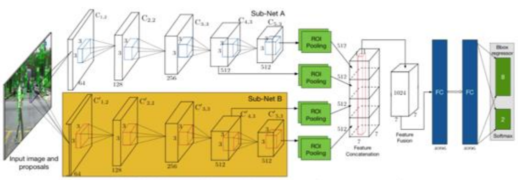
Approach



- RRN
  - RGB domain to Thermal domain
  - weakly supervised reconstruction
  - region-based instead of frame-level based

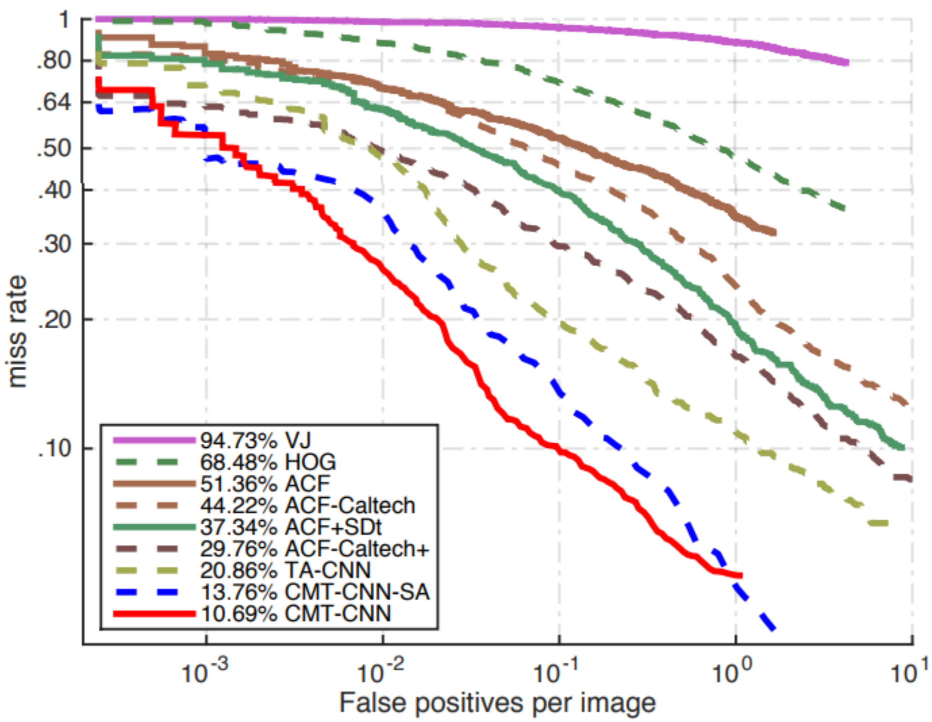


- MSDN
  - cross-modal multi-scale feature fusion
  - the parameters of subnetwork in yellow box are transferred from RRN



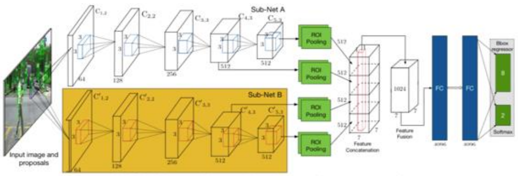
MSDN: multi-scale detection network

Results - Caltech



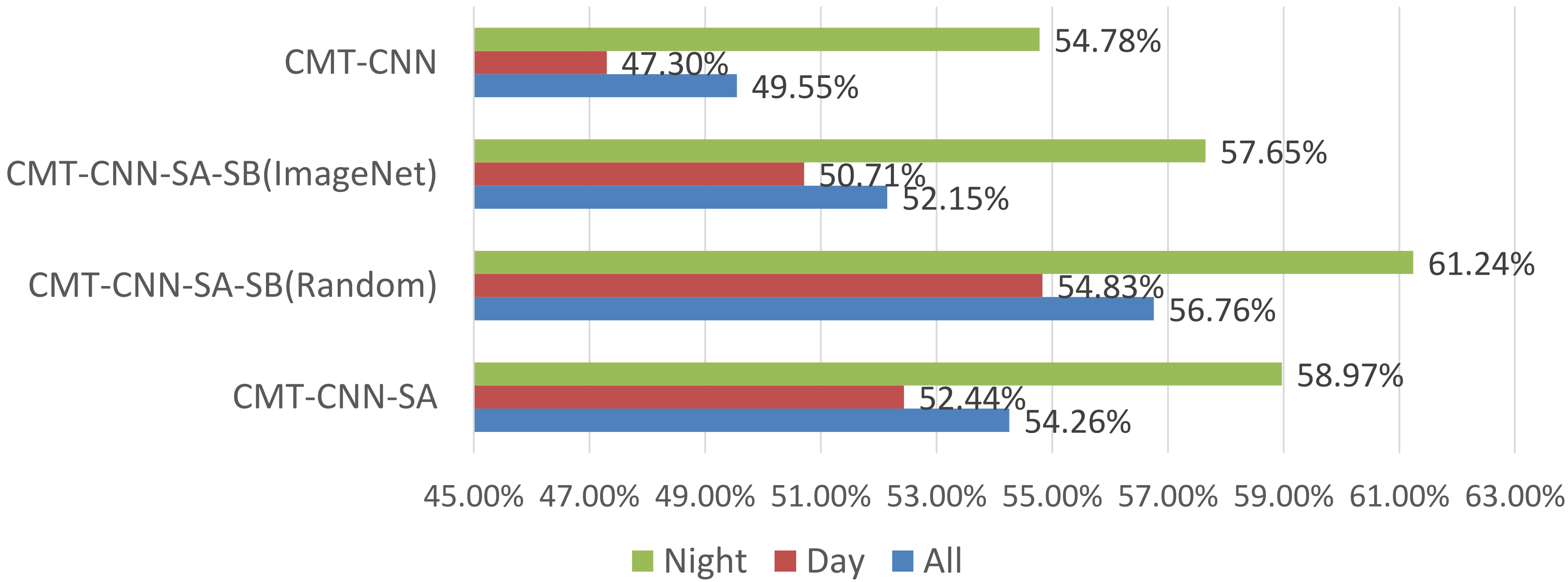
	Average Miss rate
CMT-CNN-SA	13.76%
CMT-CNN	10.69%

- Demonstrated the effectiveness of the learned cross-modal representations
- Achieved superior detection performance

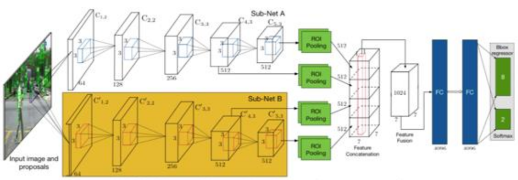


MSDN: multi-scale detection network

Results - KAIST



- Demonstrated the effectiveness of the learned cross-modal representations
- Achieved superior detection performance



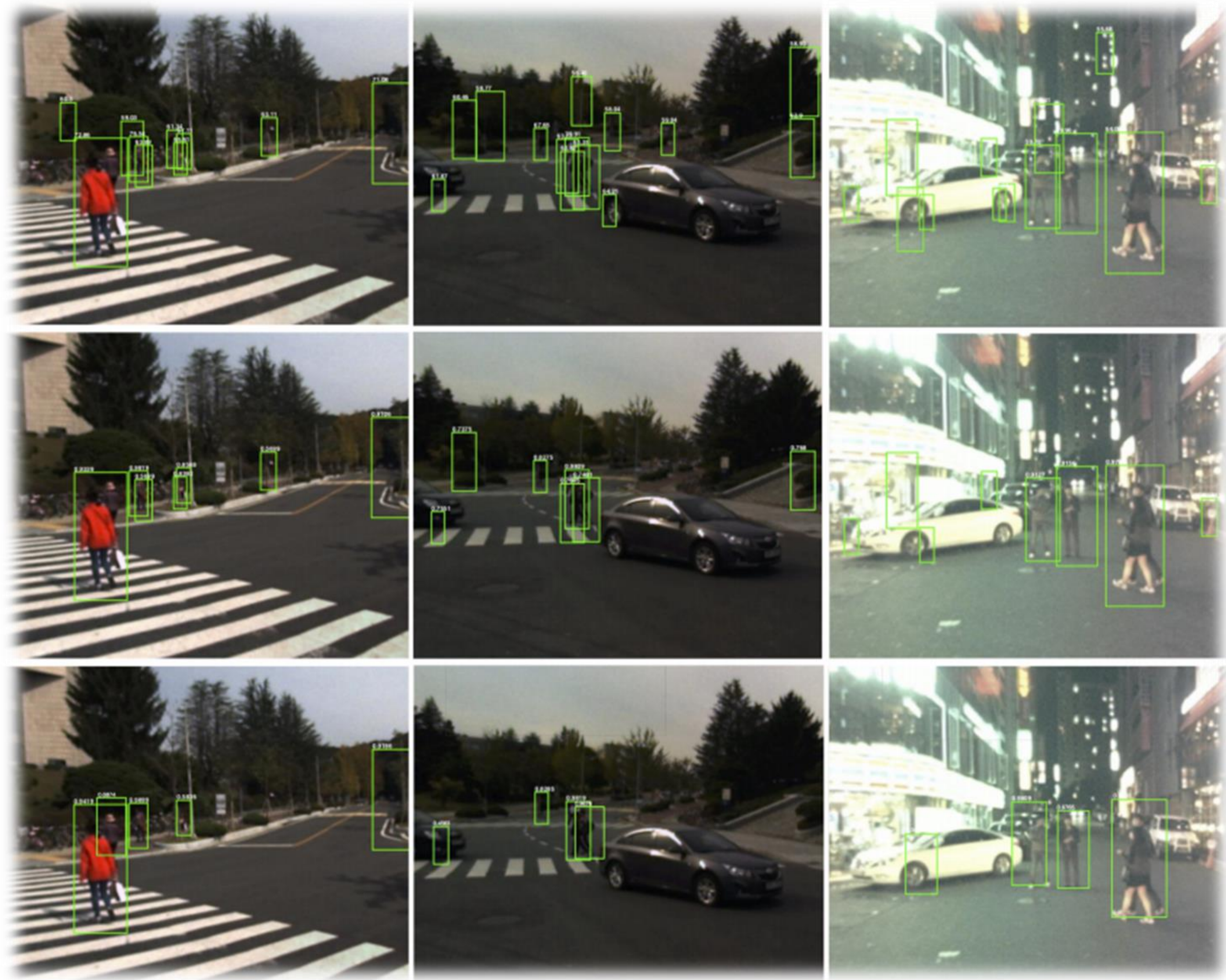
MSDN: multi-scale detection network

Qualitative results

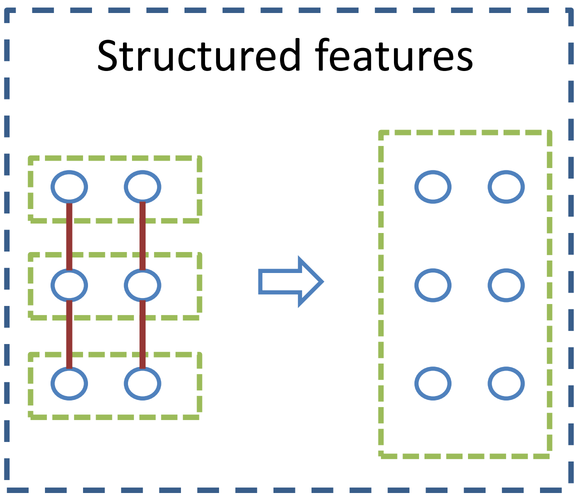
ACF

w/o reconstruction network

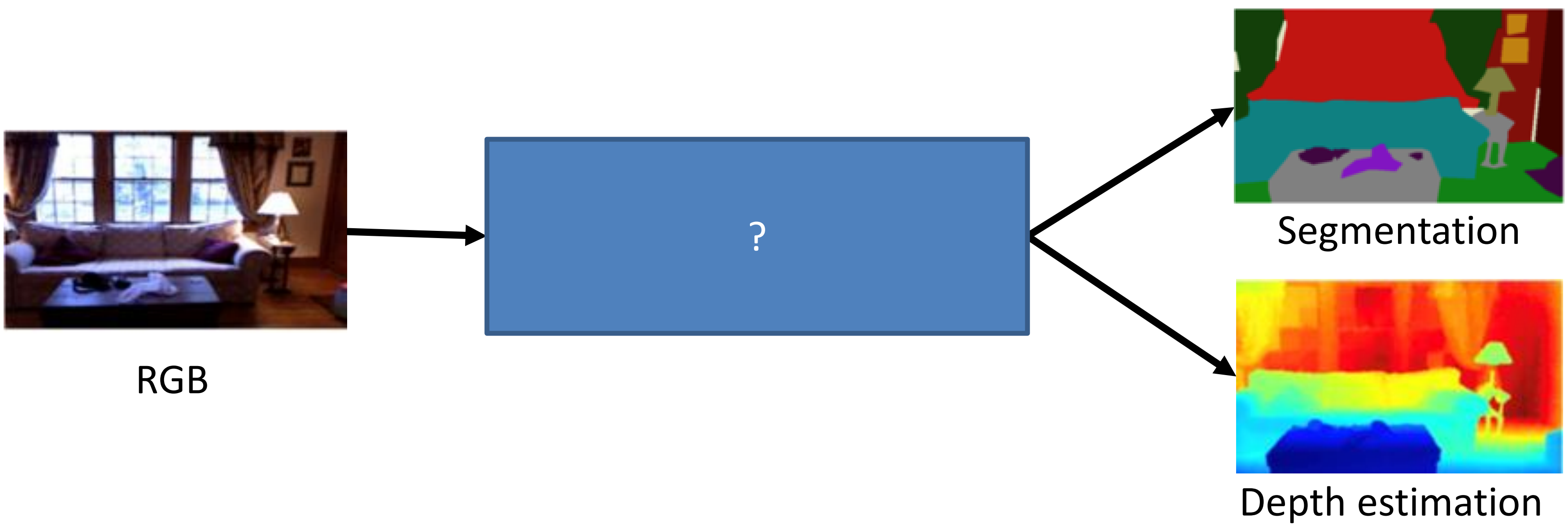
With reconstruction network



Effectively using high performance images

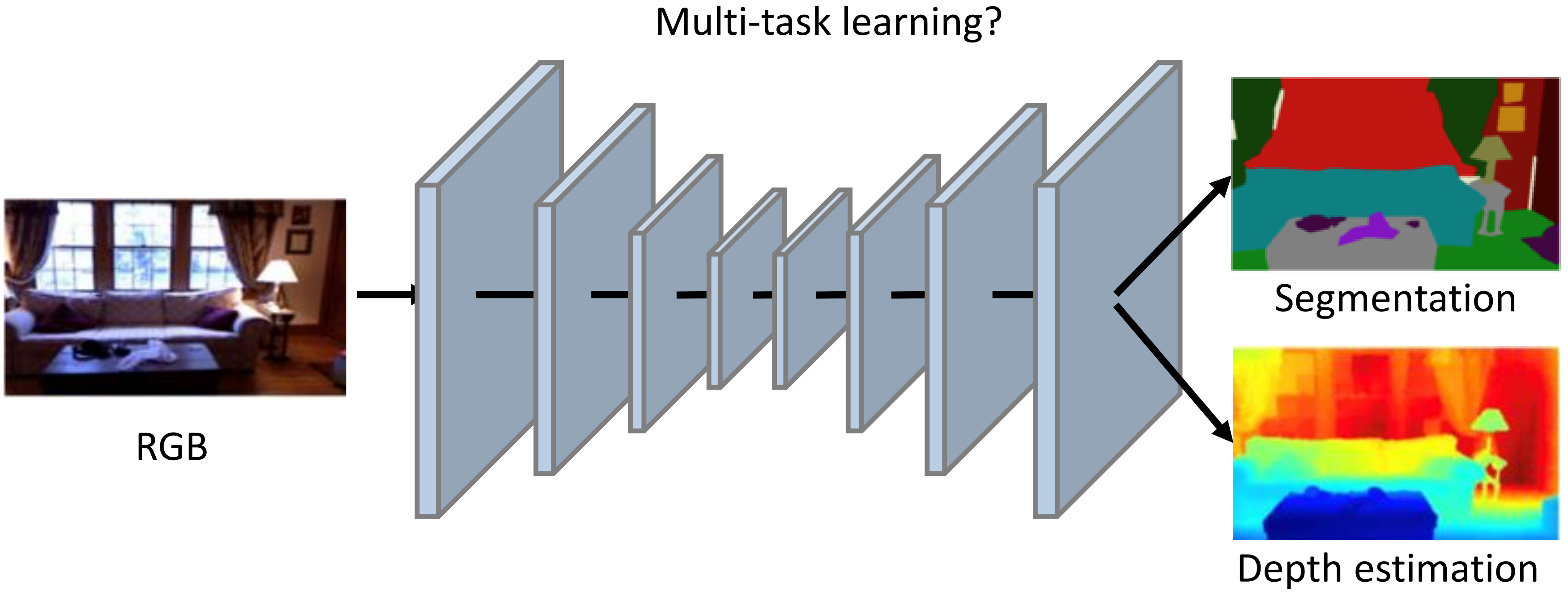


Motivation



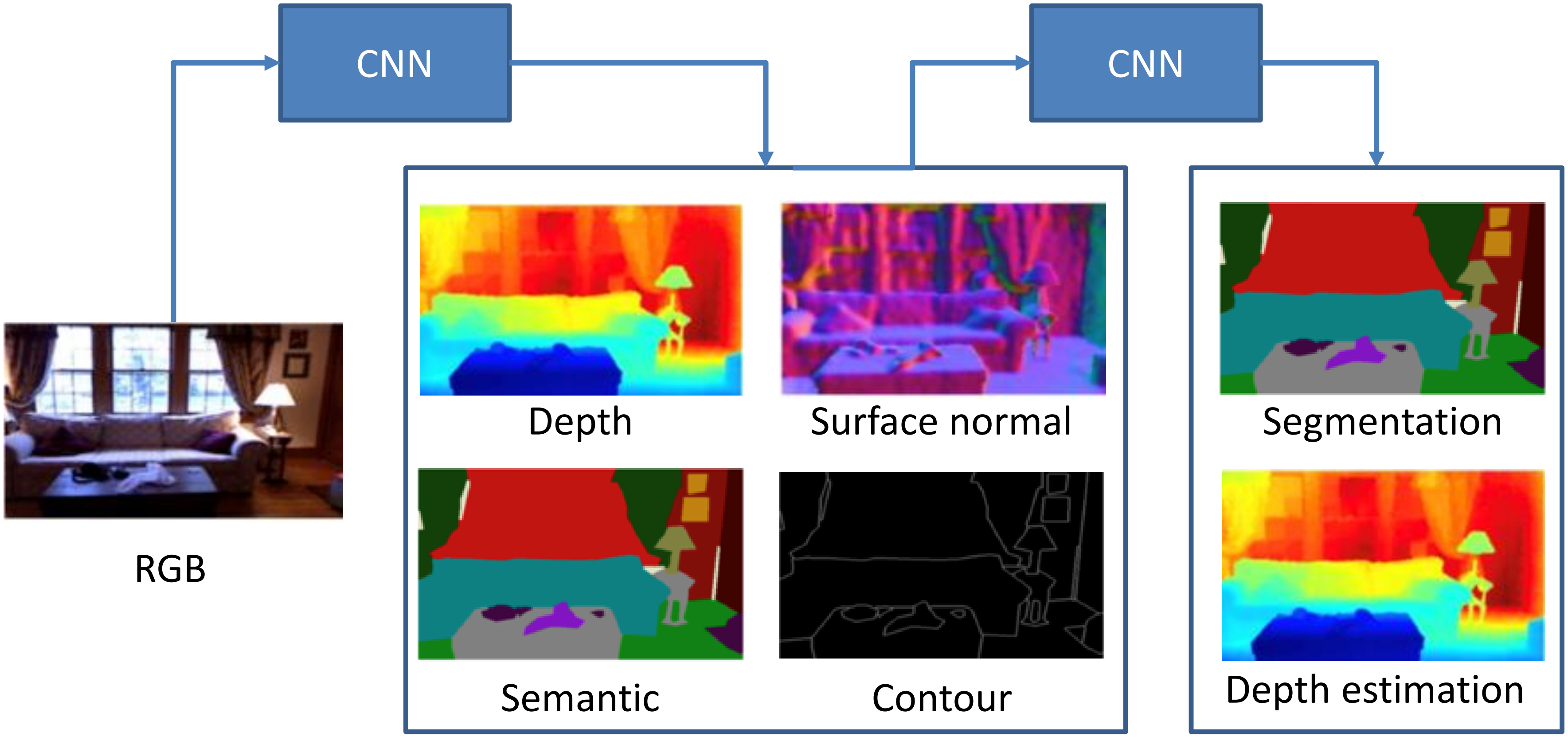
Dan Xu, Wanli Ouyang, Xiaogang Wang, Nicu Sebe, “PAD-Net: Multi-Tasks Guided Prediction-and-Distillation Network for Simultaneous Depth Estimation and Scene Parsing”, IEEE Conference on Computer Vision and Pattern Recognition (***CVPR 2018***)

Motivation



- Directly optimizing multiple tasks given input training data does not guarantee consistent gain on all the tasks

Motivation



- Multi-modal input data improve training of deep networks
- Facilitate final tasks via leveraging intermediate multiple predictions while only one single modal data are required?

