

# Deep Residual Learning

MSRA @ ILSVRC & COCO 2015 competitions

Kaiming He

with Xiangyu Zhang, Shaoqing Ren, Jifeng Dai, & Jian Sun

Microsoft Research Asia (MSRA)

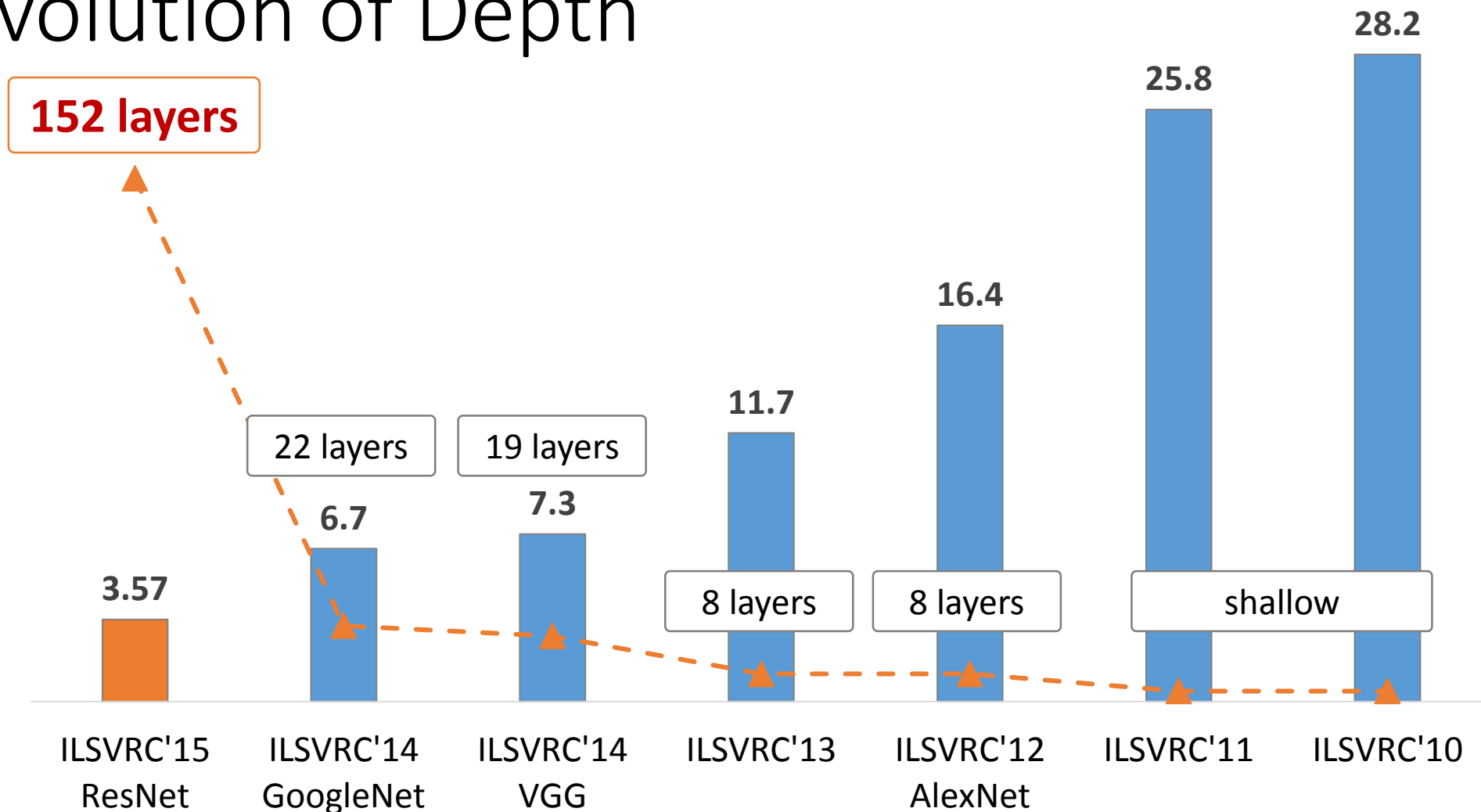
# MSRA @ ILSVRC & COCO 2015 Competitions

- **1st places in all five main tracks**

- ImageNet Classification: “*Ultra-deep*” (quote Yann) **152-layer** nets
- ImageNet Detection: **16%** better than 2nd
- ImageNet Localization: **27%** better than 2nd
- COCO Detection: **11%** better than 2nd
- COCO Segmentation: **12%** better than 2nd

\*improvements are relative numbers

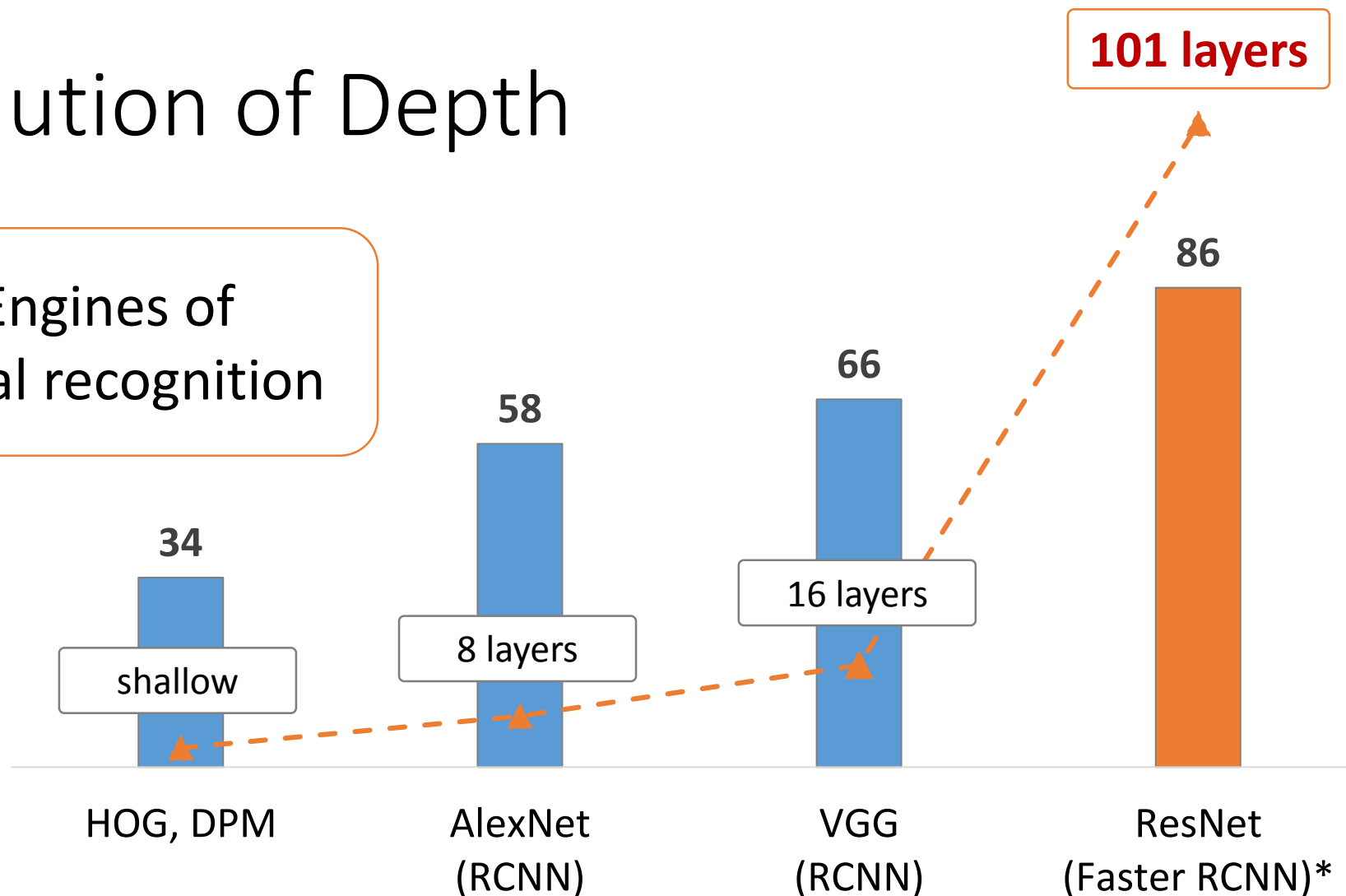
# Revolution of Depth



ImageNet Classification top-5 error (%)

# Revolution of Depth

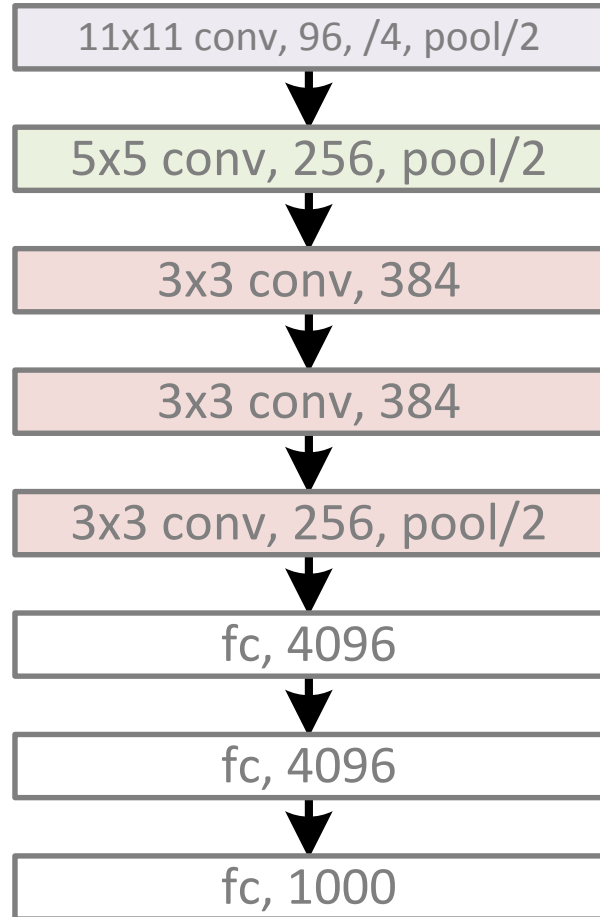
Engines of  
visual recognition



PASCAL VOC 2007 **Object Detection** mAP (%)

# Revolution of Depth

AlexNet, 8 layers  
(ILSVRC 2012)





# Revolution of Depth

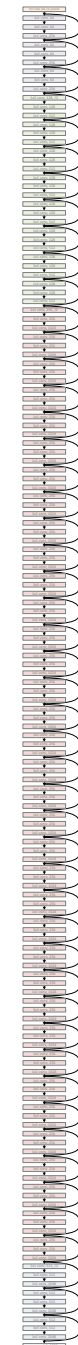
AlexNet, 8 layers  
(ILSVRC 2012)



VGG, 19 layers  
(ILSVRC 2014)

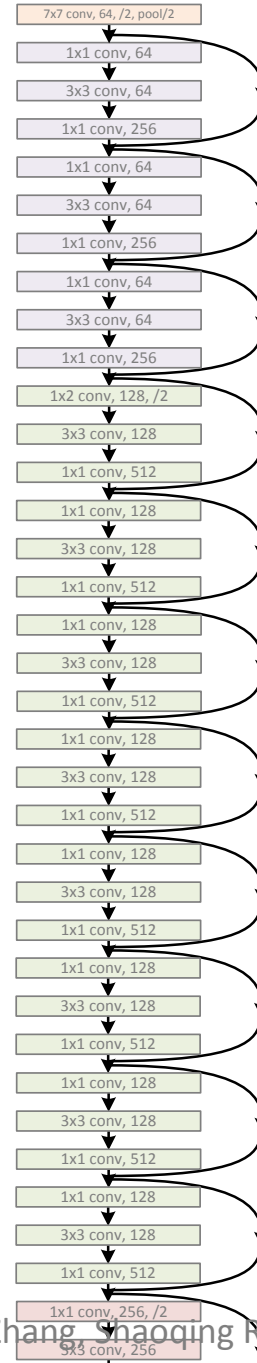


ResNet, 152 layers  
(ILSVRC 2015)



# Revolution of Depth

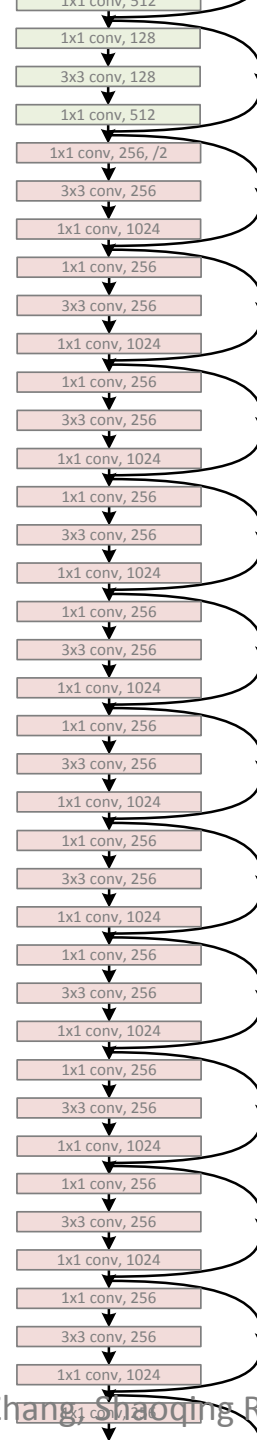
ResNet, 152 layers



(there was an animation here)

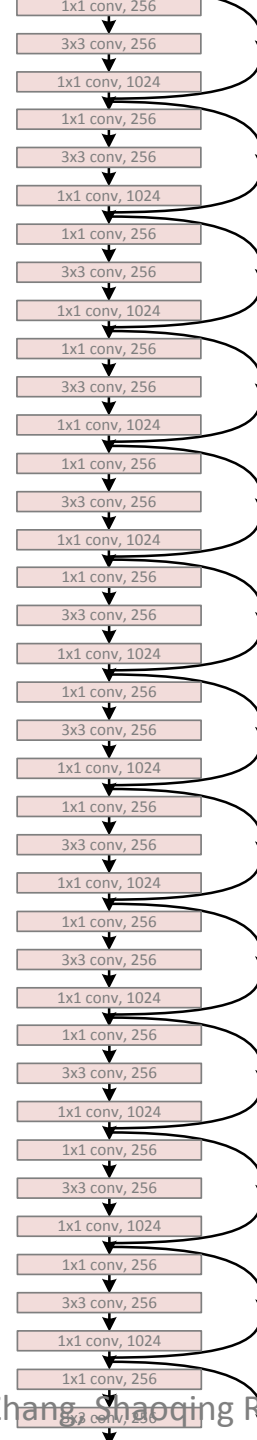


## ResNet, 152 layers



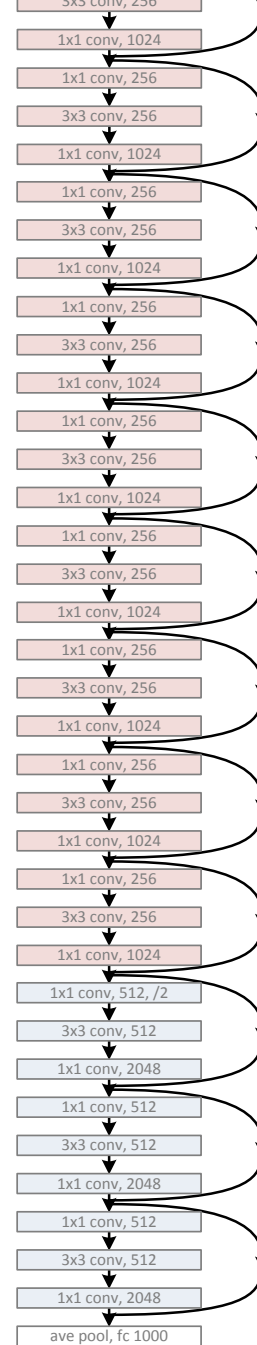
(there was an animation here)

## ResNet, 152 layers



(there was an animation here)

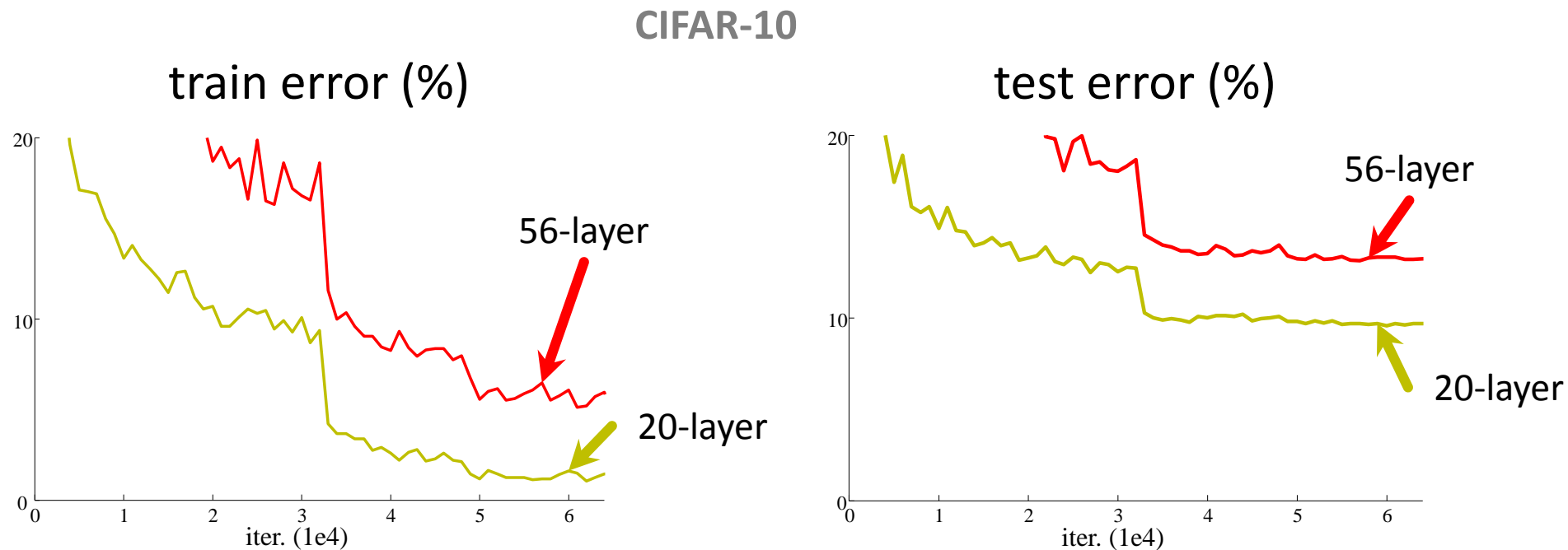
## ResNet, 152 layers



(there was an animation here)