Approach

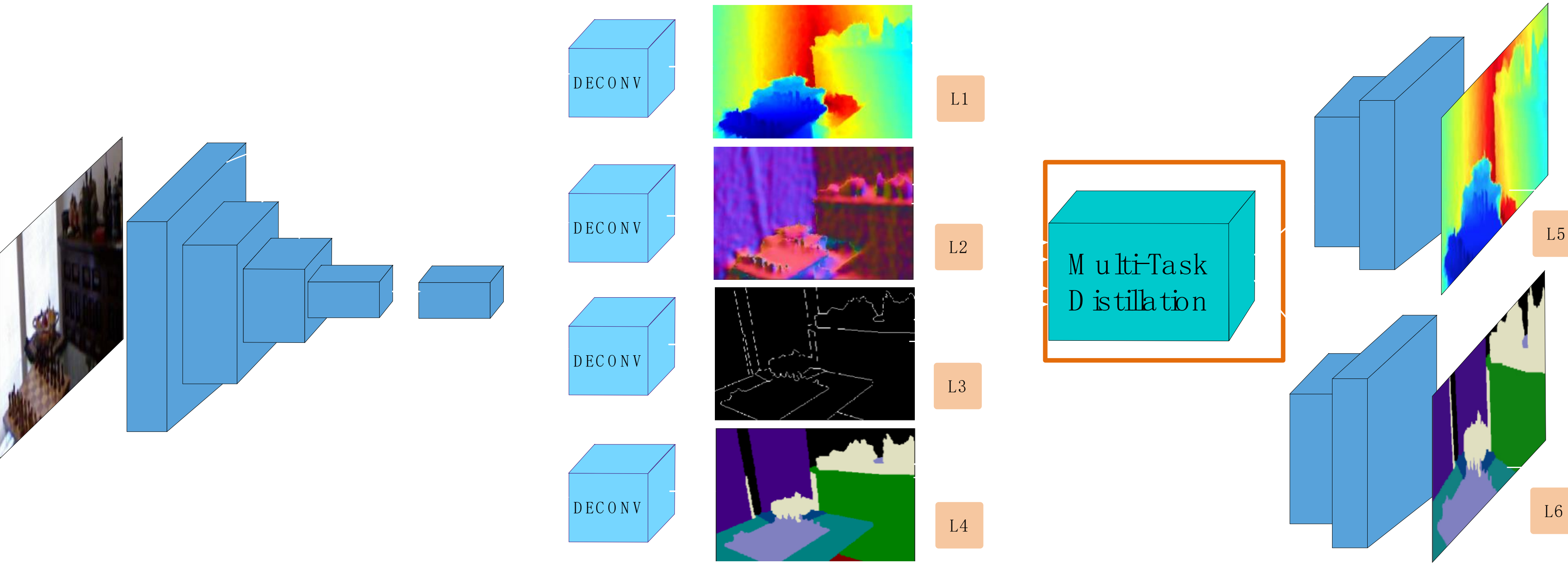
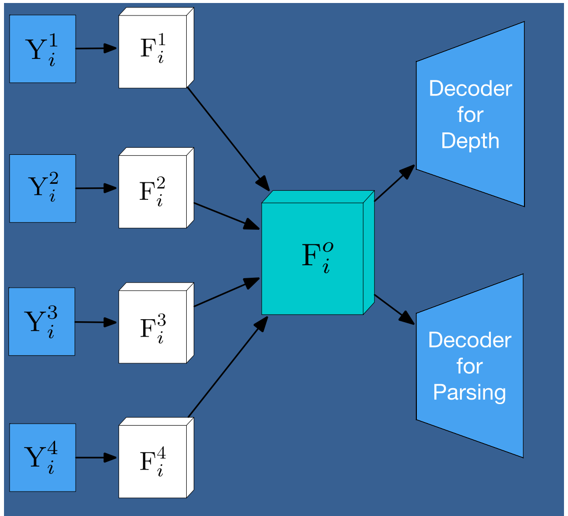


Illustration of the proposed multi-task distillation network for simultaneous depth estimation and scene parsing

Approach



(a) Multi-task Distillation Module A

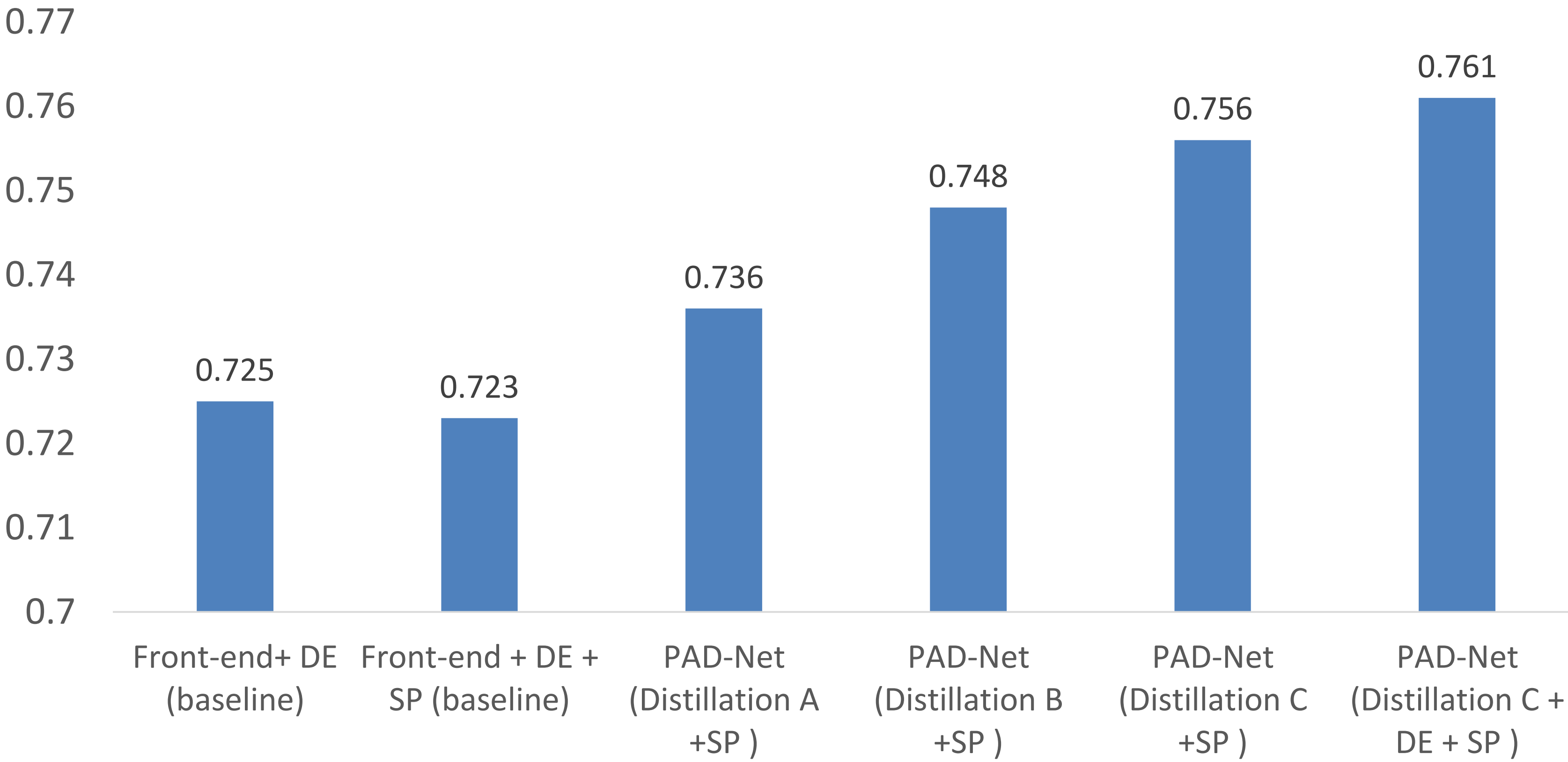
- **Different multi-task distillation modules:**

  Naive implementation via feature concatenation

  Passing message between feature maps

  Attention mechanism guided message passing module

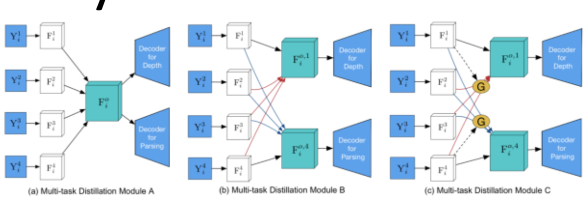
Results for scene parsing on Cityscapes



Multi-task learning results in decrease of IOU accuracy

Distillation improves accuracy

Message passing with attention performs better



Results

- **Datasets: NYUD-V2 and Cityscapes**
- **Ablation study**:
- (i) PAD-Net (Distillation A + DE): PAD-Net performing the DE task using the distillation module A
- (ii) PAD-Net (Distillation B + DE): similar to (i) while using the distillation module B
- (iii) PAD-Net (Distillation B + DE): similar to (i) while using the distillation module C
- (iv) PAD-Net (Distillation C + DE + SP): performing DE and SP tasks simultaneously with the distillation module C
- **Effectiveness on both datasets**
- **Significant improvement over SOTA methods on joint prediction of both tasks**

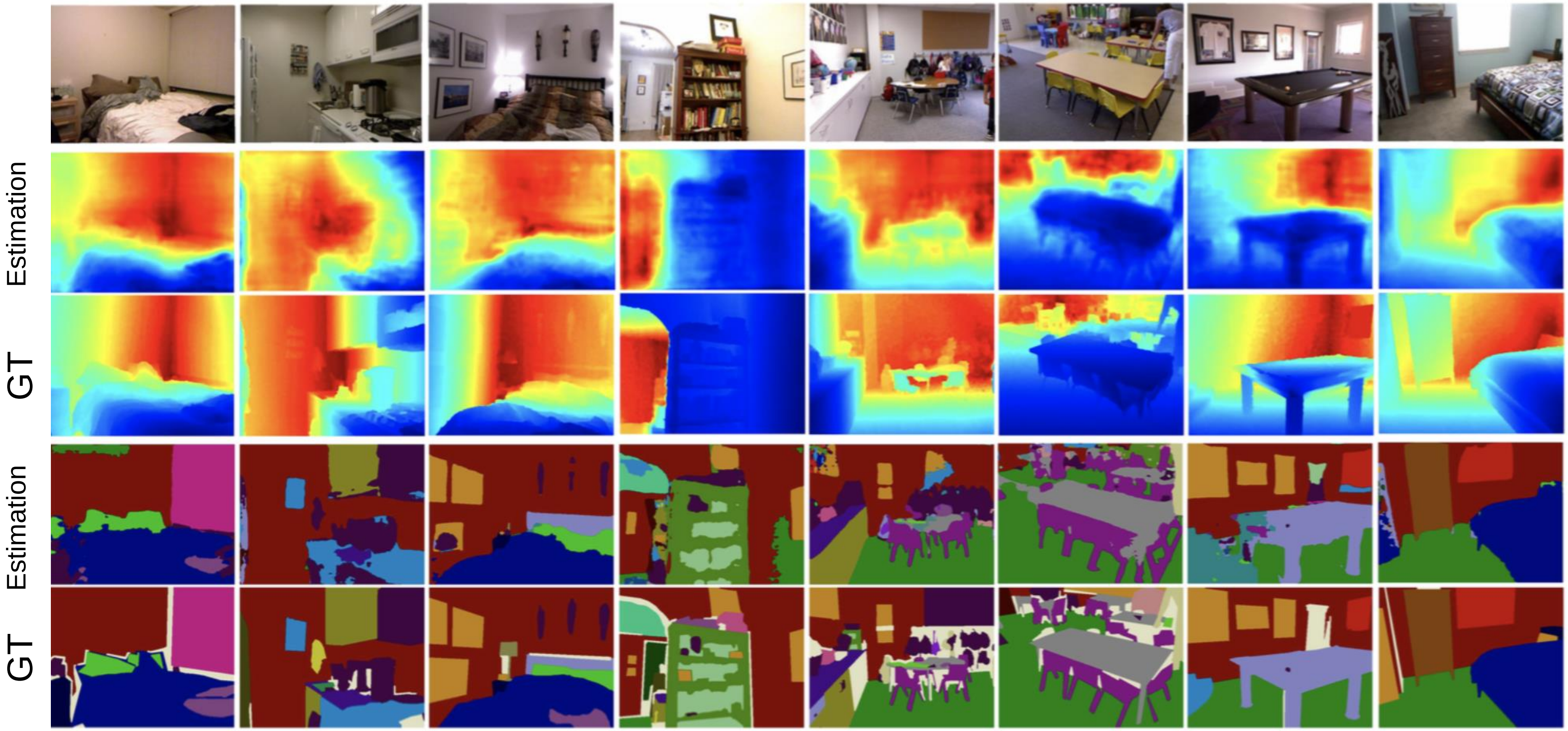
Table 1. Diagnostic experiments for the depth estimation task on NYUD V2 dataset. Distillation A, B, C represents the proposed three multi-task distillation modules.

Method	Error (lower is better)			Accuracy (higher is better)		
	rel	log10	rms	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Front-end + DE (baseline)	0.265	0.120	0.945	0.447	0.745	0.897
Front-end + DE + SP (baseline)	0.260	0.117	0.930	0.467	0.760	0.905
PAD-Net (Distillation A + DE)	0.248	0.112	0.892	0.513	0.798	0.921
PAD-Net (Distillation B + DE)	0.230	0.099	0.850	0.591	0.854	0.953
PAD-Net (Distillation C + DE)	0.221	0.094	0.813	0.619	0.882	0.965
PAD-Net (Distillation C + DE + SP)	**0.214**	**0.091**	**0.792**	**0.643**	**0.902**	**0.977**

Table 2. Diagnostic experiments for the scene parsing task on the NYUD V2 dataset.