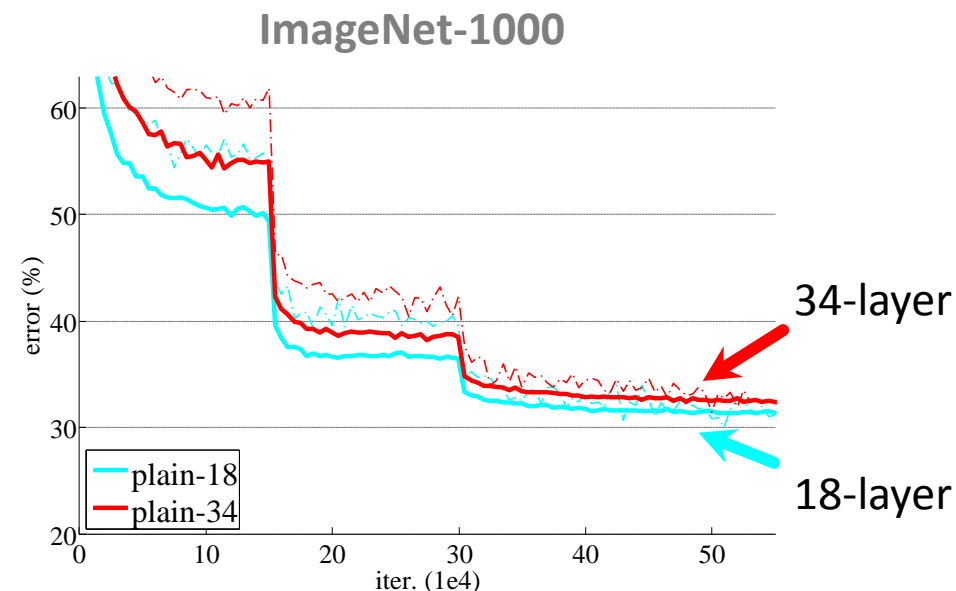
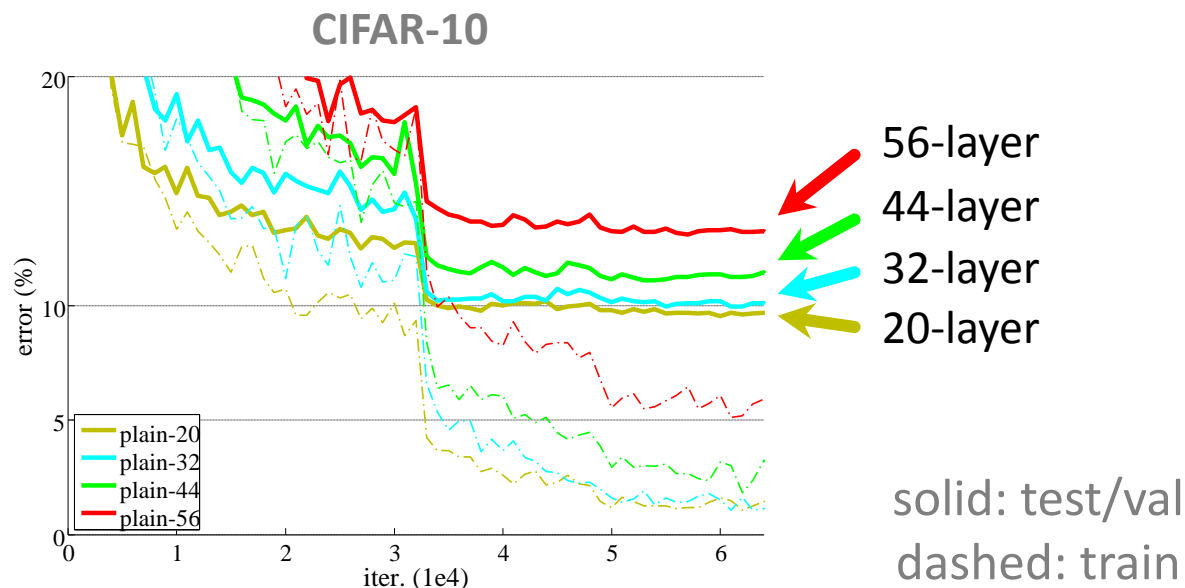
Is learning better networks  
as simple as stacking more layers?

# Simply stacking layers?



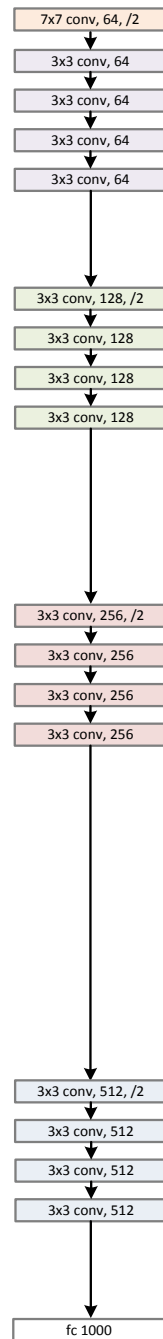
- *Plain* nets: stacking 3x3 conv layers...
- 56-layer net has **higher training error** and test error than 20-layer net

# Simply stacking layers?

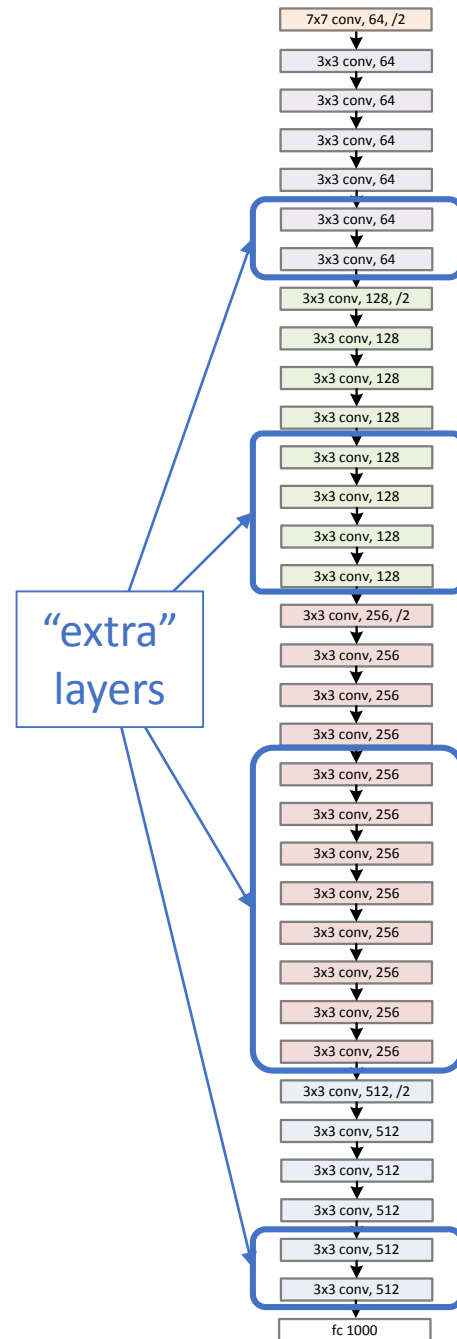


- “Overly deep” plain nets have **higher training error**
- A general phenomenon, observed in many datasets

a shallower  
model  
(18 layers)



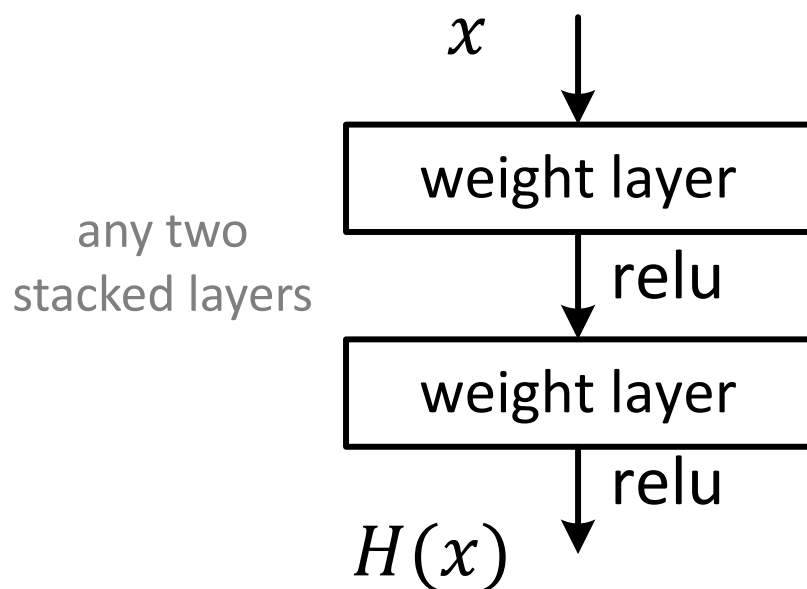
a deeper  
counterpart  
(34 layers)



- A deeper model should not have **higher training error**
- A solution *by construction*:
  - original layers: copied from a learned shallower model
  - extra layers: set as **identity**
  - at least the same training error
- **Optimization difficulties**: solvers cannot find the solution when going deeper...