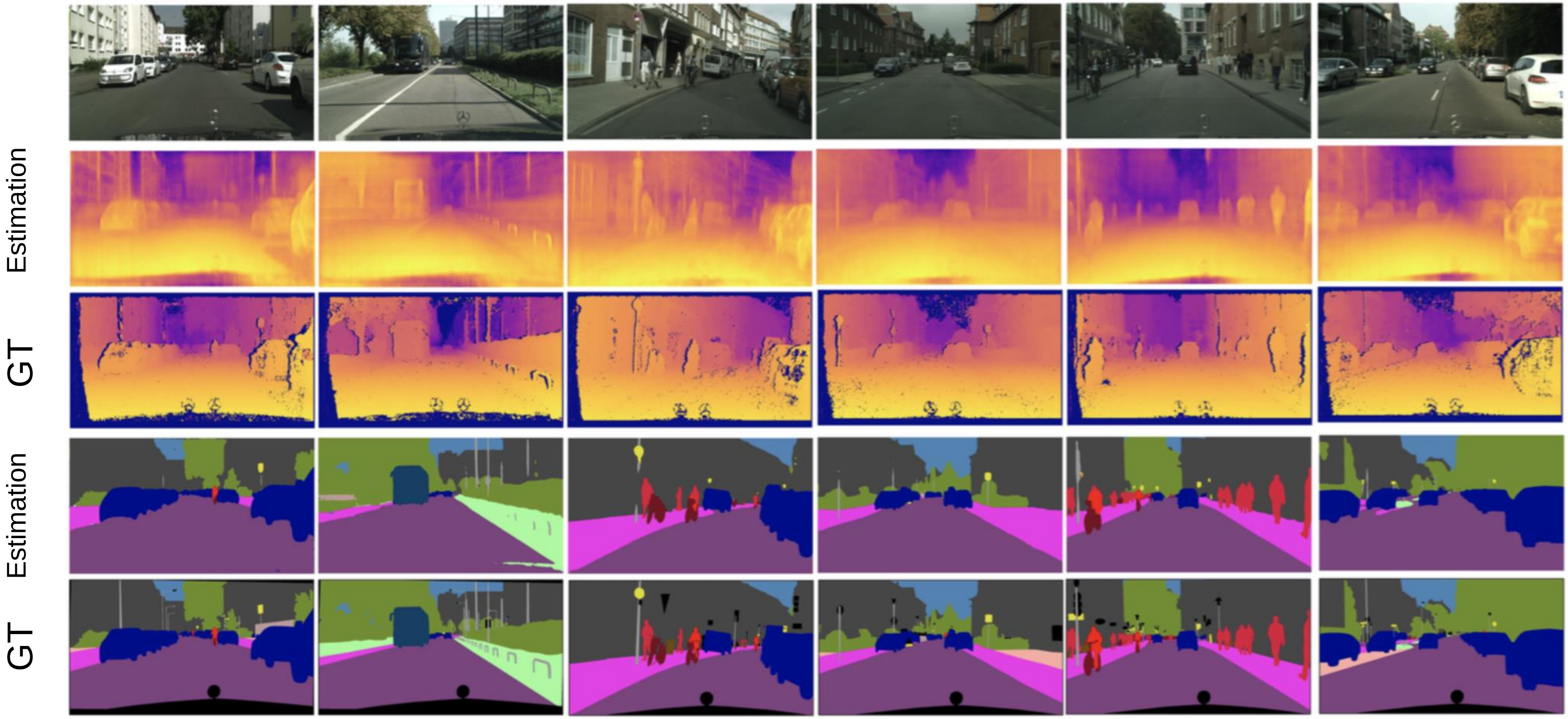
Method	Mean IoU	Mean Accuracy	Pixel Accuracy
Front-end + SP (baseline)	0.291	0.301	0.612
Front-end + SP + DE (baseline)	0.294	0.312	0.615
PAD-Net (Distillation A + SP)	0.308	0.365	0.628
PAD-Net (Distillation B + SP)	0.317	0.411	0.638
PAD-Net (Distillation C + SP)	0.325	0.432	0.645
PAD-Net (Distillation C + DE + SP)	**0.331**	**0.448**	**0.647**

Results



Qualitative results on NYUD-V2

Results



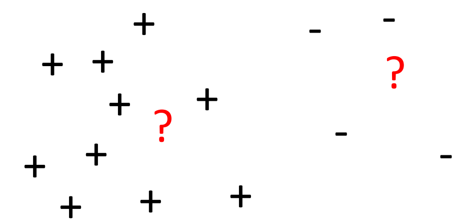
Qualitative results on Cityscapes

Effectively using high performance images

Feature fusion

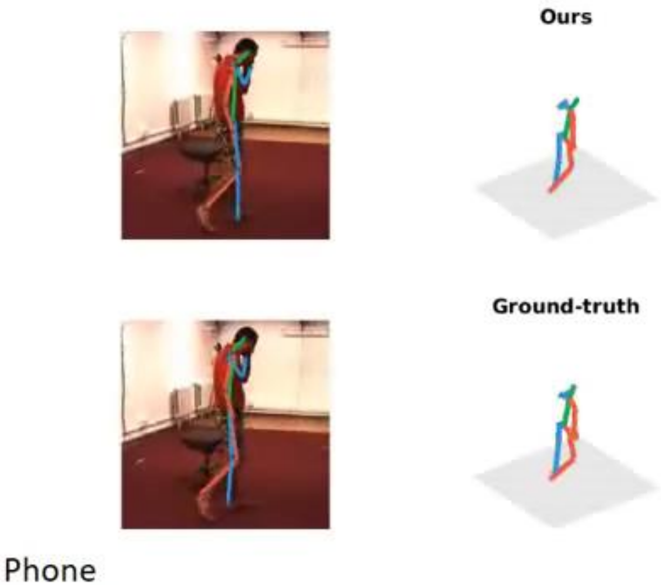
Structured features

Structured Samples



Challenges: No Annotation

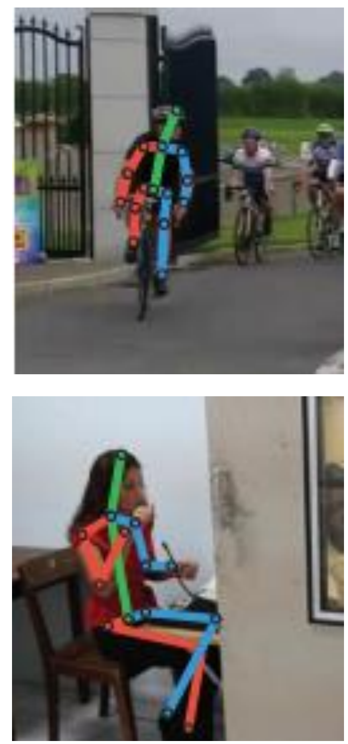
Constrained scenes



Domain

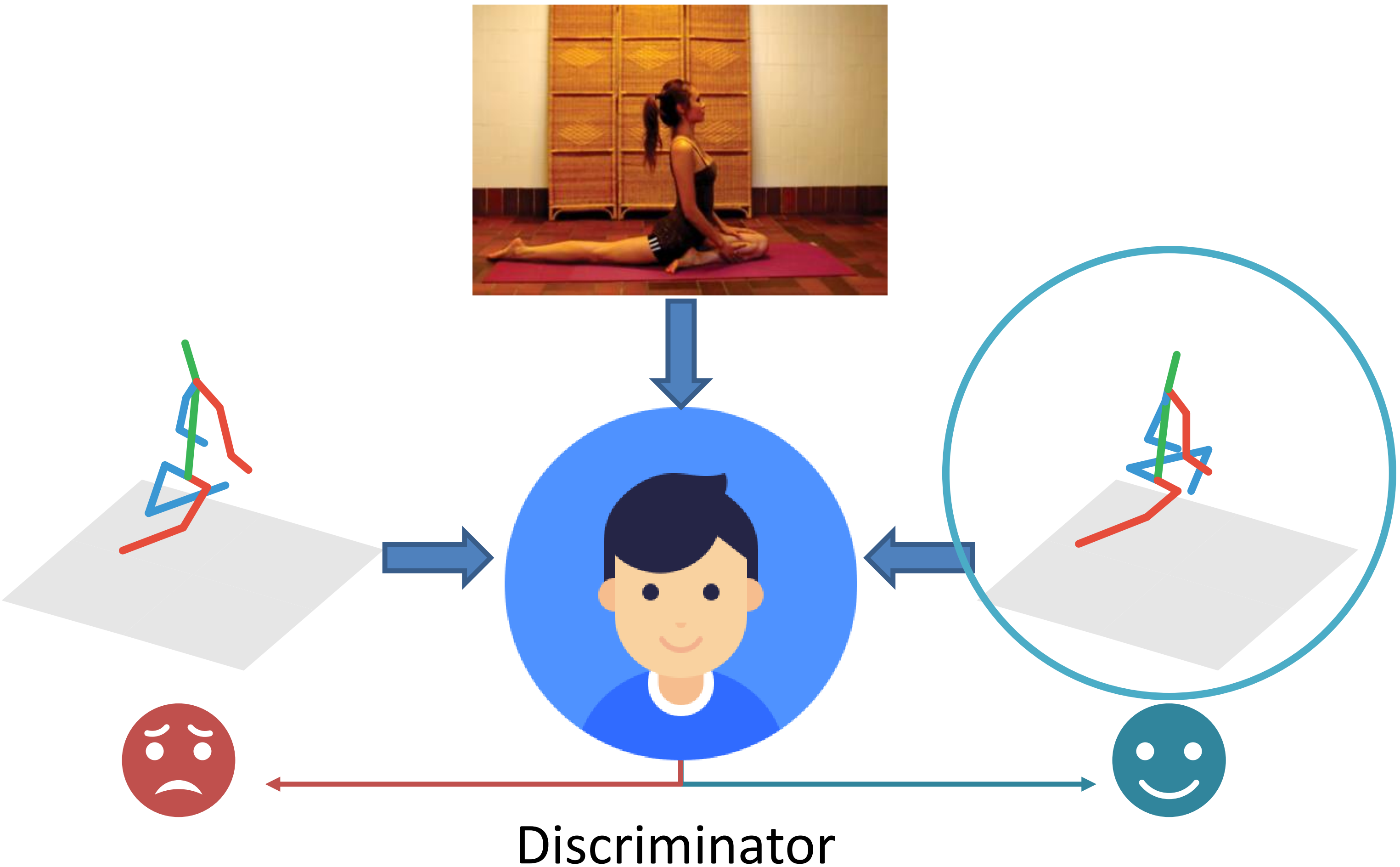
Discrepancy

In-the-wild scenes

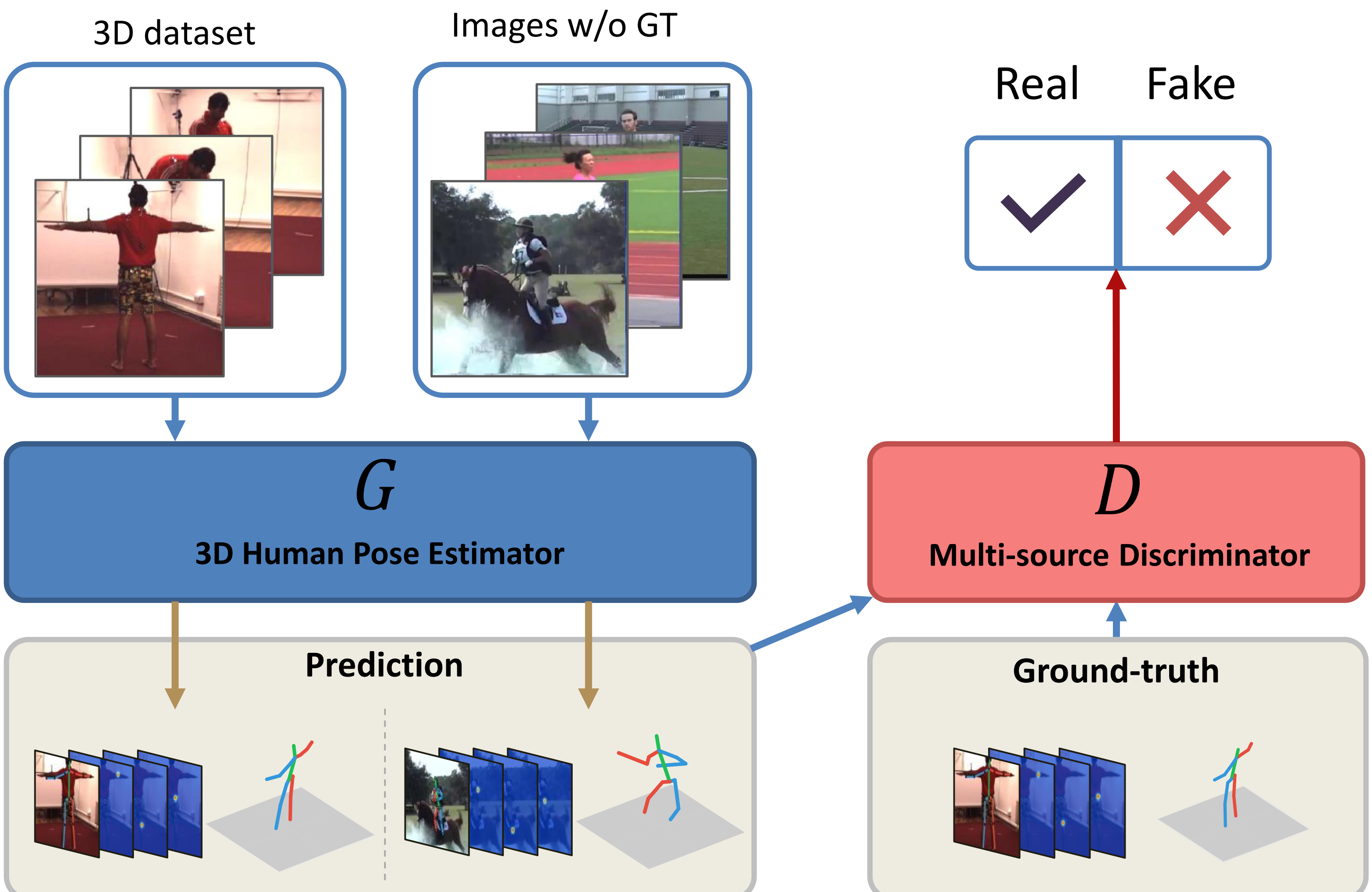


No annotation

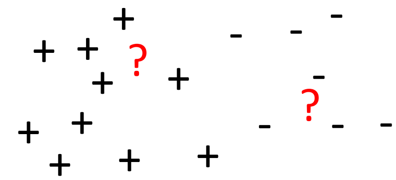
Which one is more plausible?

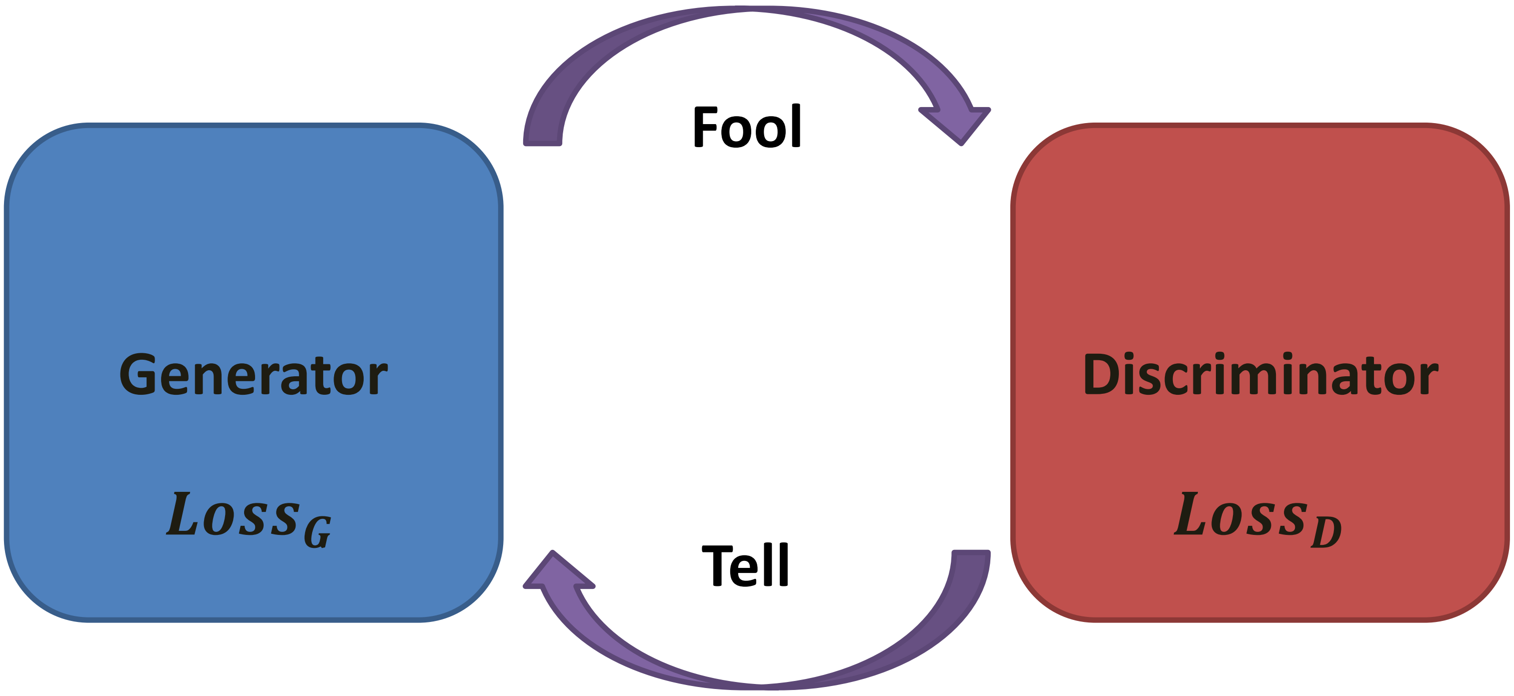


Weakly Supervised Adversarial Learning



Adversarial Learning

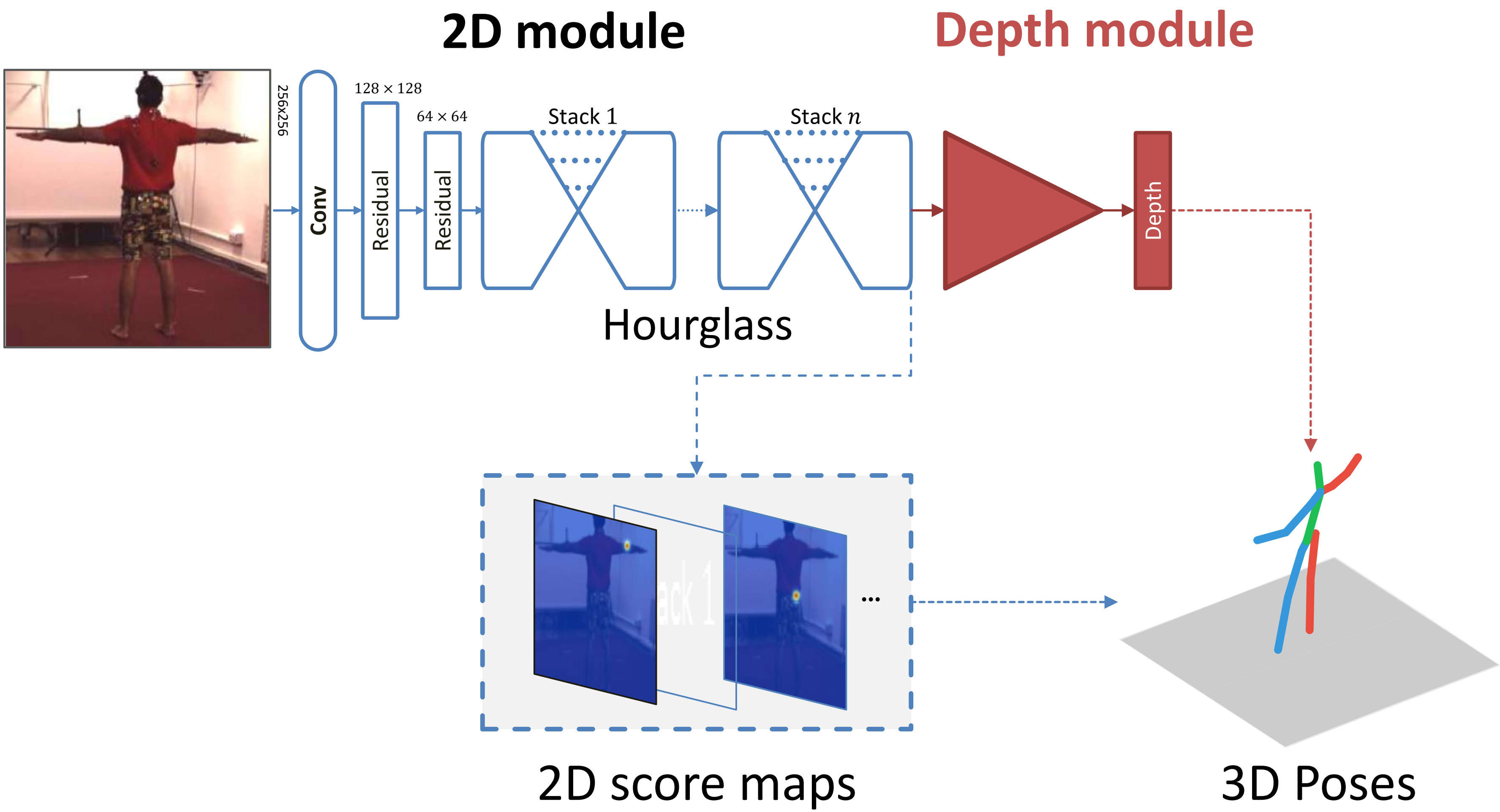




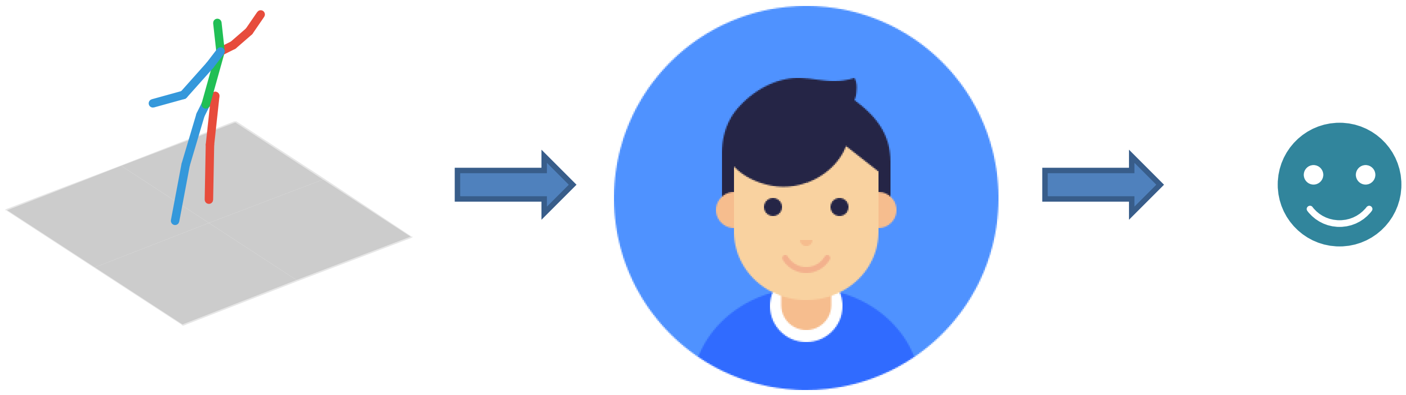
Euclidean Loss

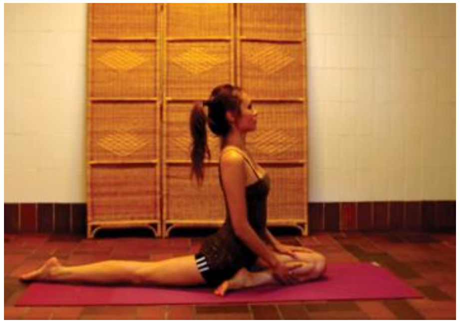
Classification Loss

Generator

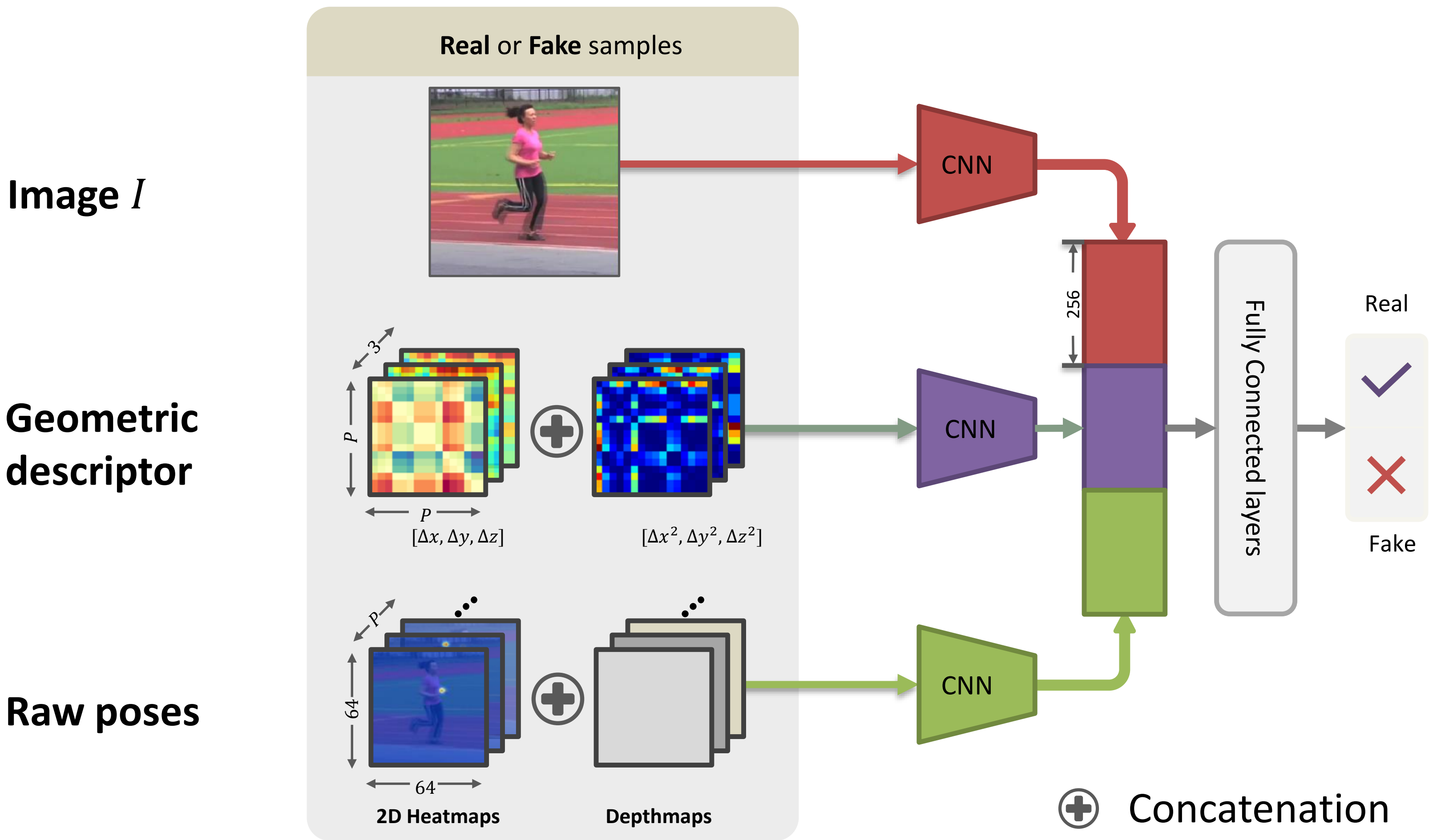

2D module
Depth module
256x256
Conv
128 × 128
Residual
64 × 64
Residual
Stack 1
Stack n
Hourglass
Depth
...
2D score maps
3D Poses