# Deep Residual Learning

- Plain net

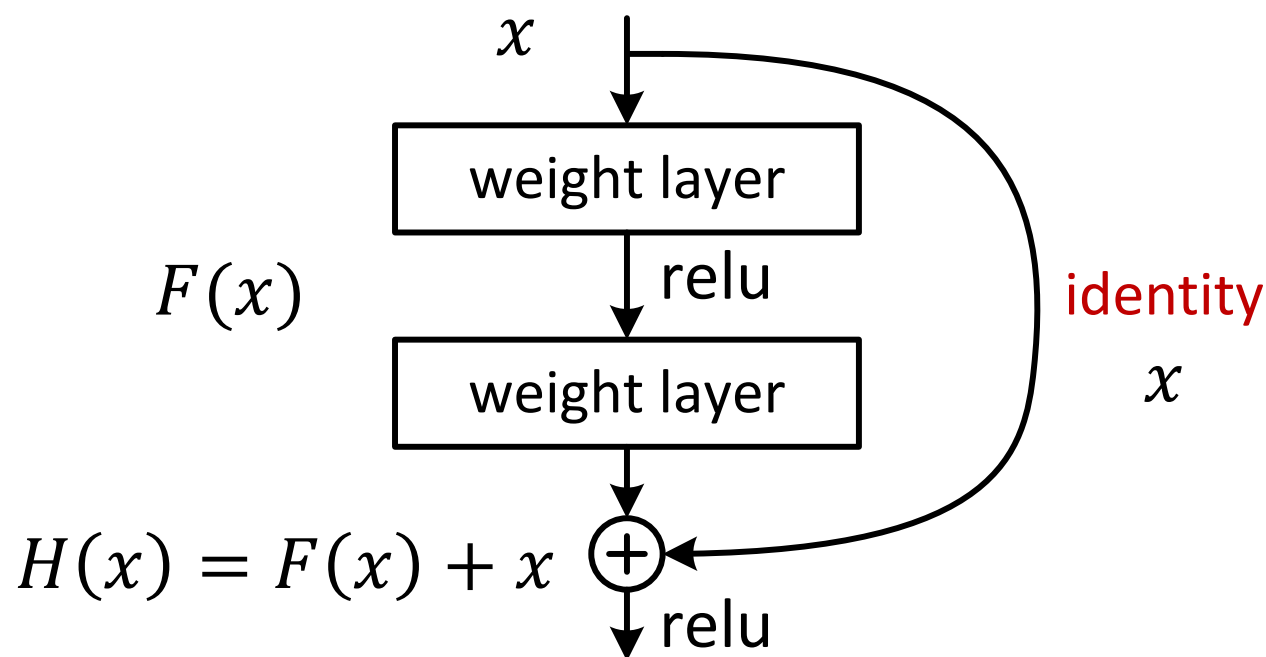


$H(x)$  is any desired mapping,  
hope the 2 weight layers fit  $H(x)$



# Deep Residual Learning

- Residual net



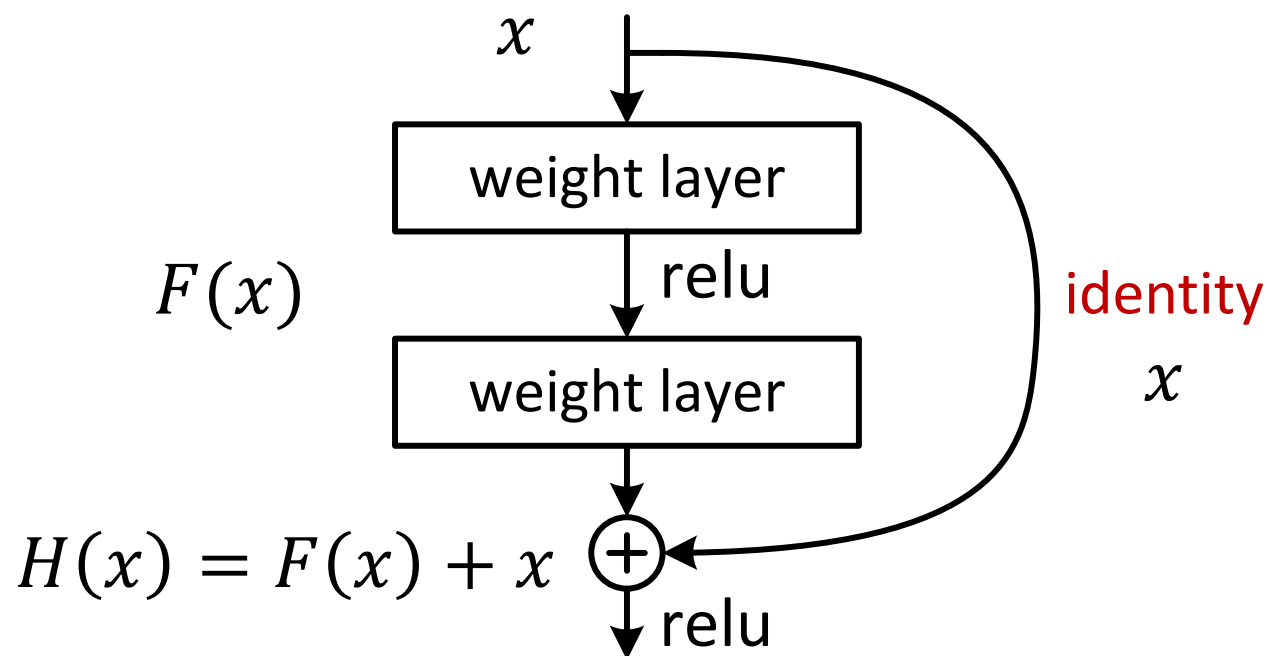
$H(x)$  is any desired mapping,  
~~hope the 2 weight layers fit  $H(x)$~~

hope the 2 weight layers fit  $F(x)$

$$\text{let } H(x) = F(x) + x$$

# Deep Residual Learning

- $F(x)$  is a **residual** mapping w.r.t. **identity**



- If identity were optimal, easy to set weights as 0
- If optimal mapping is closer to identity, easier to find small fluctuations

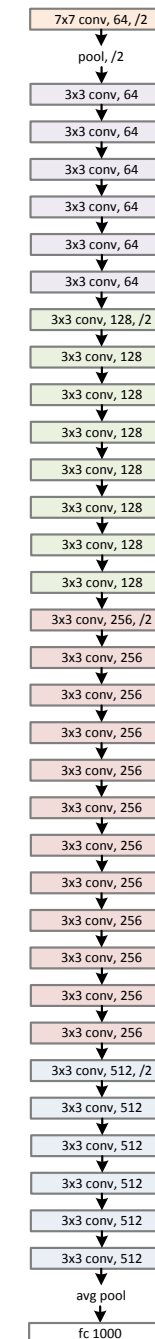
# Related Works – Residual Representations

- VLAD & Fisher Vector [Jegou et al 2010], [Perronnin et al 2007]
  - Encoding residual vectors; powerful shallower representations.
- Product Quantization (IVF-ADC) [Jegou et al 2011]
  - Quantizing residual vectors; efficient nearest-neighbor search.
- MultiGrid & Hierarchical Precondition [Briggs, et al 2000], [Szeliski 1990, 2006]
  - Solving residual sub-problems; efficient PDE solvers.

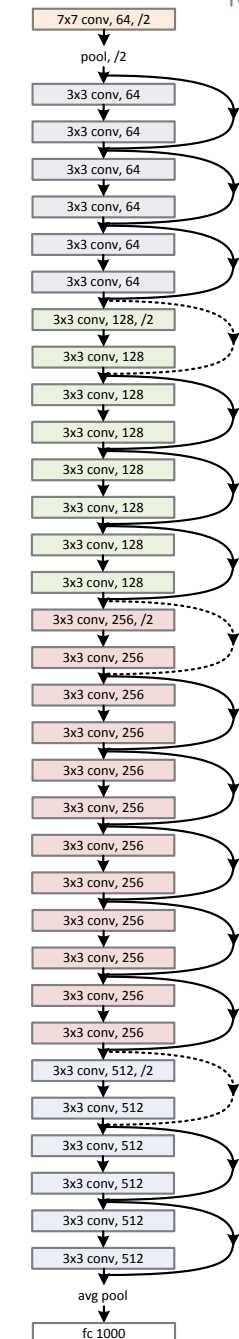
# Network “Design”

- Keep it simple
- Our basic design (VGG-style)
  - all 3x3 conv (almost)
  - spatial size /2 => # filters x2
  - **Simple design; just deep!**
- Other remarks:
  - no max pooling (almost)
  - no hidden fc
  - no dropout

plain net



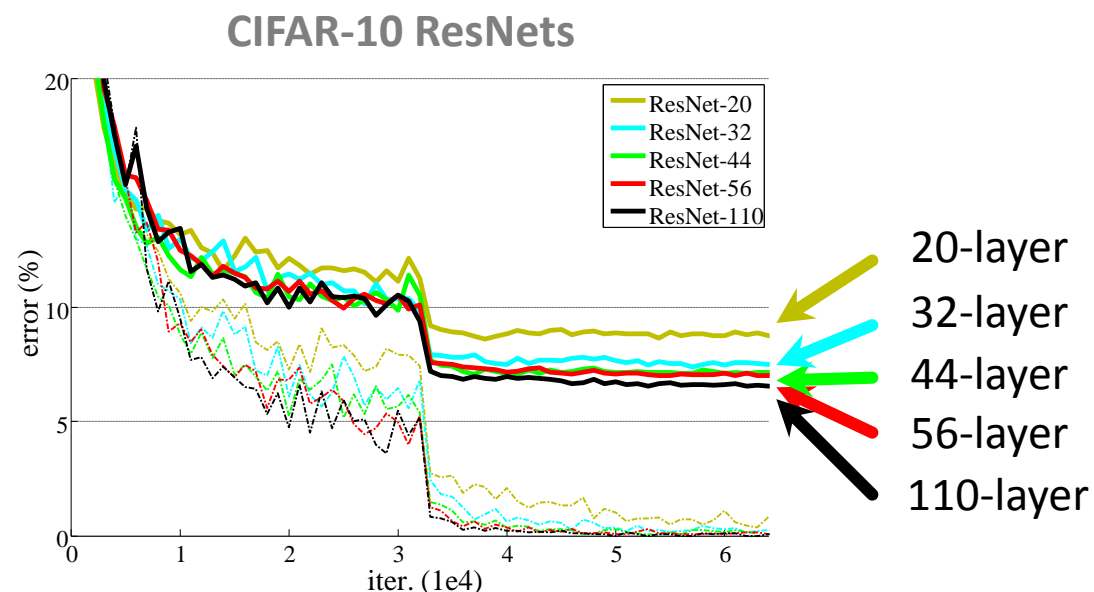
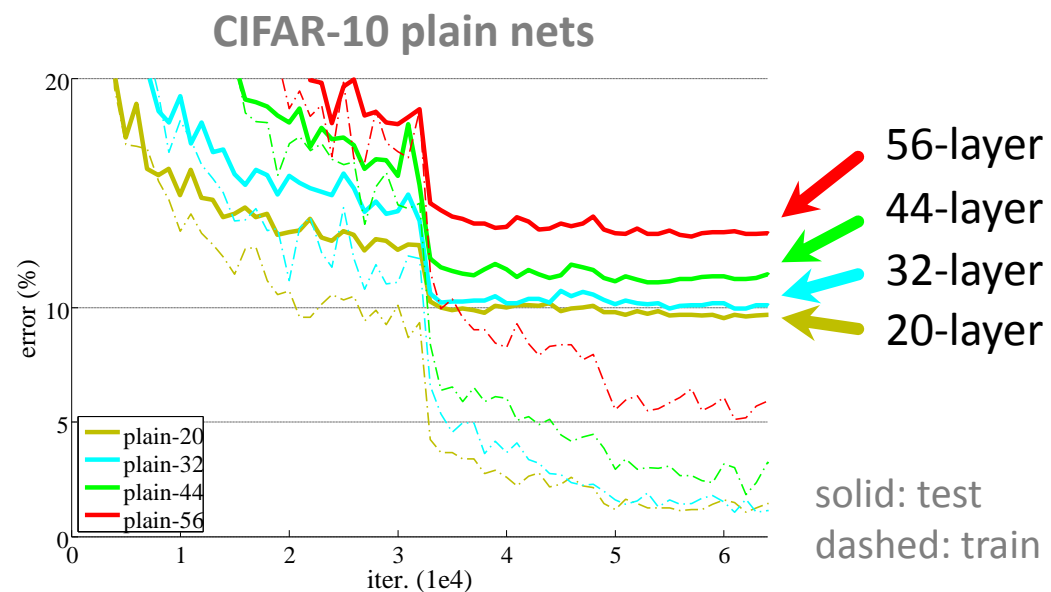
ResNet