Discriminator

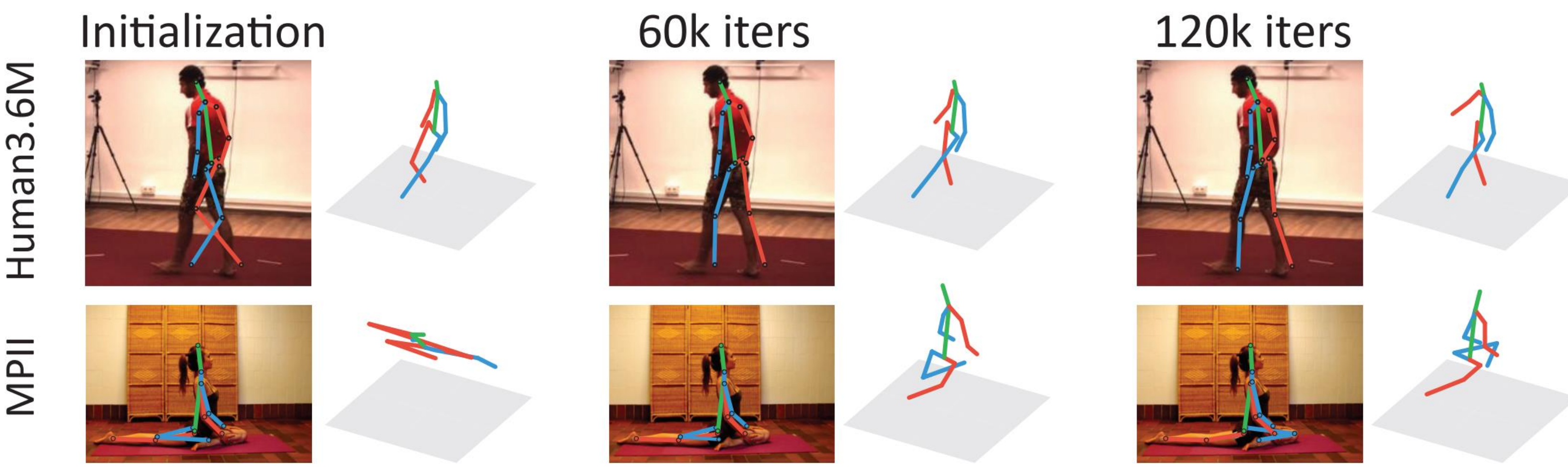




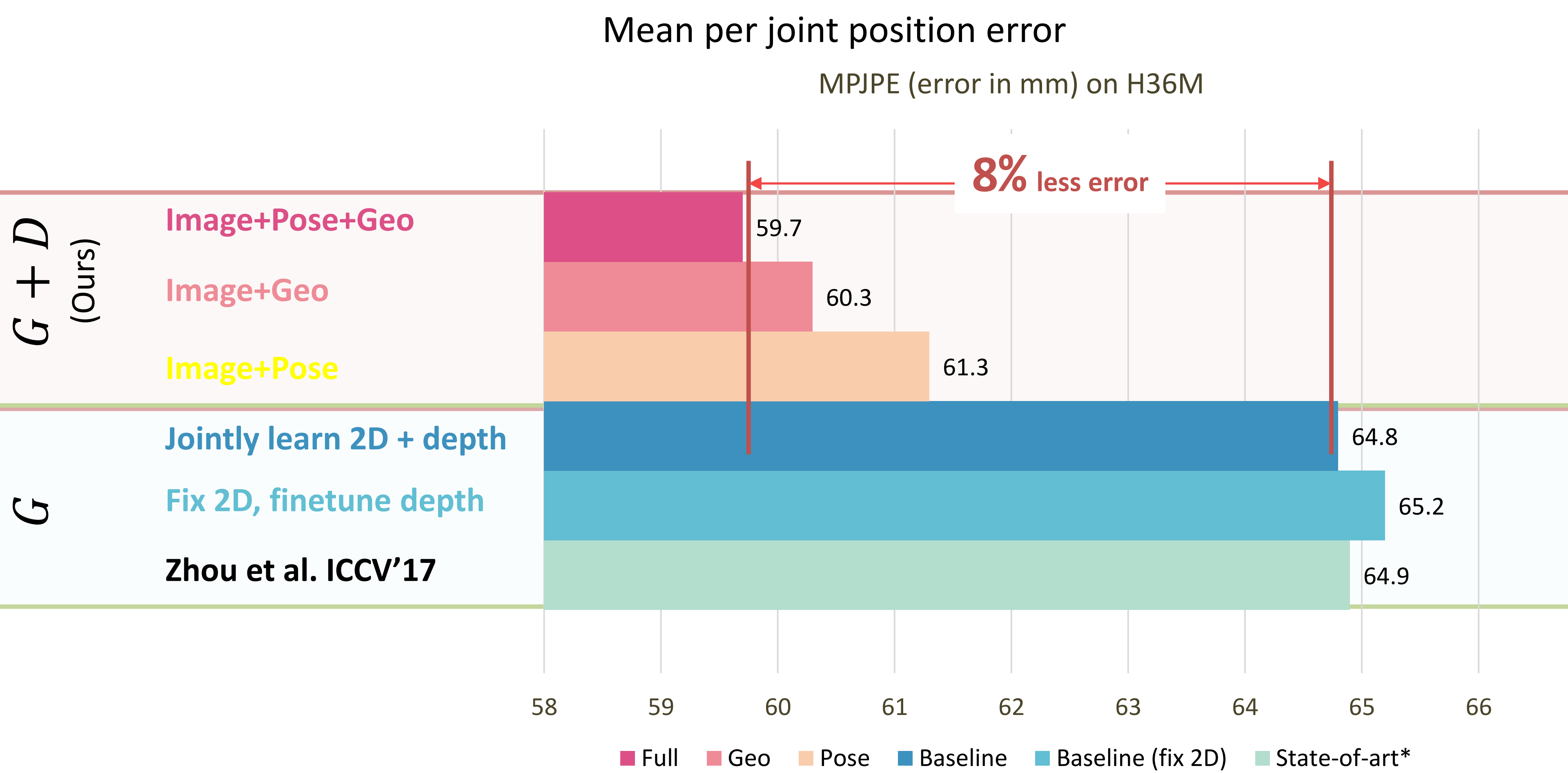
Multi-Source Discriminator



Effectiveness of Adversarial Learning



Ablation Study on H36M Dataset

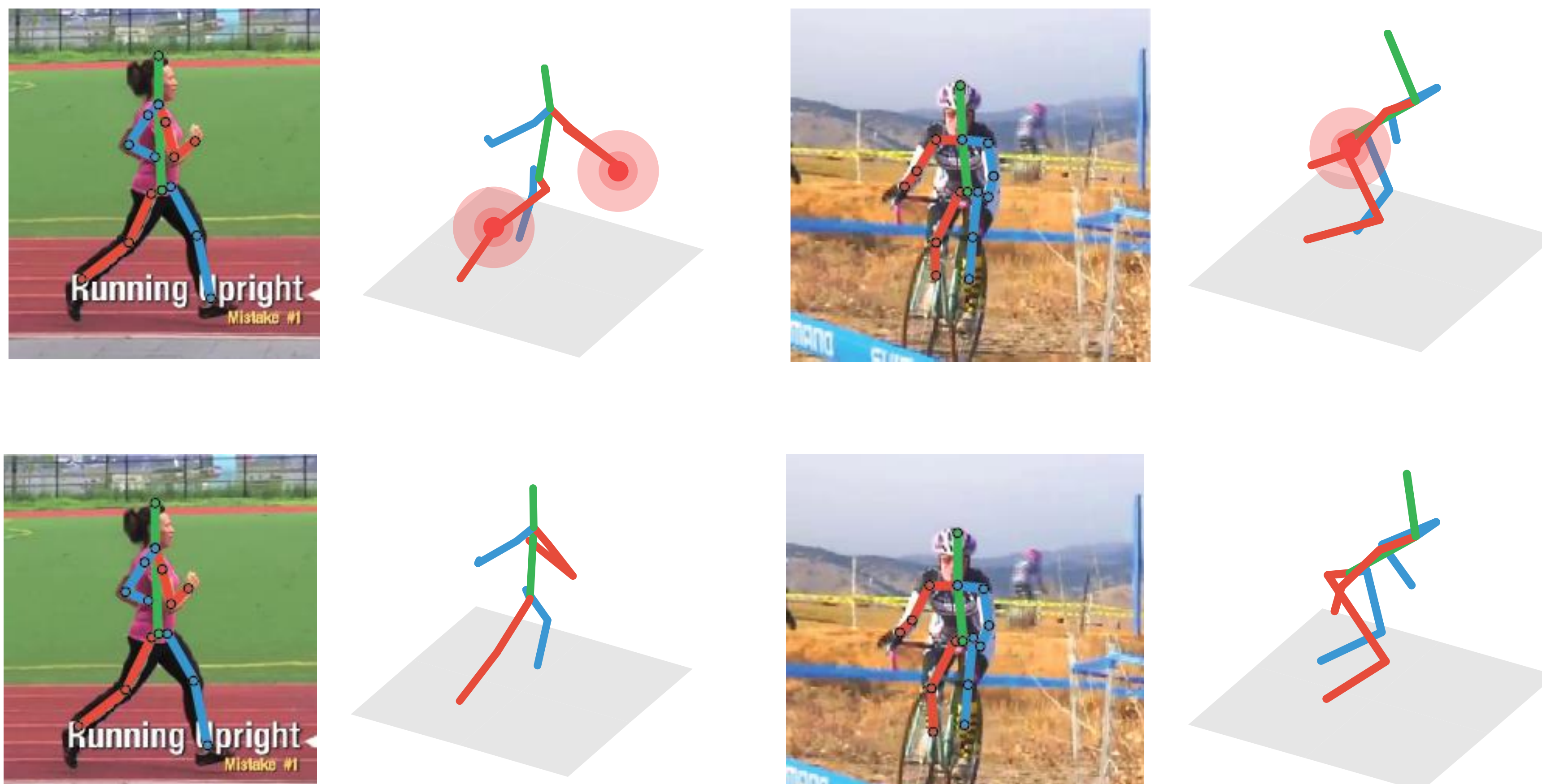


*Zhou et al. ICCV'17

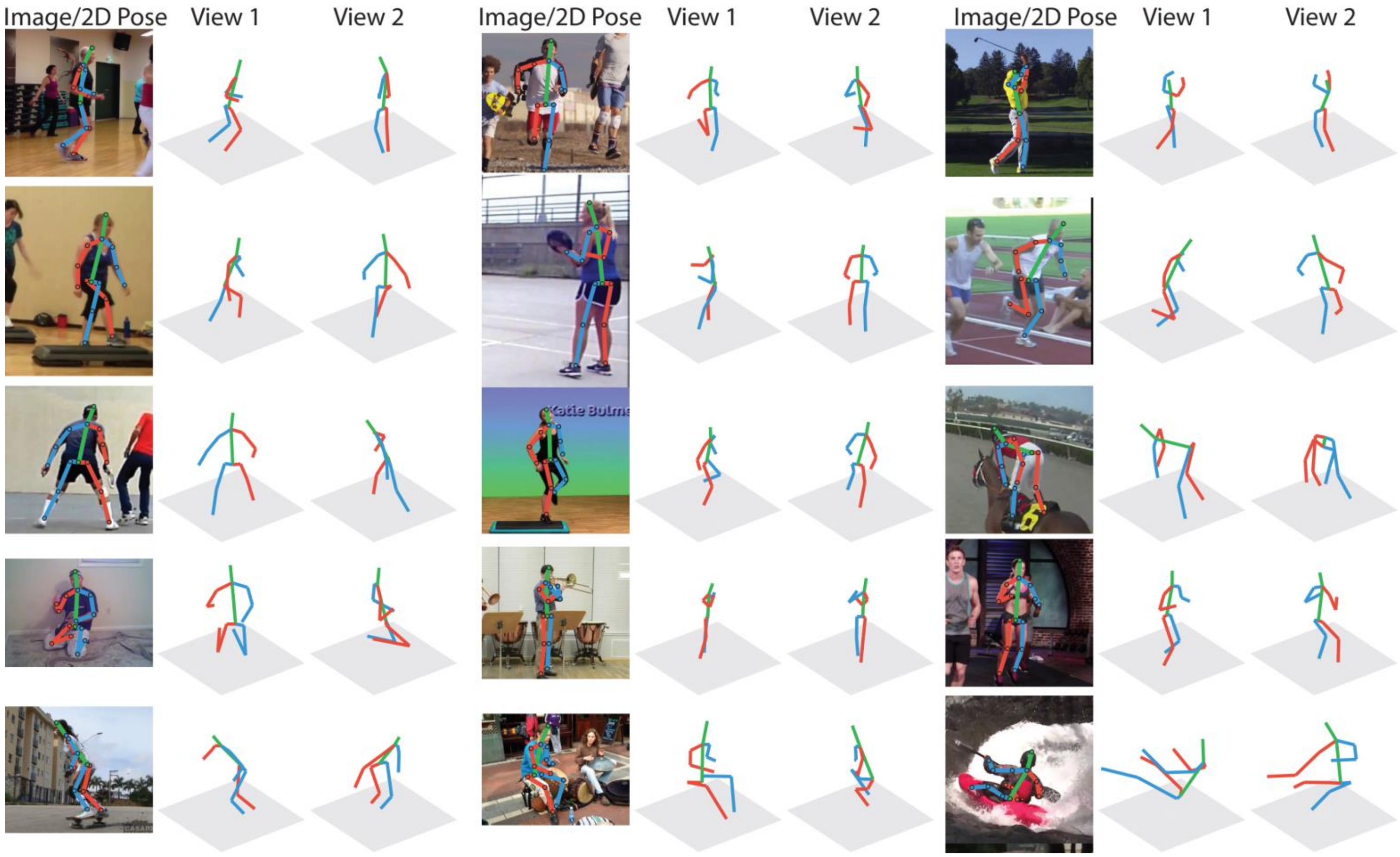
Results on Images in the Wild

baseline

Ours



Multi-view Results

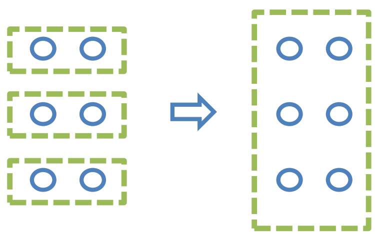


Outline

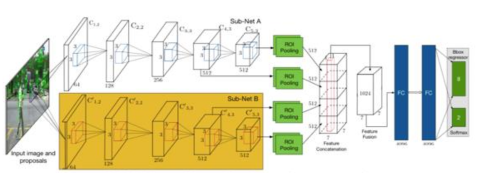
Introduction

Effectively using high performance images

Feature fusion

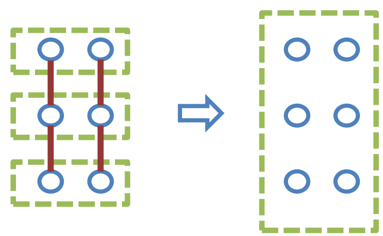


Pedestrian detection (CVPR'17)

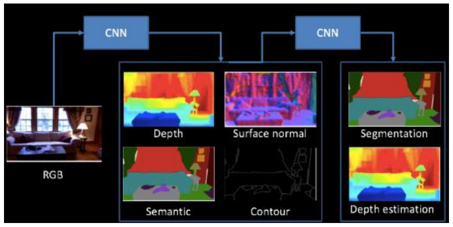


MSDN: multi-scale detection network

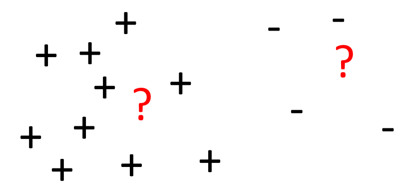
Structured Features



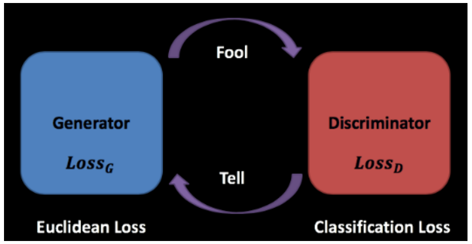
Scene parsing and depth estimation(CVPR18)



Structured Samples



3D human pose estimation (CVPR'18)



Conclusion

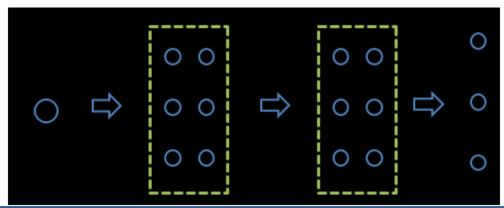
Does structure only exist for specific task?

Outline

Introduction

Effectively using high performance images

Back-bone model design



FishNet
(NIPS 18)

Optical guided feature
(CVPR18)

Low-level and high-level features

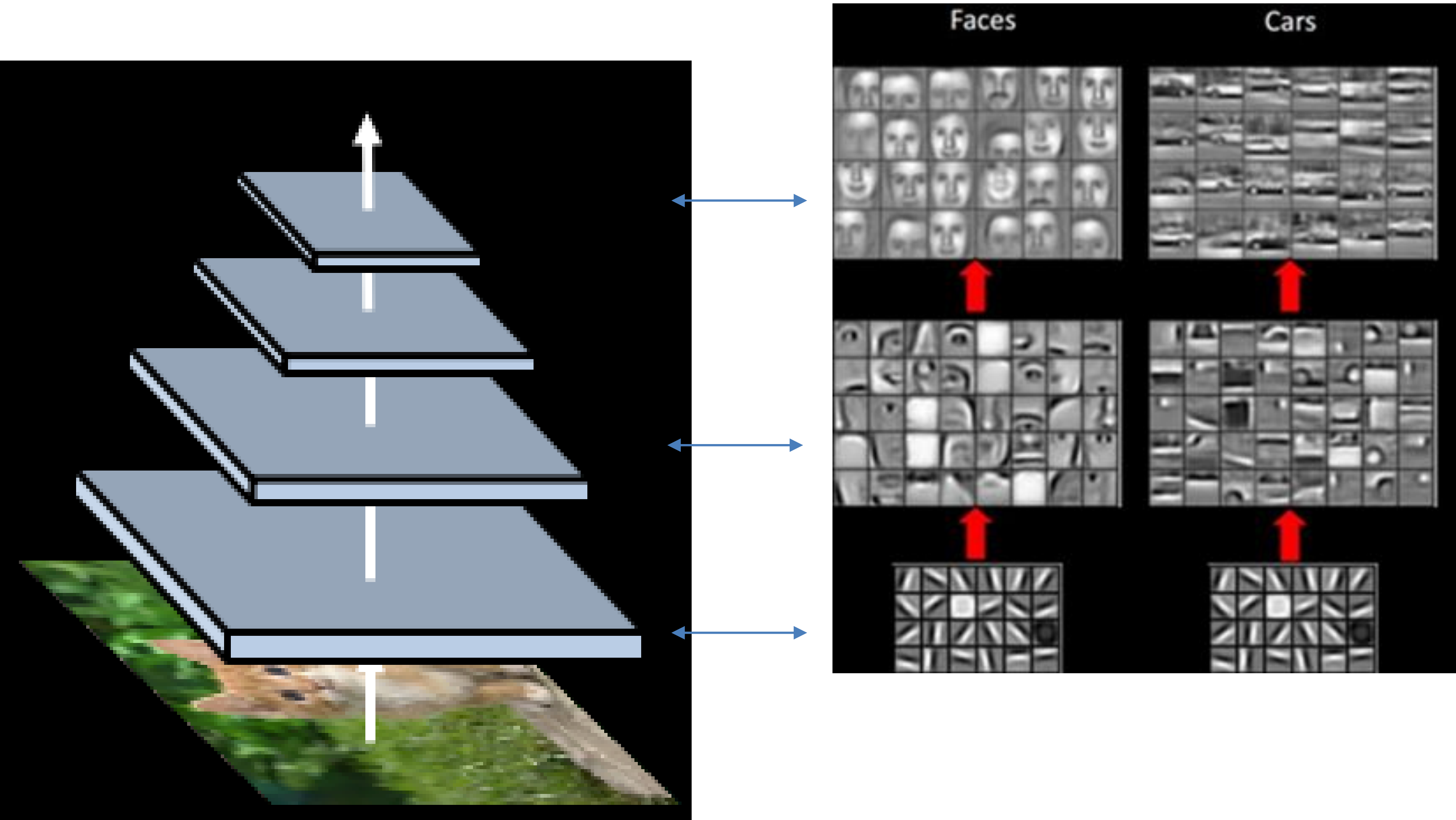


Image from Andrew Ng’s slides

Current CNN Structures

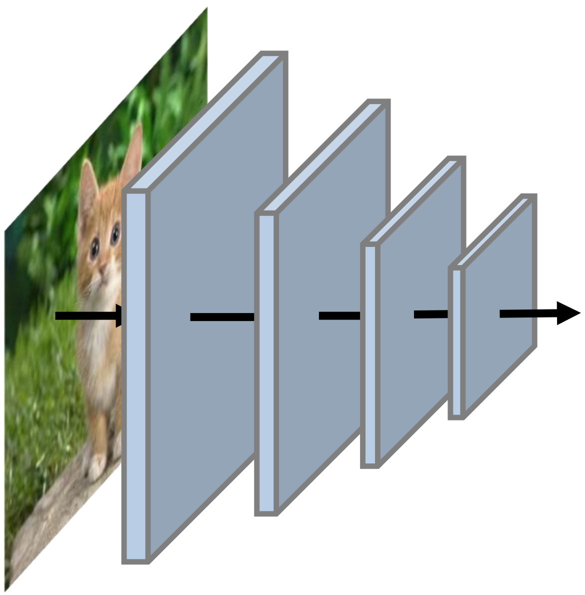
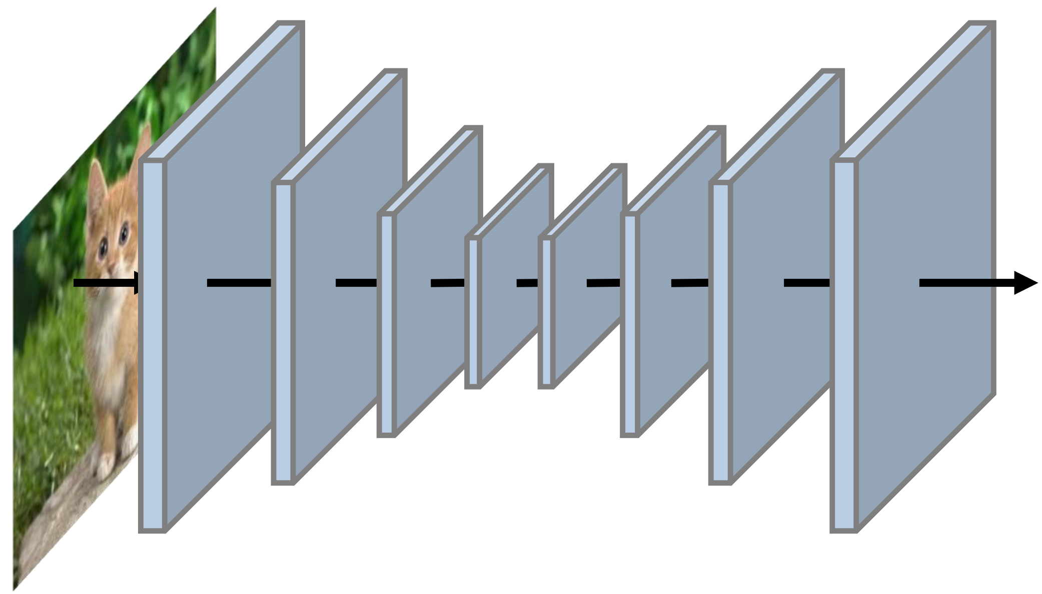


Image Classification:
Summarize high-level semantic information of the whole image.

Called U-Net, Hourglass, or Conv-deconv



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

Architectures designed for different granularities are **DIVERGING**

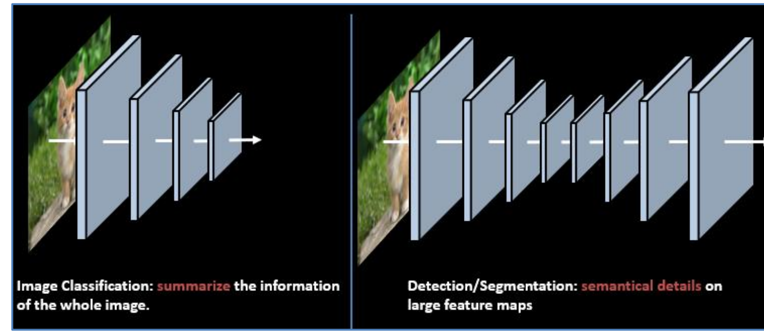
Unify the advantages of networks for pixel-level, region-level, and image-level tasks

Hourglass for Classification

Features with high-level semantics and high resolution is good

Directly applying hourglass for classification?

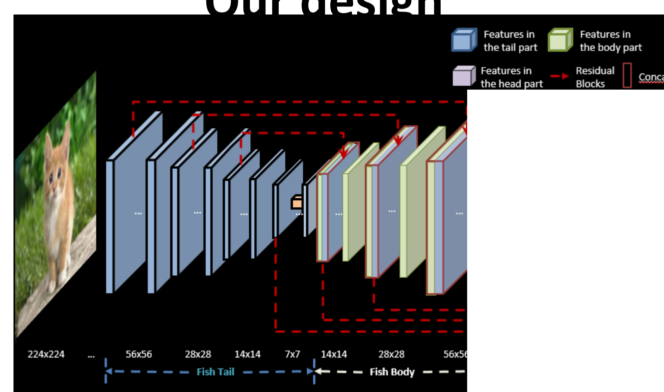
Poor performance.



So what is the **problem**?

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

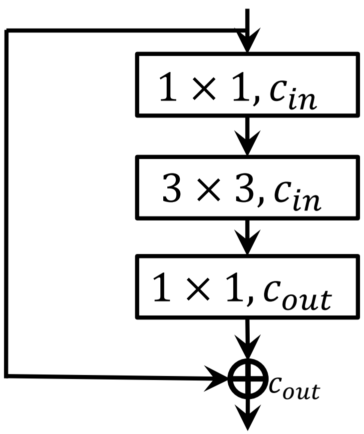
Our design



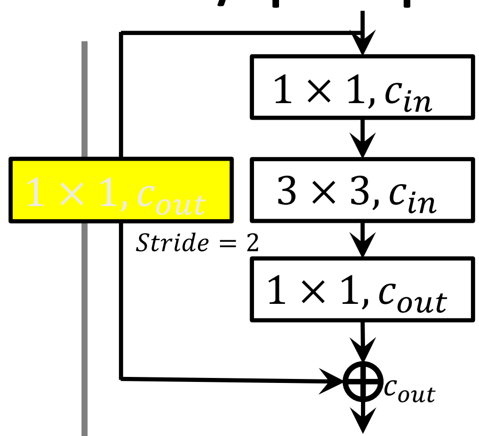
Hourglass for Classification

Ⓒ Concat

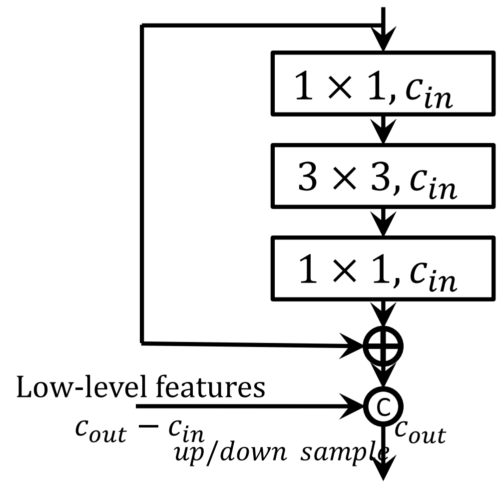
Normal Res-Block



Res-Block for down/up sampling



Our design



- **Different tasks require different resolutions of feature**
- **Hourglass may bring more isolated convolutions than ResNet**

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

Observation and design

Our observation

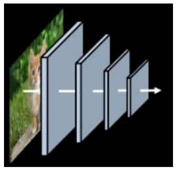
1. Diverged structures for tasks requiring different resolutions.
2. Isolated Conv blocks the direct back-propagation
3. Features with different depths are not fully explored, or mixed but not preserved

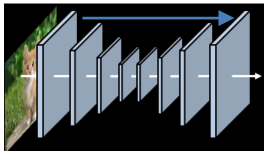
Solution

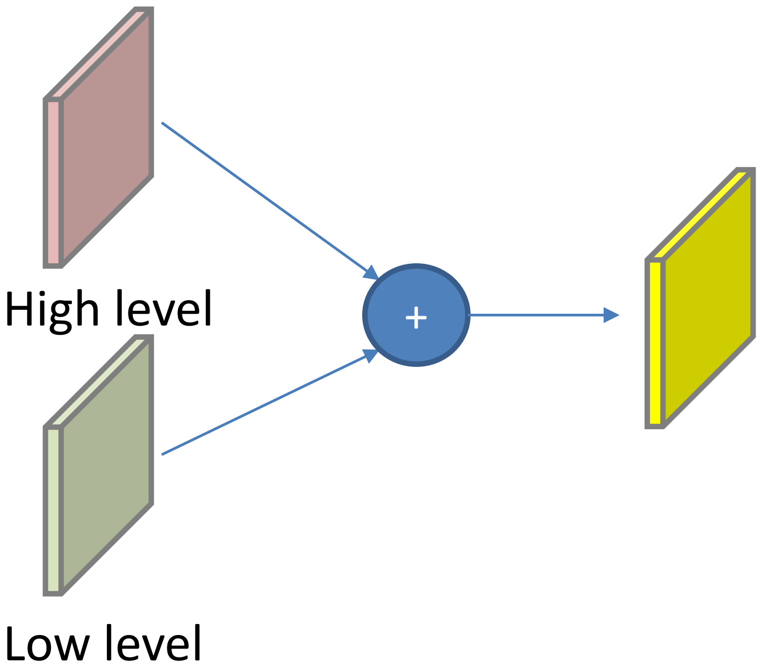
1. Unify the advantages of networks for pixel-level, region-level, and image-level tasks.
2. Design a network that does not need isolated convolution
3. Features from varying depths are **preserved and refined** from each other.

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.