# Training

- All plain/residual nets are trained **from scratch**
- All plain/residual nets use Batch Normalization
- Standard hyper-parameters & augmentation

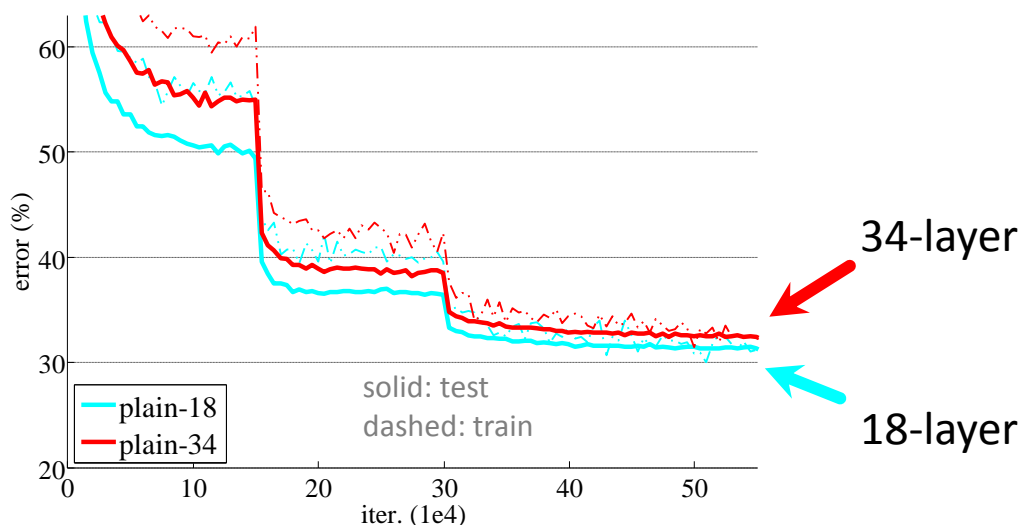
# CIFAR-10 experiments



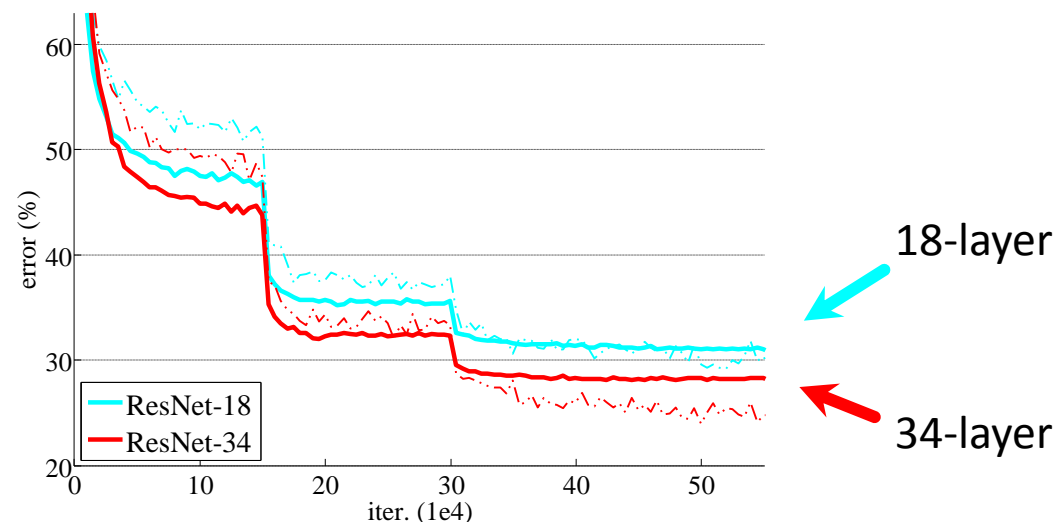
- Deep ResNets can be trained without difficulties
- Deeper ResNets have **lower training error**, and also lower test error

# ImageNet experiments

ImageNet plain nets



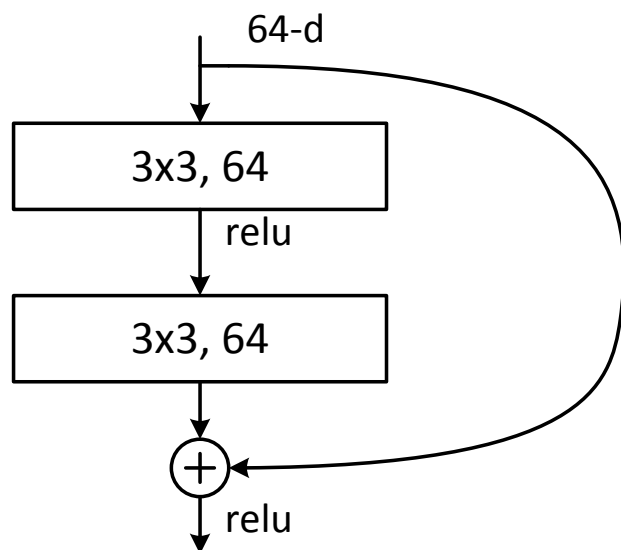
ImageNet ResNets



- Deep ResNets can be trained without difficulties
- Deeper ResNets have **lower training error**, and also lower test error

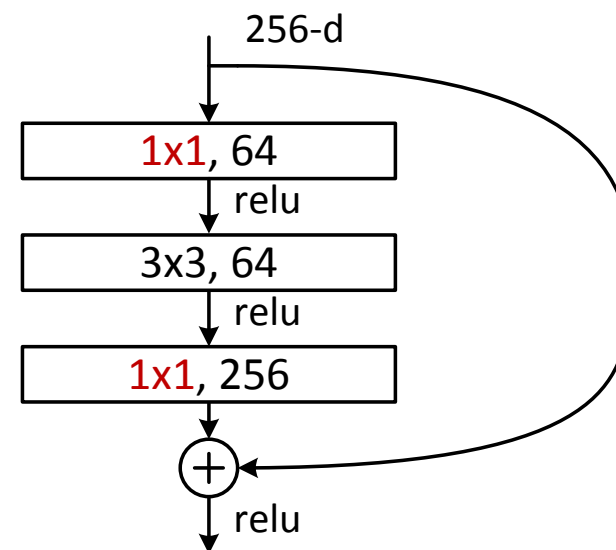
# ImageNet experiments

- A practical design of going deeper



all-3x3

similar  
complexity



**bottleneck**

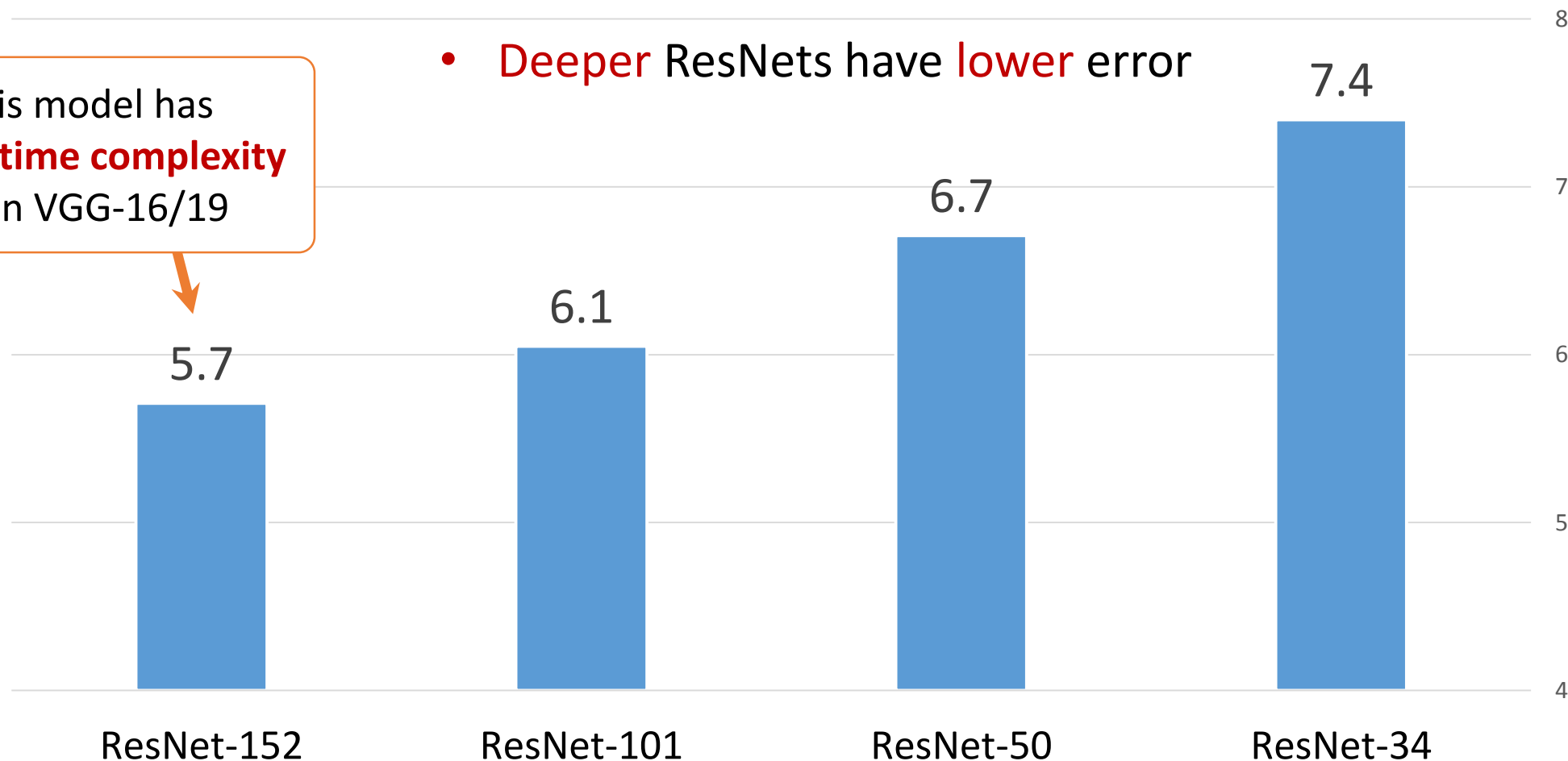
(for ResNet-50/101/152)



# ImageNet experiments

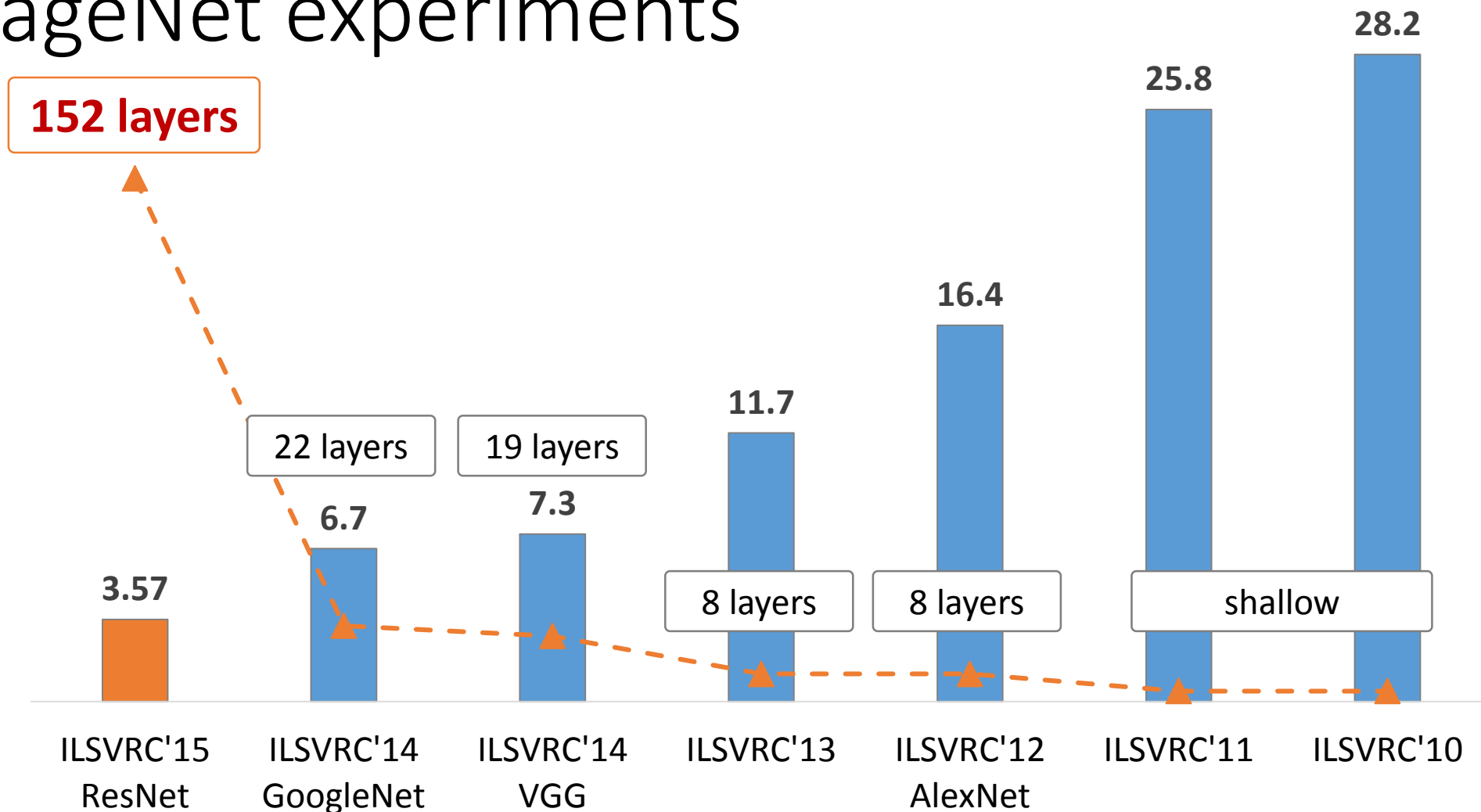
- Deeper ResNets have lower error

this model has  
**lower time complexity**  
than VGG-16/19



**10-crop** testing, top-5 val error (%)

# ImageNet experiments



ImageNet Classification top-5 error (%)

# Just classification?

**A treasure from ImageNet is on **learning features**.**

*“Features matter.”* (quote [Girshick et al. 2014], the R-CNN paper)

task	2nd-place winner	MSRA	margin (relative)
ImageNet Localization (top-5 error)	12.0	9.0	<b>27%</b>
ImageNet Detection (mAP@.5)	53.6	62.1	<b>16%</b>
COCO Detection (mAP@.5:.95)	33.5	37.3	<b>11%</b>
COCO Segmentation (mAP@.5:.95)	25.1	28.2	<b>12%</b>

**absolute  
8.5% better!**

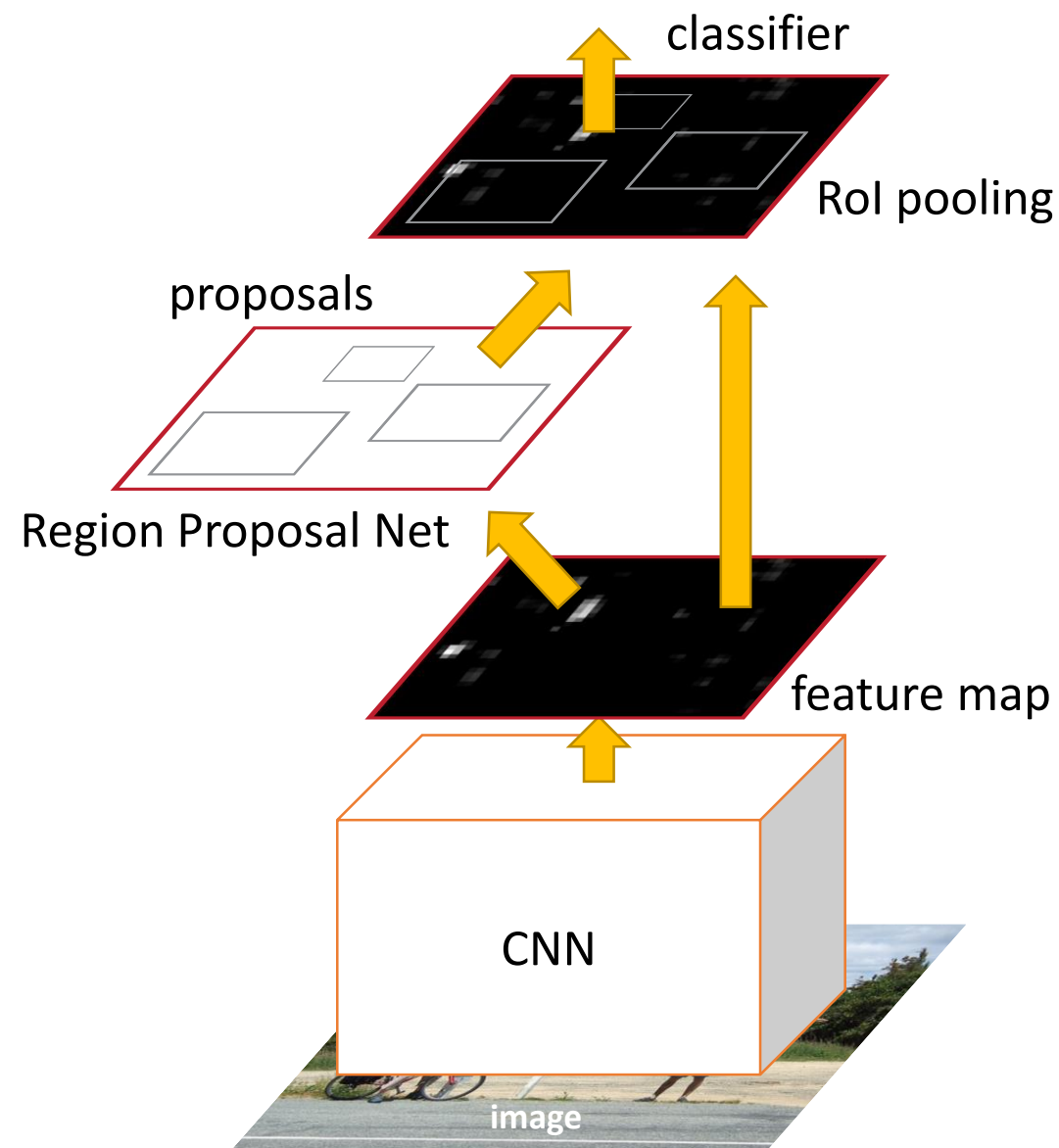
- Our results are all based on **ResNet-101**
- Our features are **well transferrable**

# Object Detection (brief)

- Simply “Faster R-CNN + ResNet”