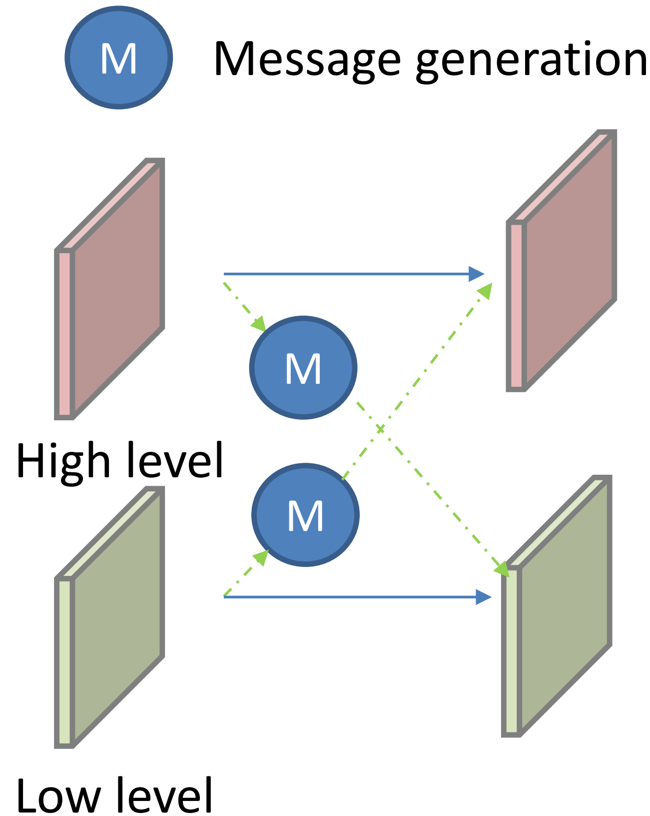
Difference between mix and preserve and refine





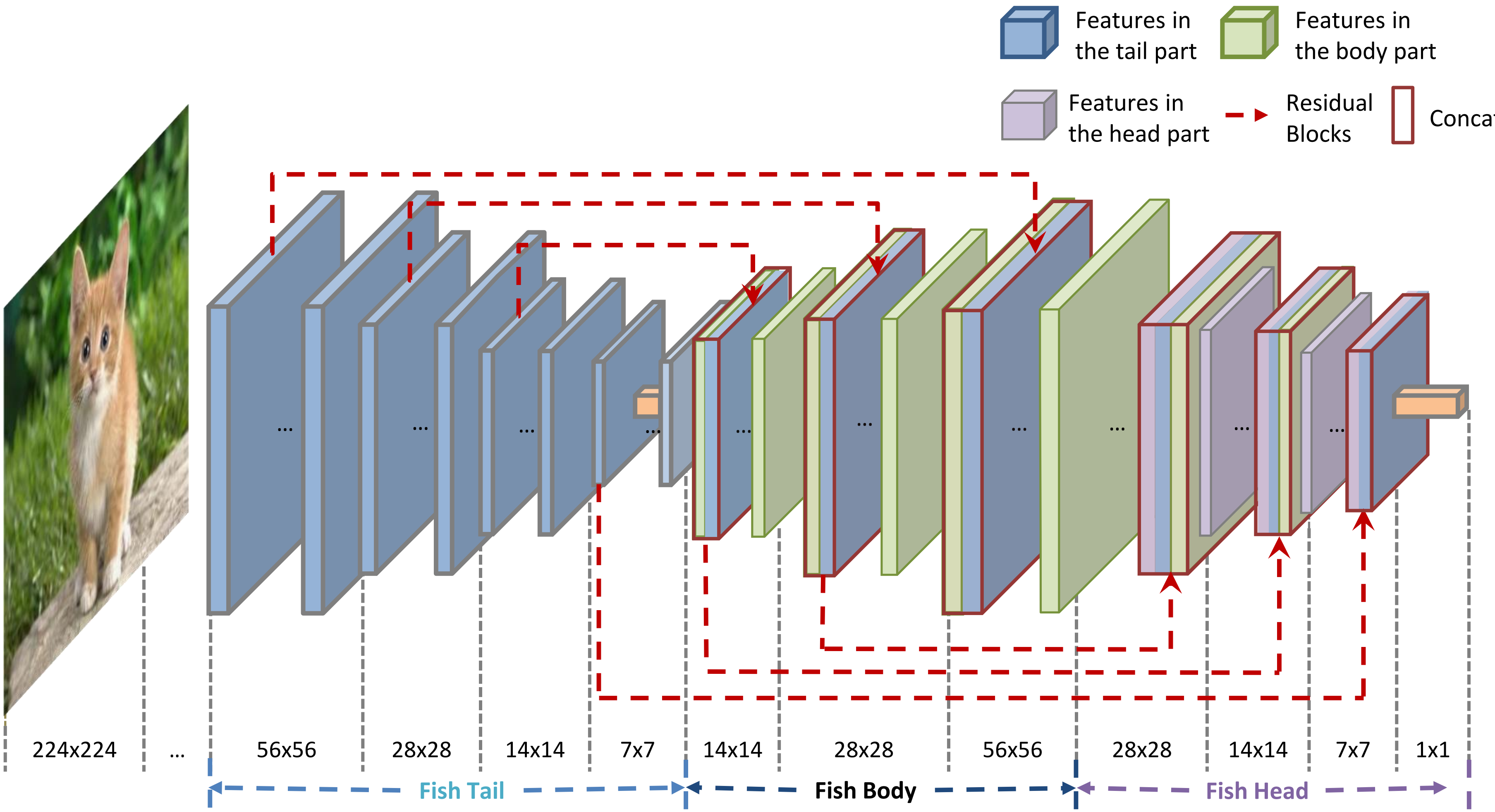


Mixed features

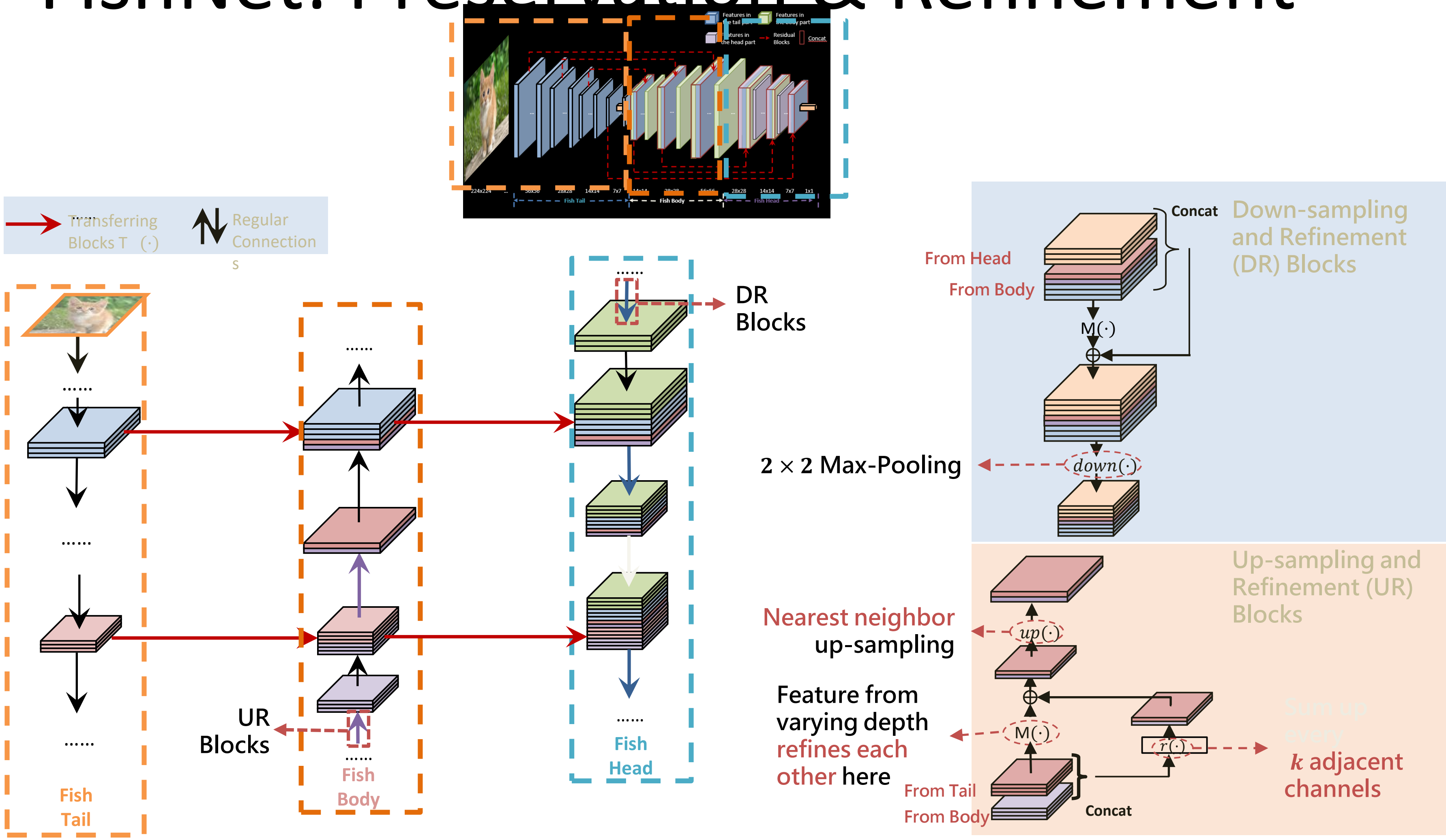


Preserve and refine

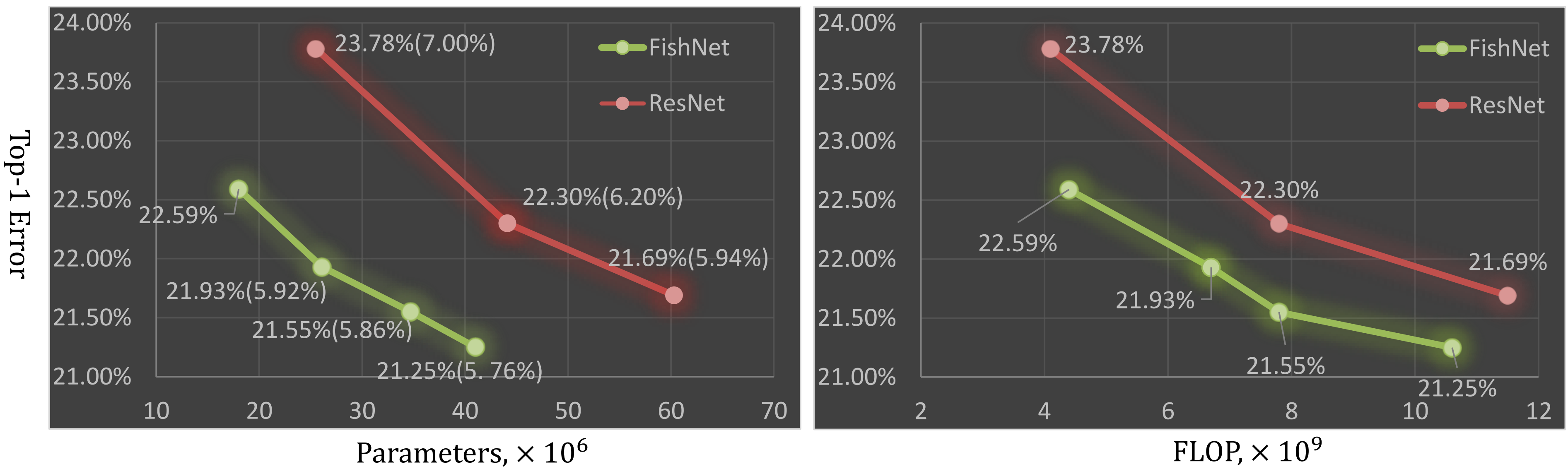
FishNet: Overview



FishNet: Preservation & Refinement

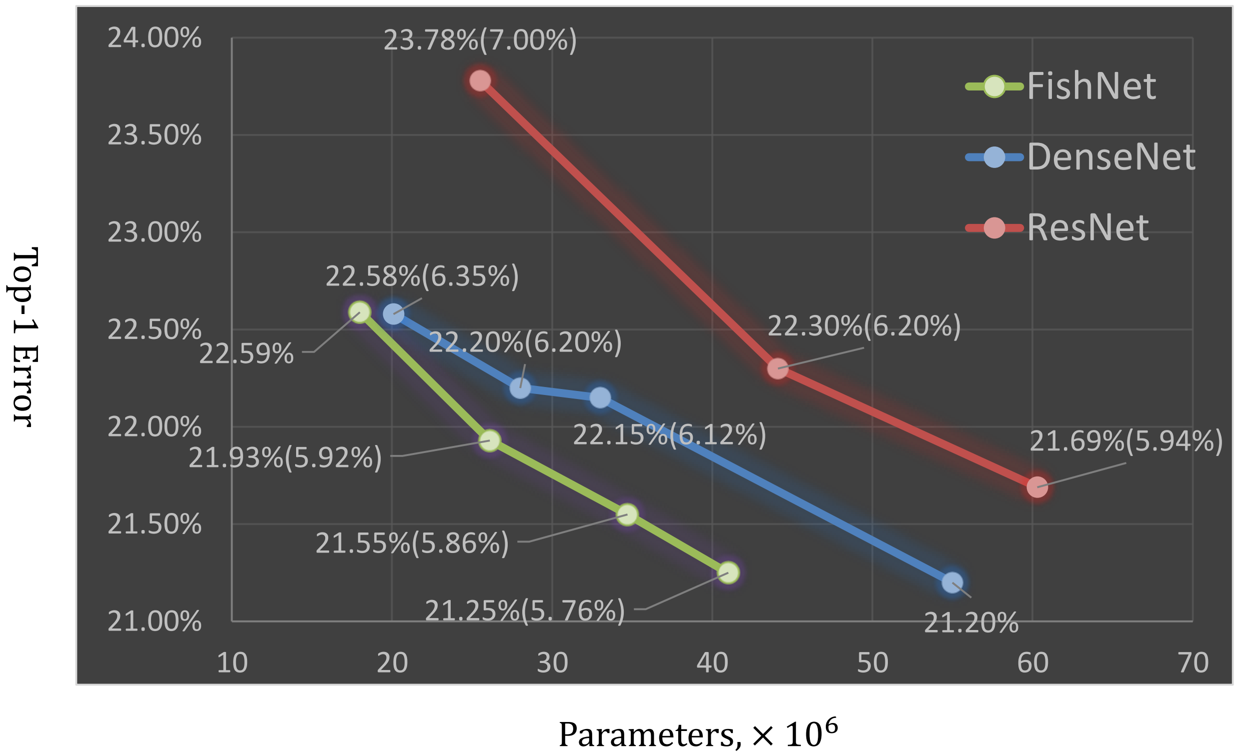


FishNet: Performance-ImageNet



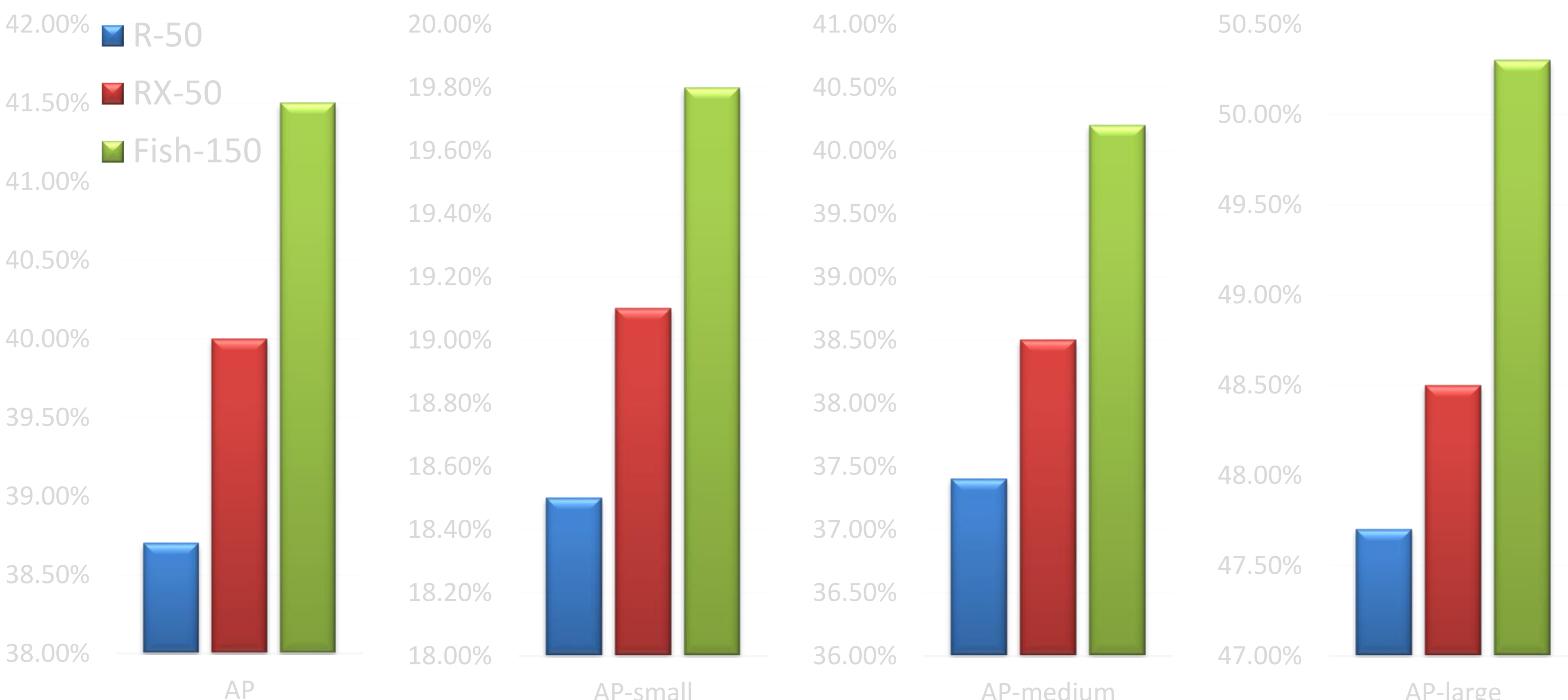
Code
https://github.com/kevin-ssy/FishNet

FishNet: Performance-ImageNet



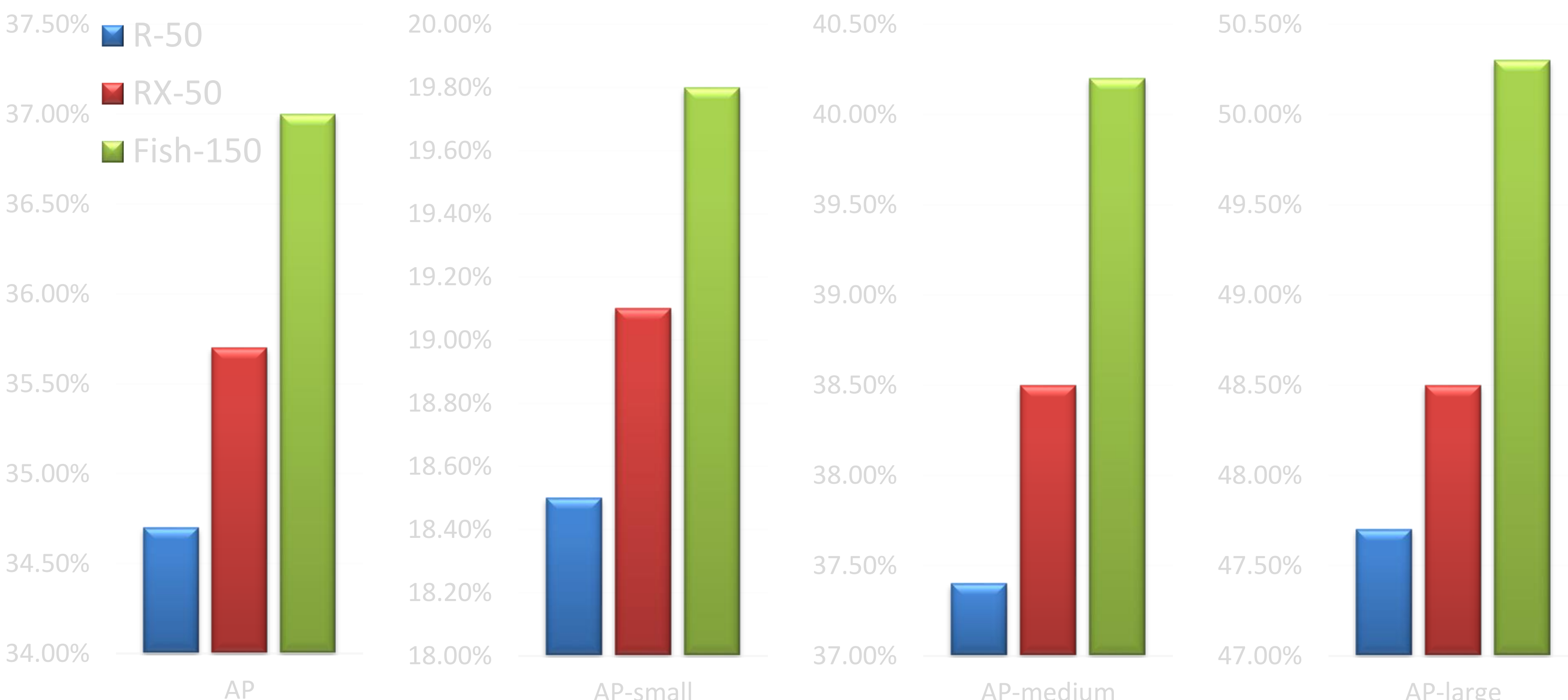
Code
https://github.com/kevin-ssy/FishNet

FishNet: Performance on COCO Detection



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

FishNet: Performance on COCO Instance Segmentation

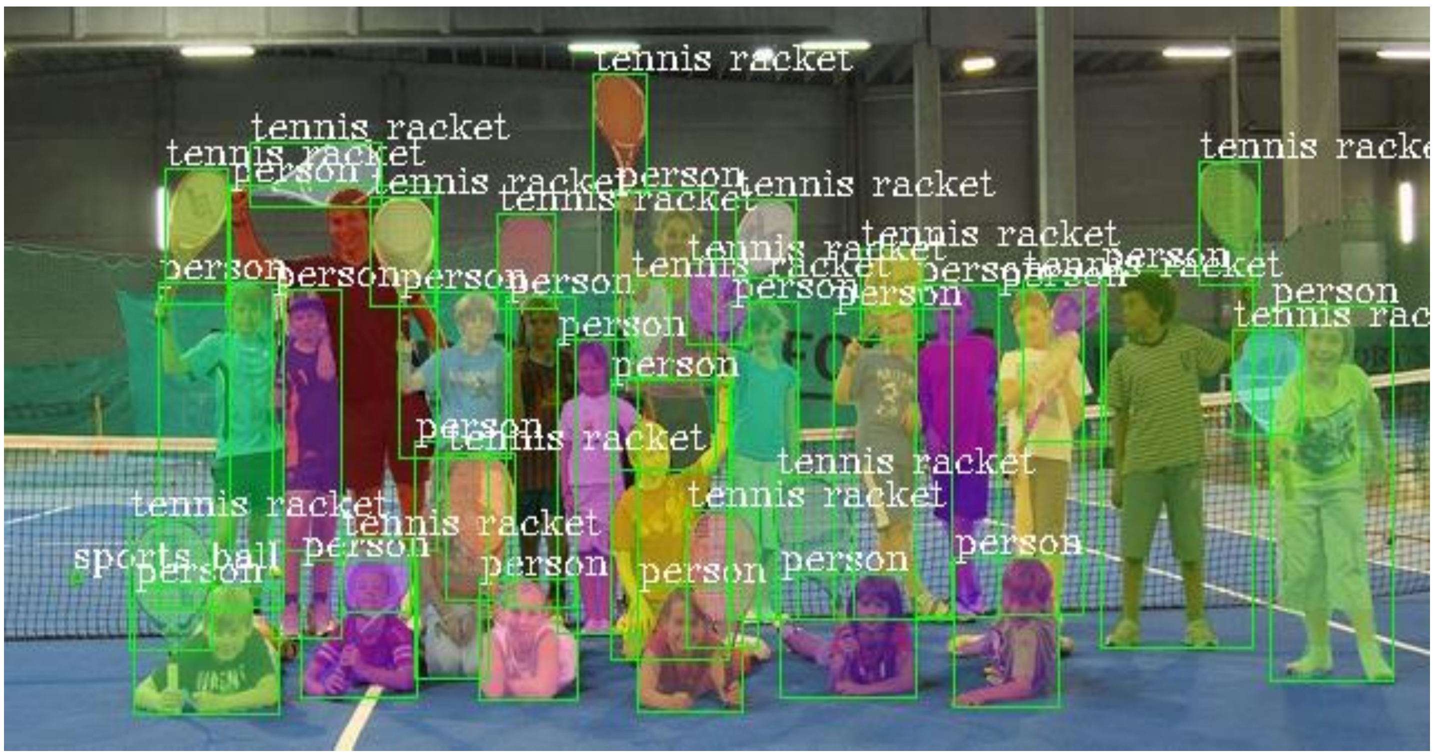


Code
https://github.com/kevin-ssy/FishNet

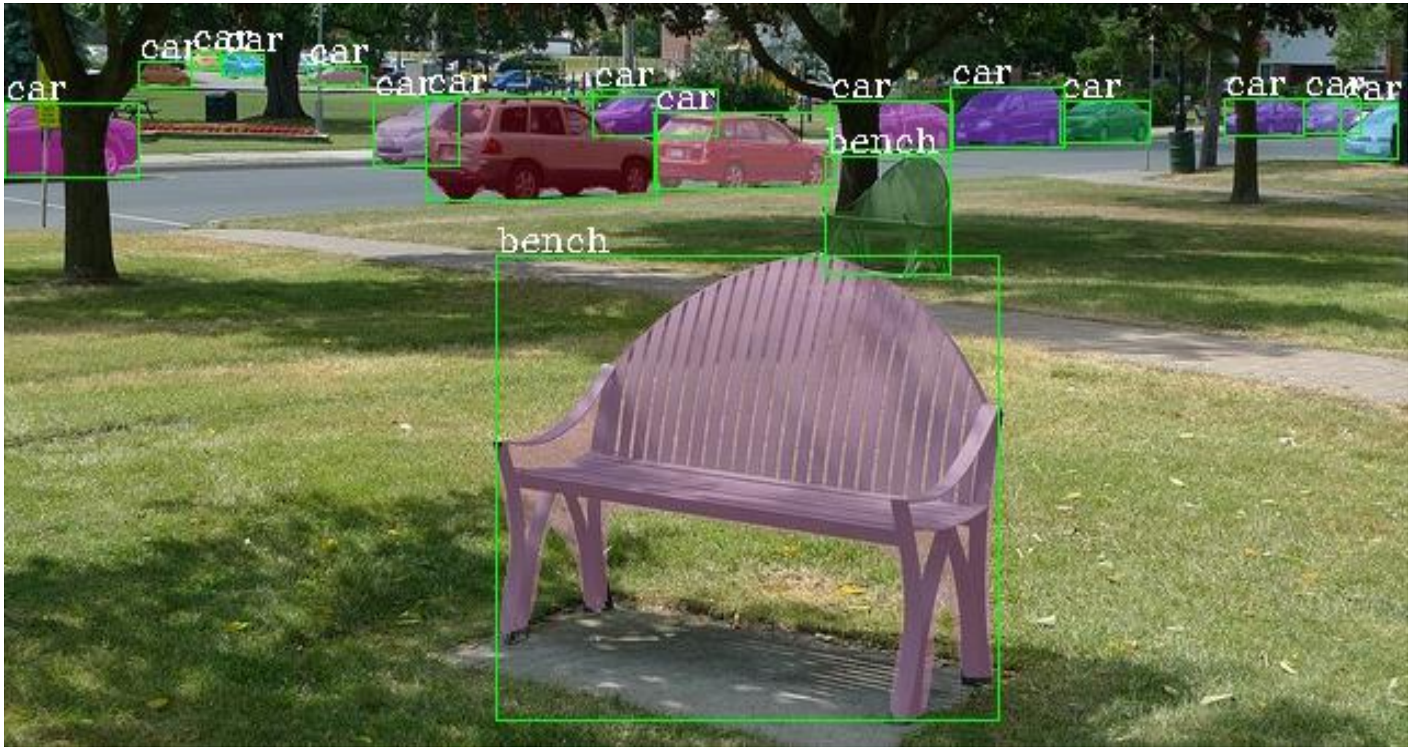
Winning COCO 2018 Instance Segmentation Task

	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AP^{L}	AR^{1}	AR^{10}	AR^{100}	AR^{S}	AR^{M}	AR^{L}	date
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

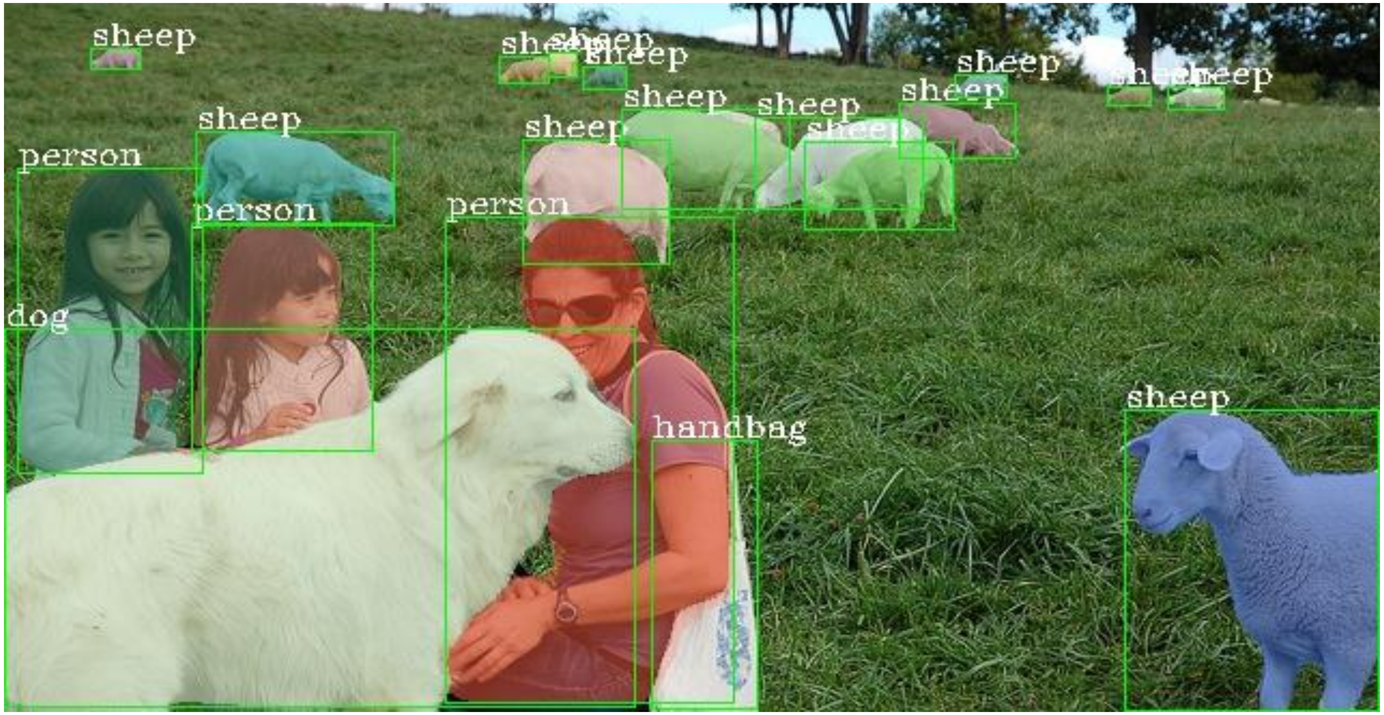
Visualization



Visualization

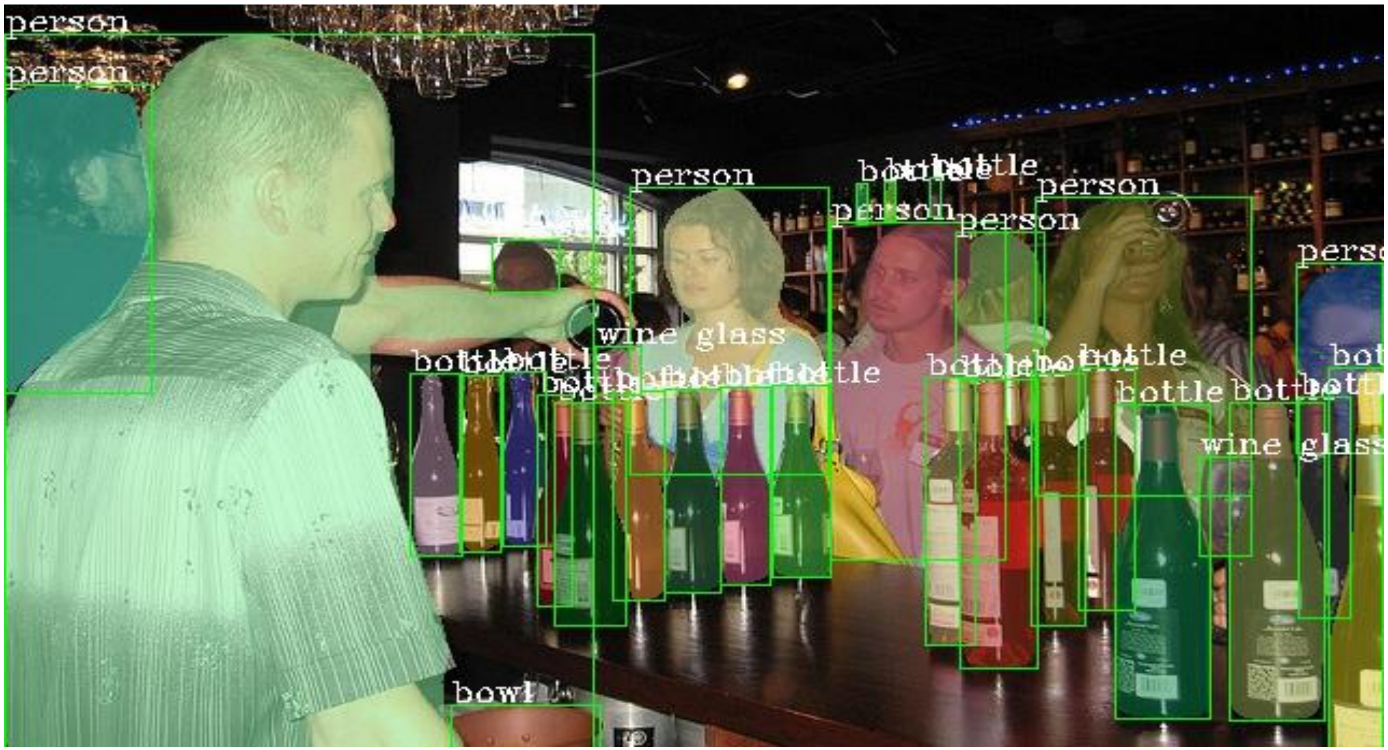


Visualization



Visualization

Visualization



Codebase

- Comprehensive
  - [x] RPN
  - [x] Mask R-CNN
  - [x] Cascade R-CNN
  - [ ] More … …
  - [x] Fast/Faster R-CNN
  - [x] FPN
  - [x] RetinaNet

- **High performance**
  - [x] Better performance
  - [x] Optimized memory consumption
  - [x] Faster speed

- **Handy to develop**
  - [x] Written with PyTorch
  - [x] Modular design

GitHub: mmdet

FishNet: Advantages

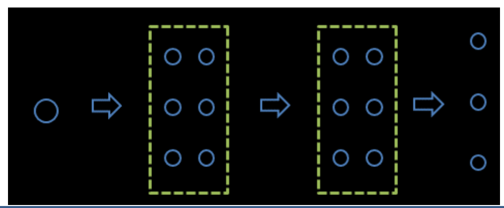
1. Better gradient flow to shallow layers
2. High-resolution features contain rich low-level and high-level semantics
3. Feature from varying depth are preserved and refined from each other

Outline

Introduction

Effectively using high performance images

Back-bone model design



FishNet
(NIPS 18)

Optical guided feature
(CVPR18)