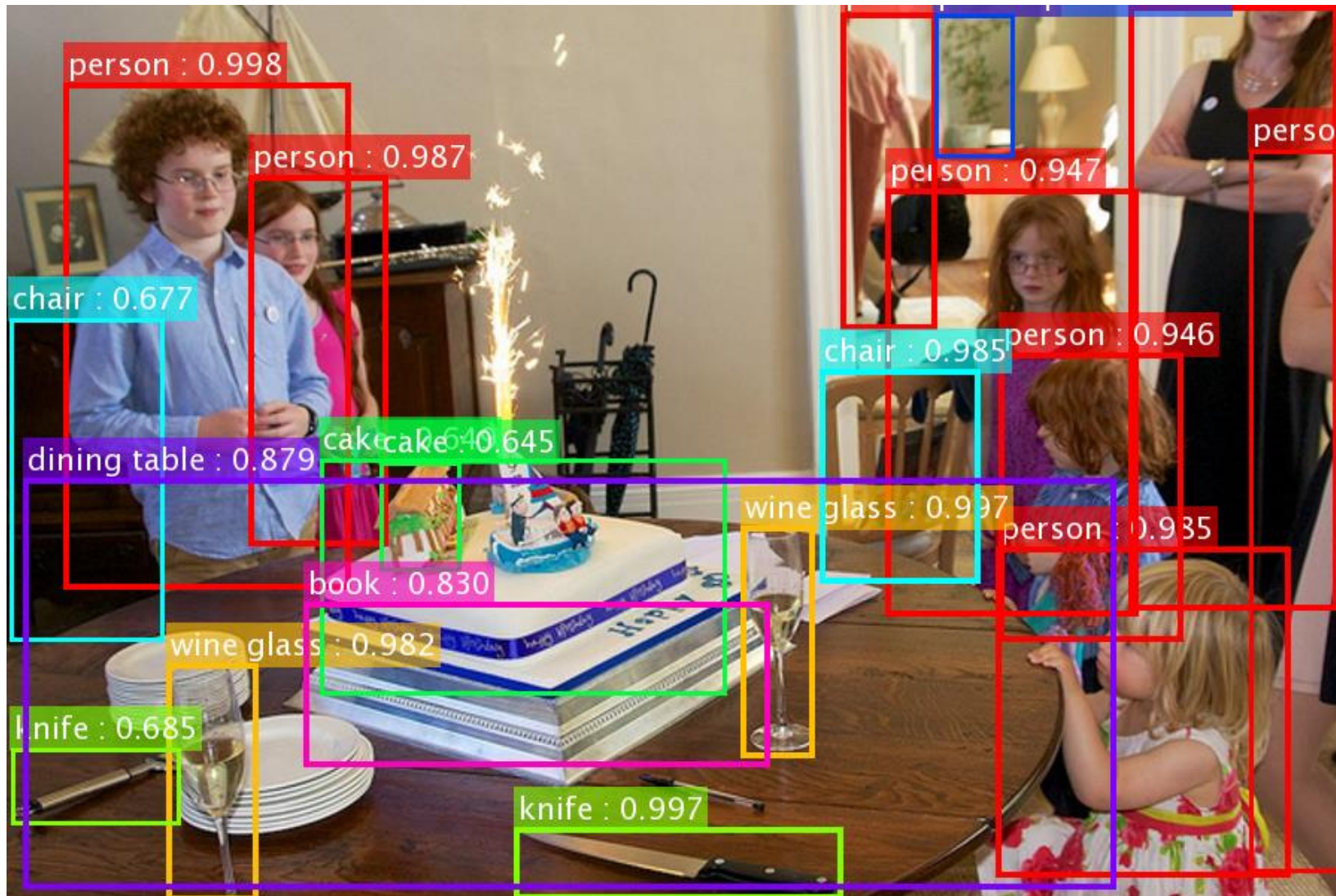
Faster R-CNN baseline	mAP@.5	mAP@.5:.95
VGG-16	41.5	21.5
ResNet-101	<b>48.4</b>	<b>27.2</b>

coco detection results  
(ResNet has 28% relative gain)



# Object Detection (brief)

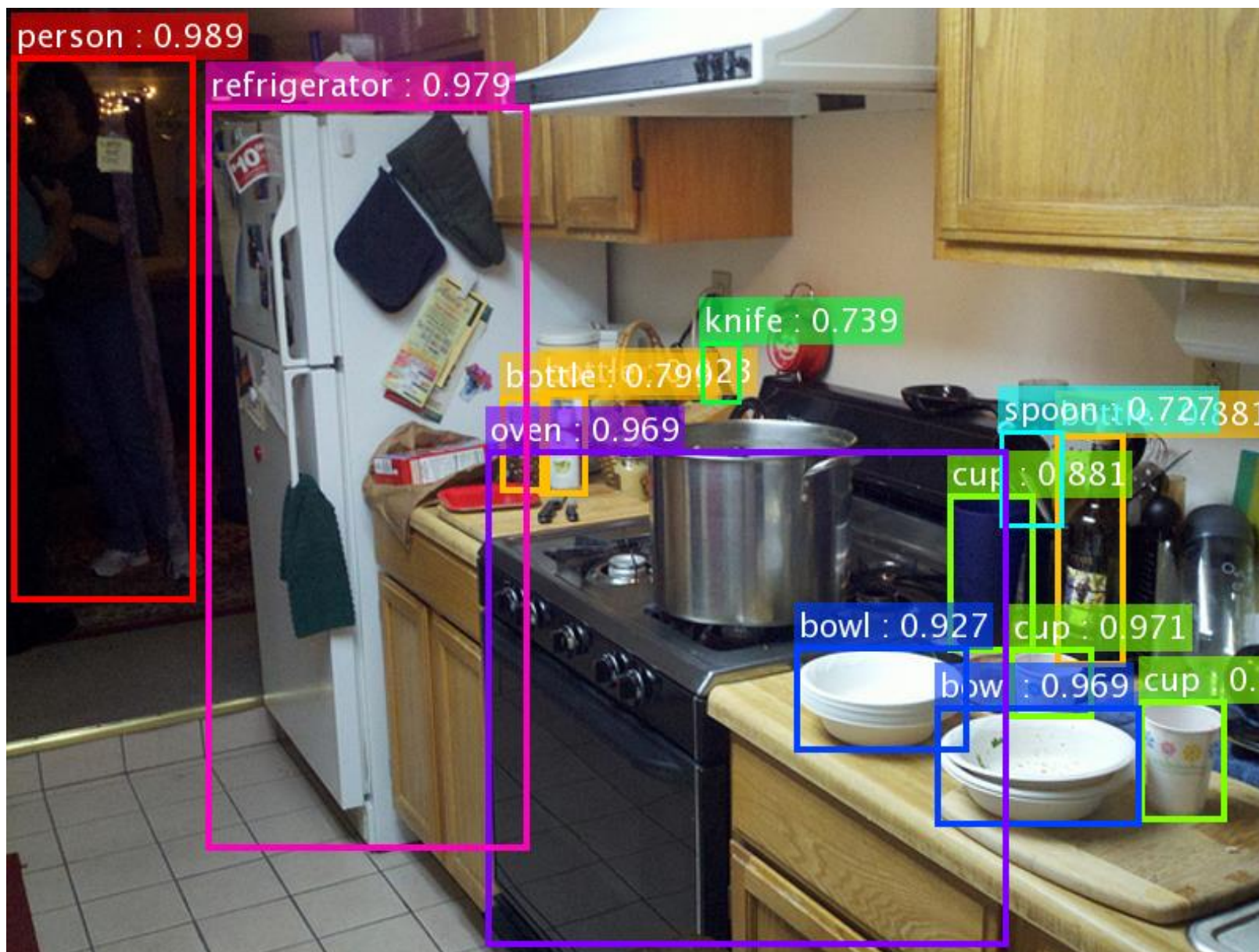
- RPN **learns** proposals by extremely deep nets
  - We use **only 300 proposals** (no SS/EB/MCG!)
- Add what is just missing in Faster R-CNN...
  - Iterative localization
  - Context modeling
  - Multi-scale testing
- All are based on CNN features; all are end-to-end (train and/or inference)
- All benefit **more** from **deeper** features – cumulative gains!



Our results on COCO – too many objects, let's check carefully!

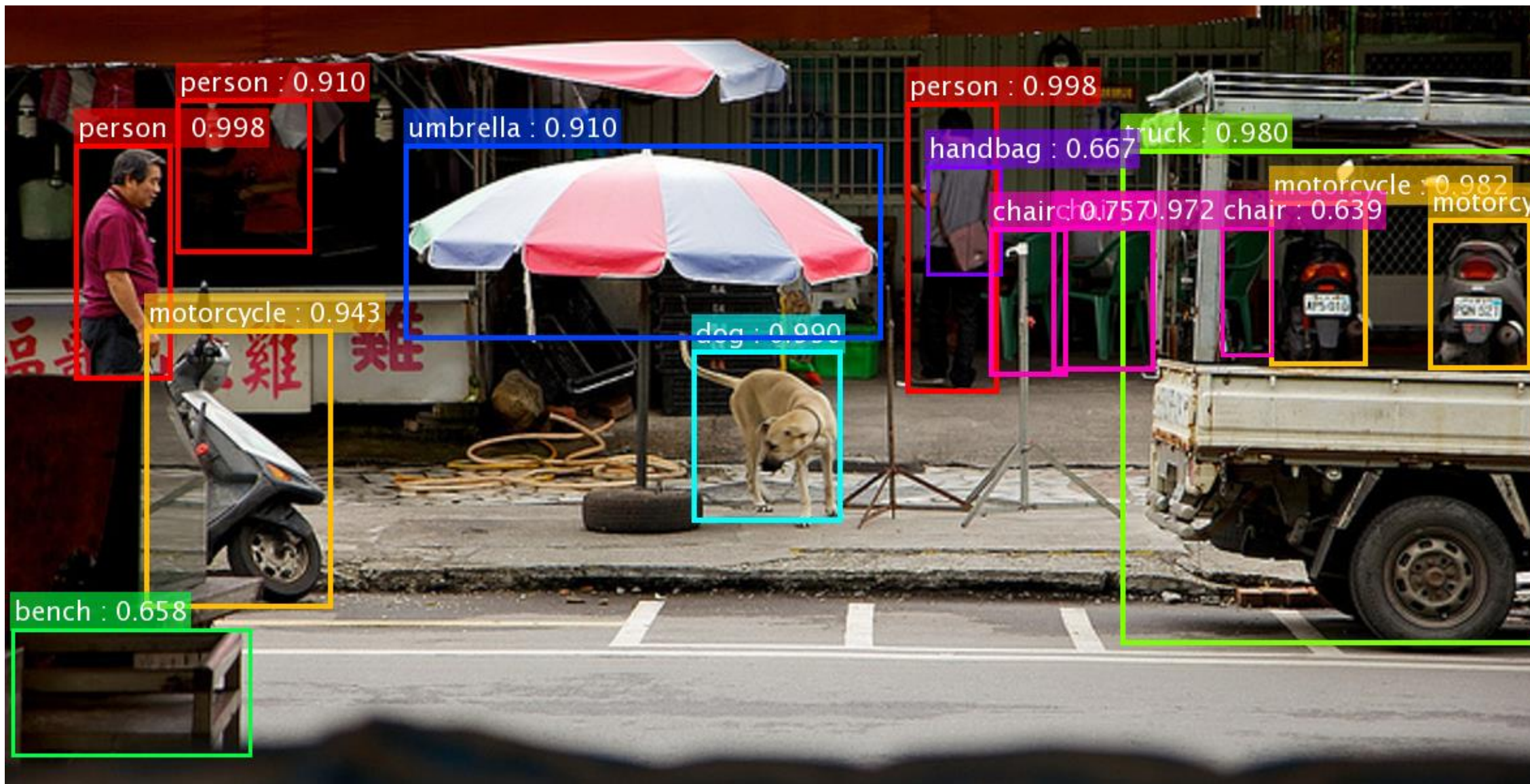
\*the original image is from the COCO dataset





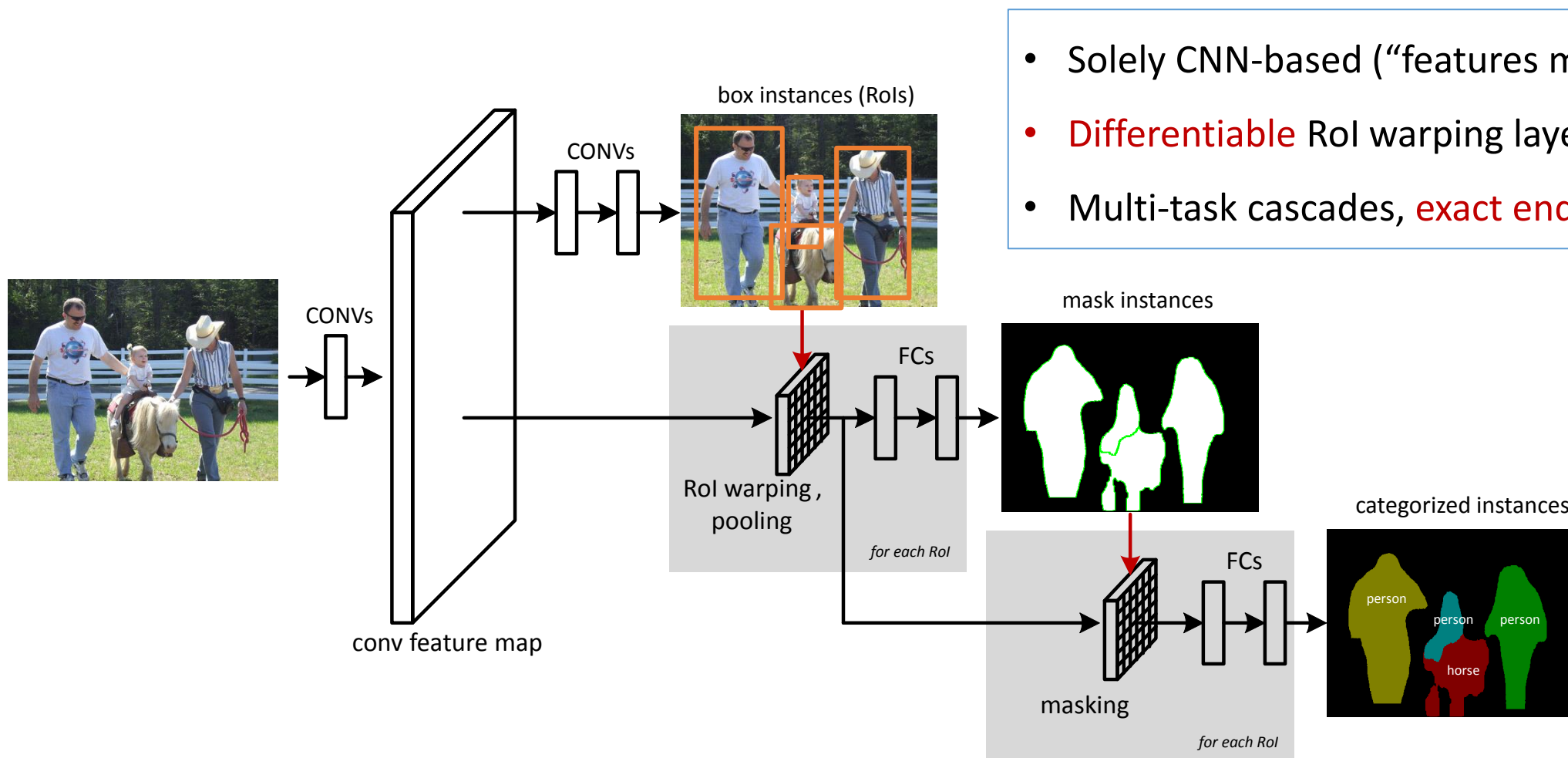
\*the original image is from the COCO dataset





\*the original image is from the COCO dataset

# Instance Segmentation (brief)

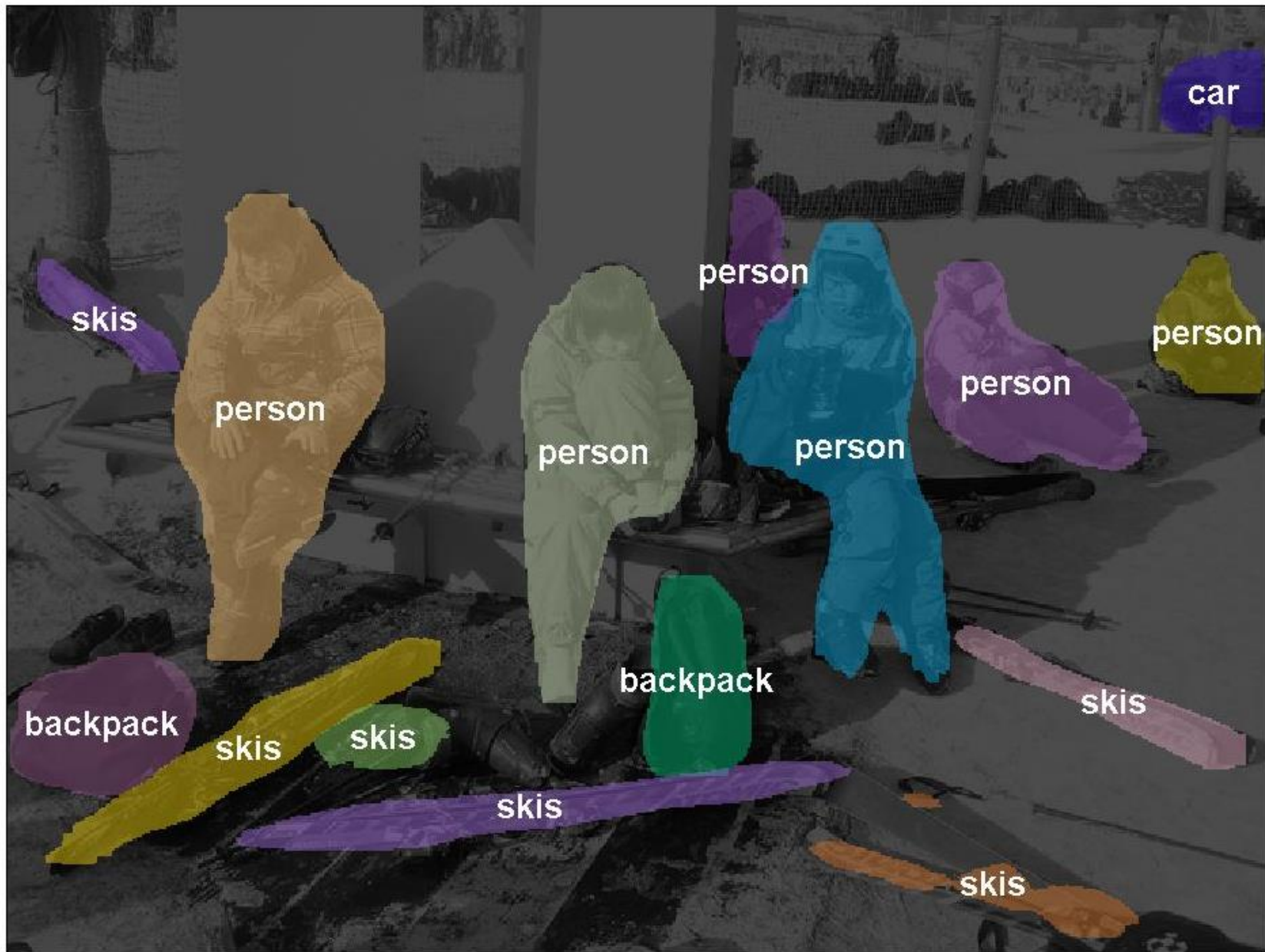


- Solely CNN-based (“features matter”)
- **Differentiable** RoI warping layer (w.r.t box coord.)
- Multi-task cascades, **exact end-to-end training**





input



\*the original image is from the COCO dataset

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.  
Jifeng Dai, Kaiming He, & Jian Sun. "Instance-aware Semantic Segmentation via Multi-task Network Cascades". arXiv 2015.

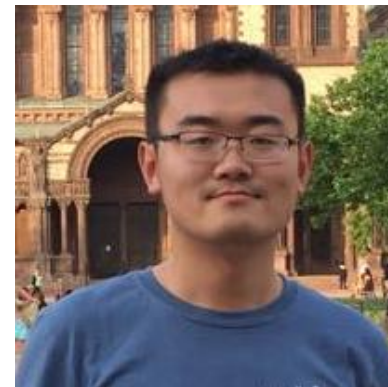
# Conclusions

- Deeper is still better
- “*Features matter*”!
- Faster R-CNN is just amazing

MSRA team



Kaiming He



Xiangyu Zhang



Shaoqing Ren



Jifeng Dai



Jian Sun