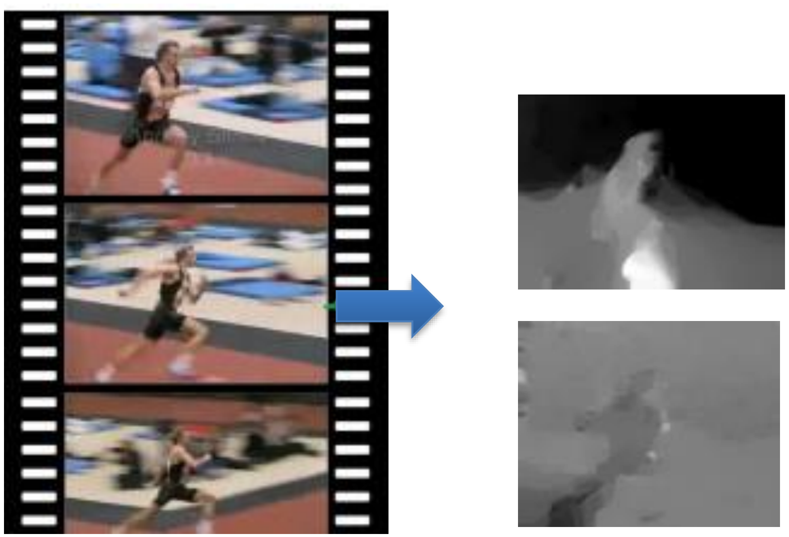
Action Recognition

- Recognize action from videos

Optical flow in Action Recognition

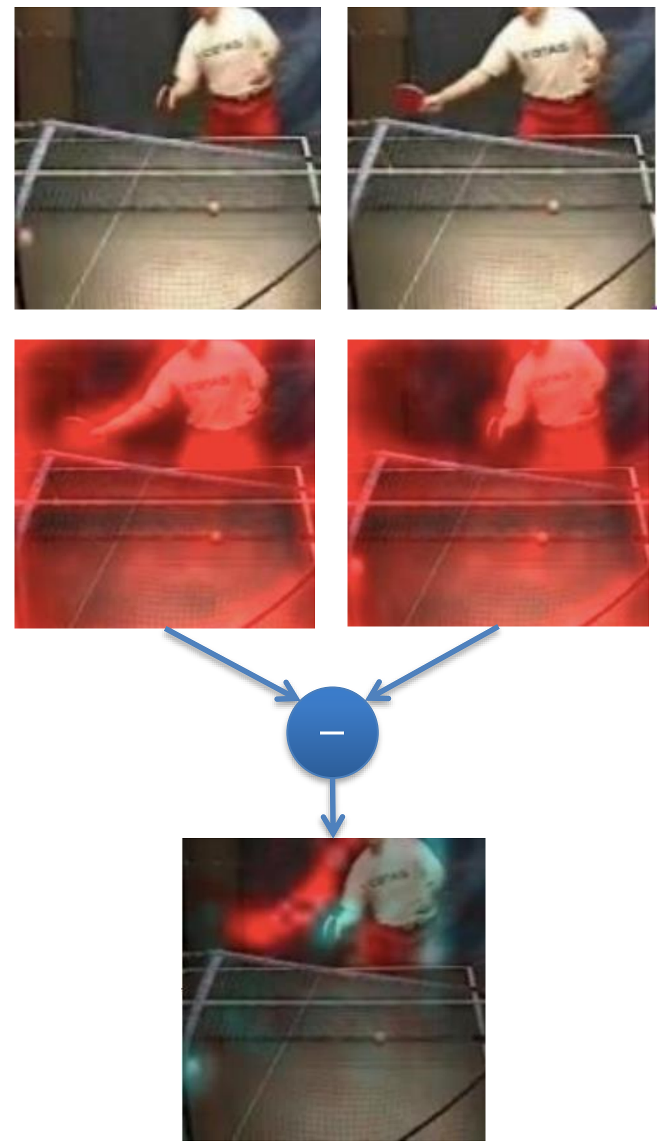
- Motion is the important information
- Optical flow
  - Effective
  - Time consuming

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:

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

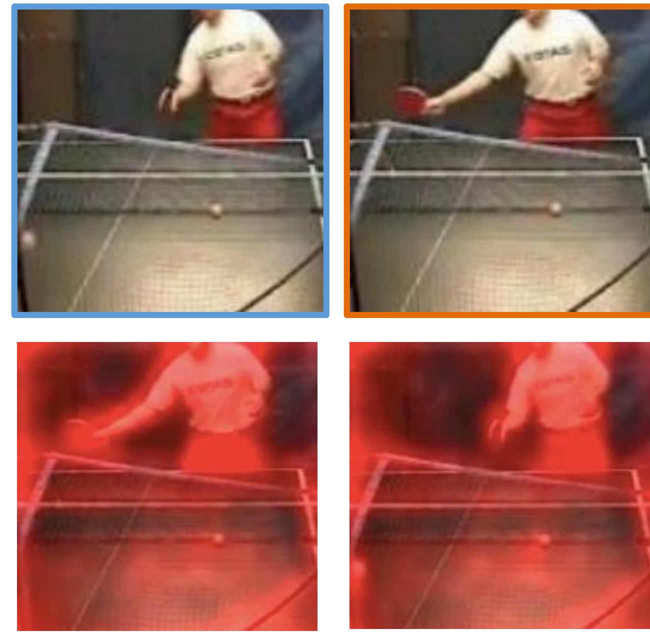
$$\frac{\partial I(x,y,t)}{\partial x} v_x + \frac{\partial I(x,y,t)}{\partial y} v_y + \frac{\partial I(x,y,t)}{\partial t} = 0$$

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

Intuitive Inspiration

Coefficient for optical flow:

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



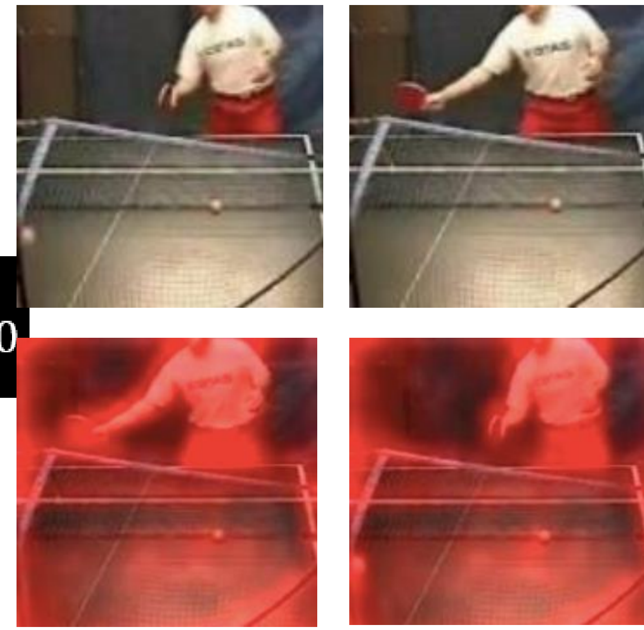
Optical flow guided feature

Feature flow:

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

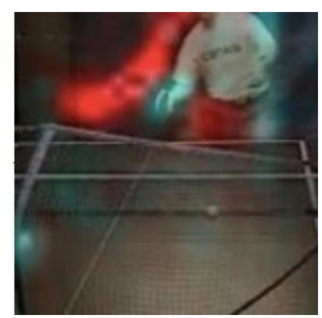
$$\frac{\partial f(I(x,y,t))}{\partial x}\tilde{v}_x+\frac{\partial f(I(x,y,t))}{\partial y}\tilde{v}_y+\frac{\partial f(I(x,y,t))}{\partial t}=0$$

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

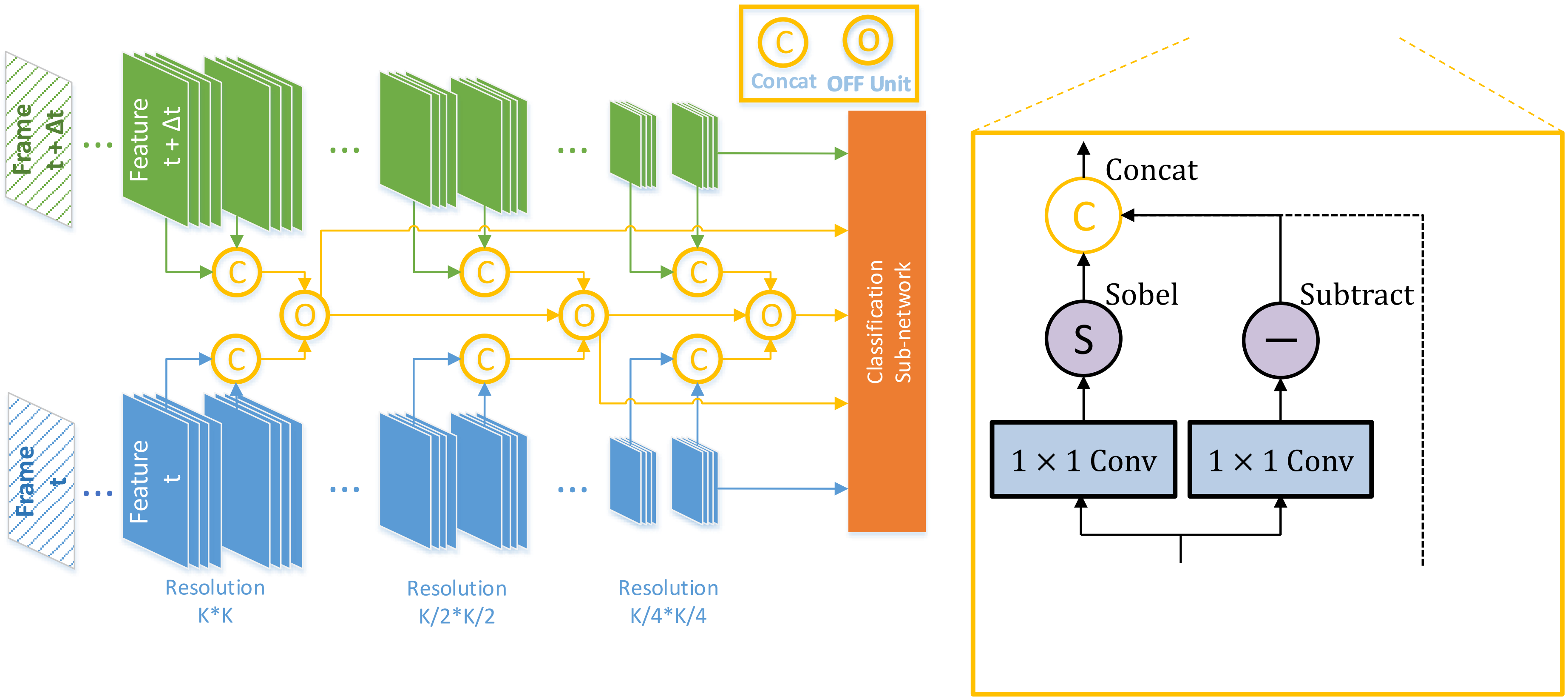


Optical flow guided feature (OFF):

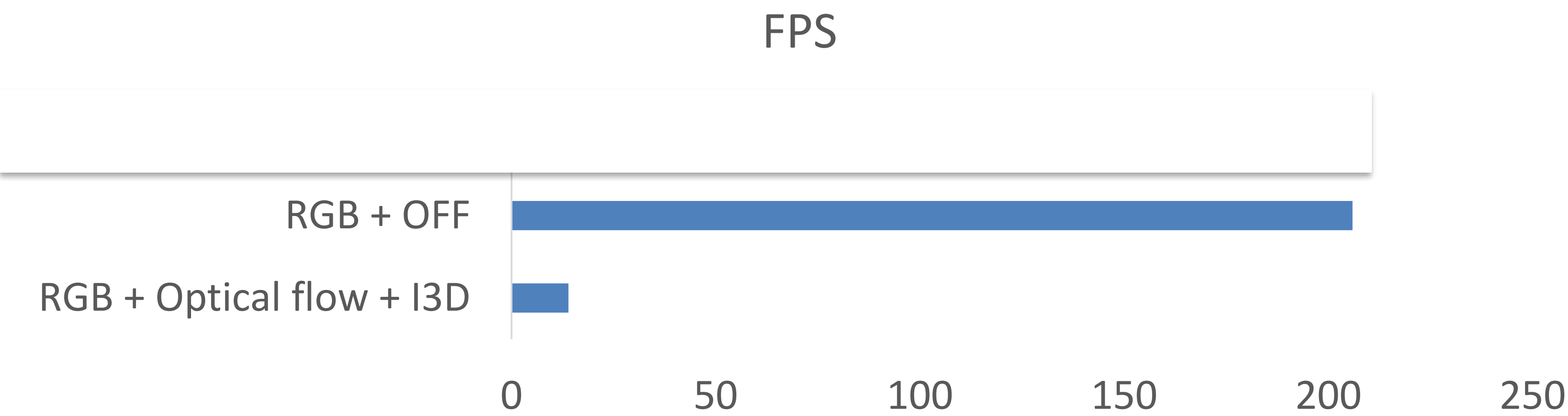
$$\{\frac{\partial f(I(x,y,t);w)}{\partial x},\frac{\partial f(I(x,y,t);w)}{\partial y},\frac{\partial f(I(x,y,t);w)}{\partial t}\}$$

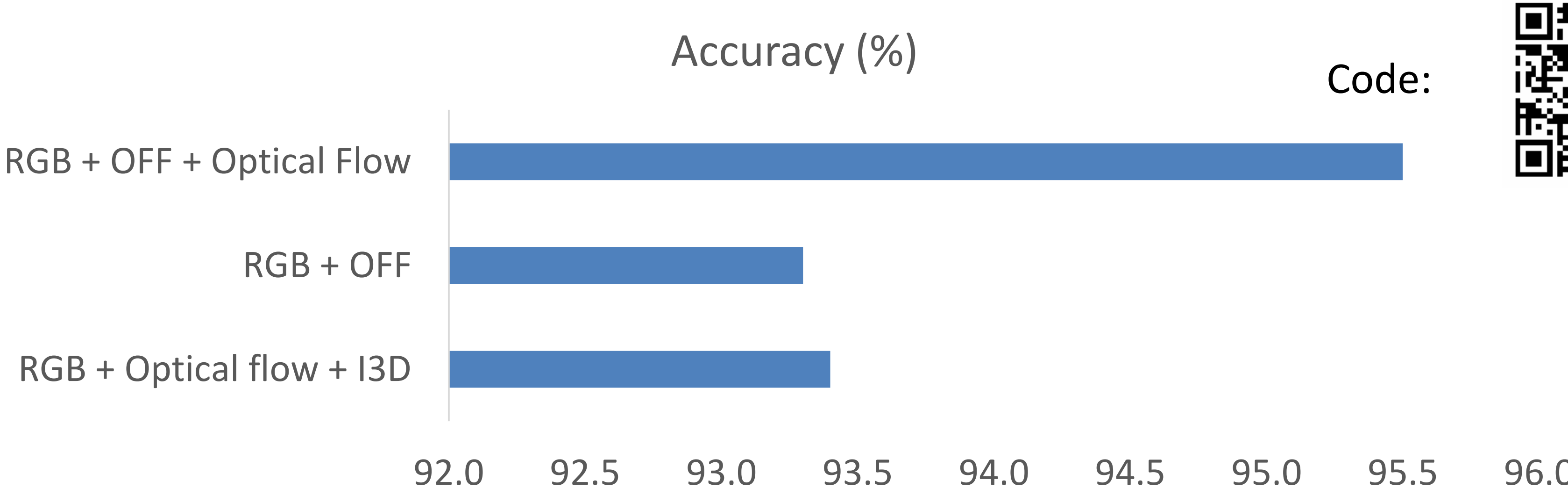


Optical flow guided feature



Optical Flow Guided Feature (OFF): Experimental results



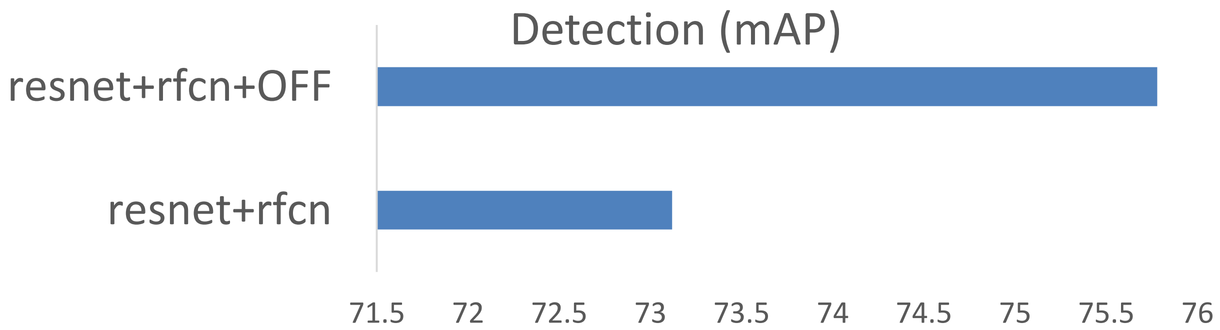


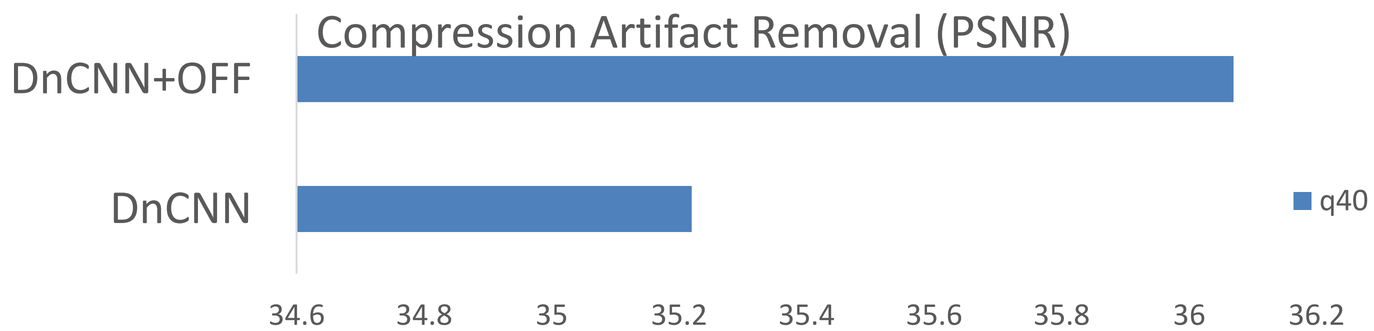
Code:

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 compression artifact removal



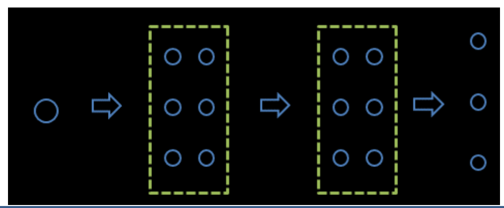


Outline

Introduction

Effectively using high performance images

Back-bone model design



FishNet
(NIPS 18)

Optical guided feature
(CVPR18)

Take home message

- Structured deep learning is
  - effective
- Effectively using high performance imaging as the privileged information by exploring the structured information at
  - Sample level
  - Feature level
- End-to-end joint training bridges the gap between structure modeling and feature learning

Joint work

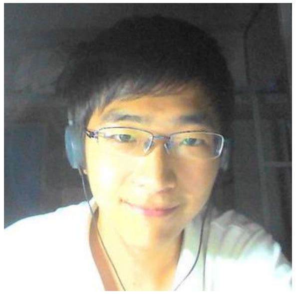
Xiaogang Wang

Nicu Sebe

Elisa Ricci

Wei Yang

Dan Xu



Shuyang Sun


Thank You