Deep Learning for Image Instance Segmentation
----RefineNet

Jianping Fan
Dept of Computer Science
UNC-Charlotte

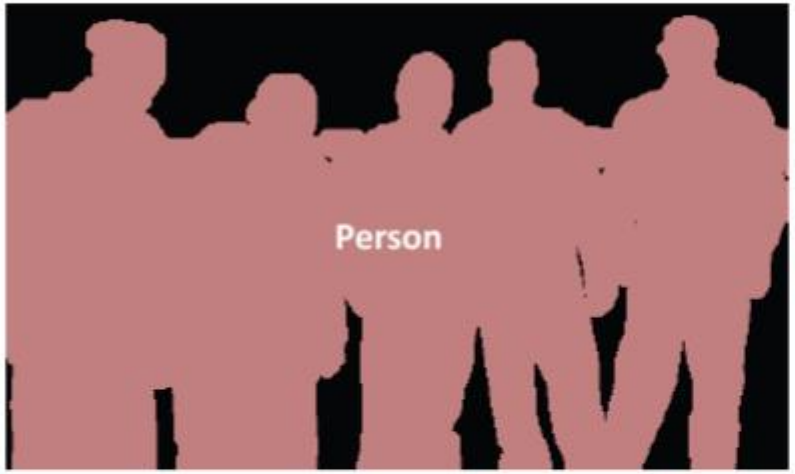
Course Website:
http://webpages.uncc.edu/jfan/itcs5152.html

Guosheng Lin, Anton Milan, Chunhua Shen, Ian Reid, **RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation,** arXiv:1611.06612, IEEE CVPR 2017

Definition of Image Instance Segmentation

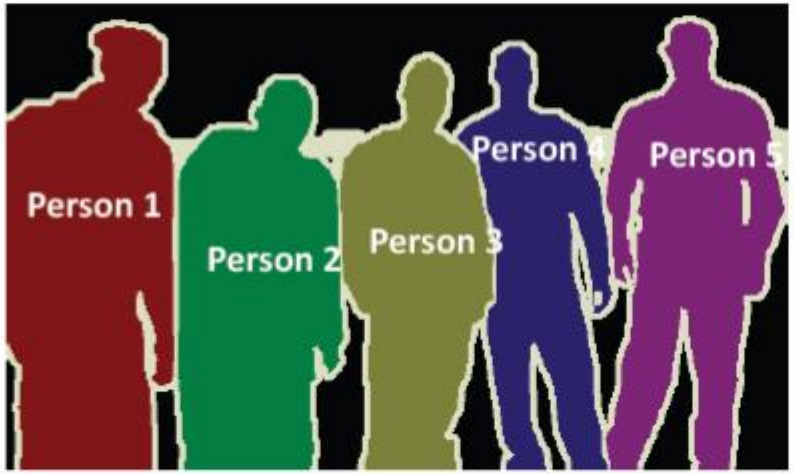
Object Detection

✓



Semantic Segmentation

✓



Instance Segmentation

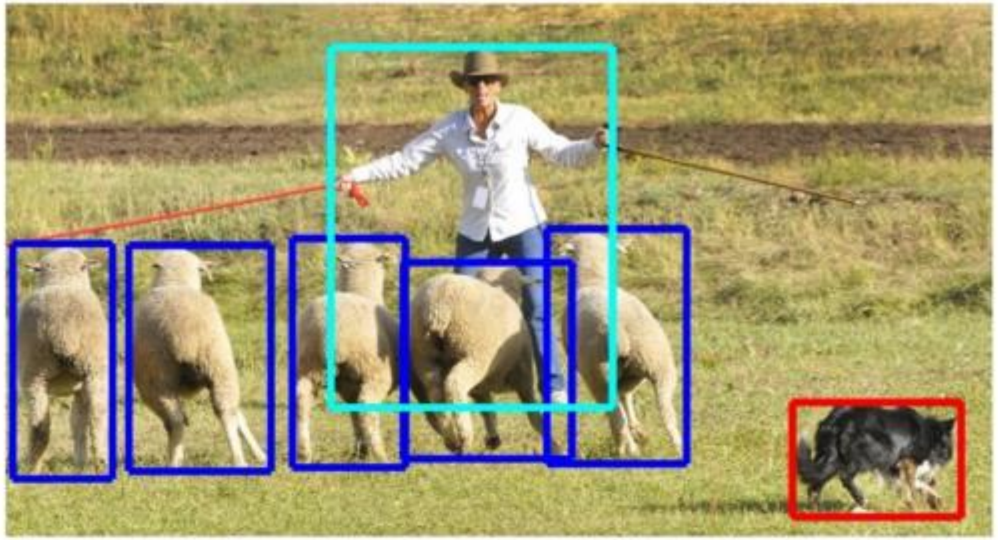
?

Instance segmentation = object detection + semantic segmentation?

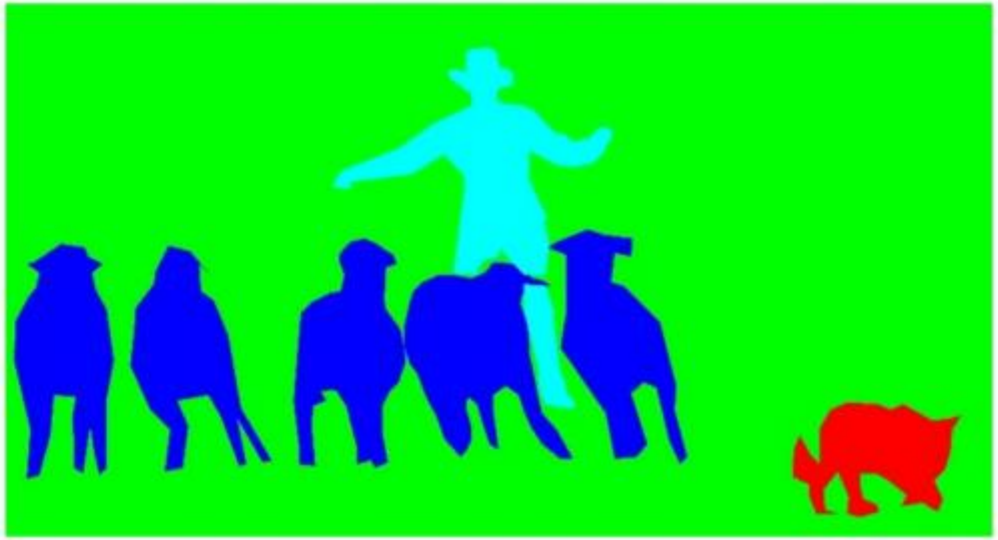
Slide from *Mask R-CNN Tutorial*, K. He. ICCV 2017

Scene understanding

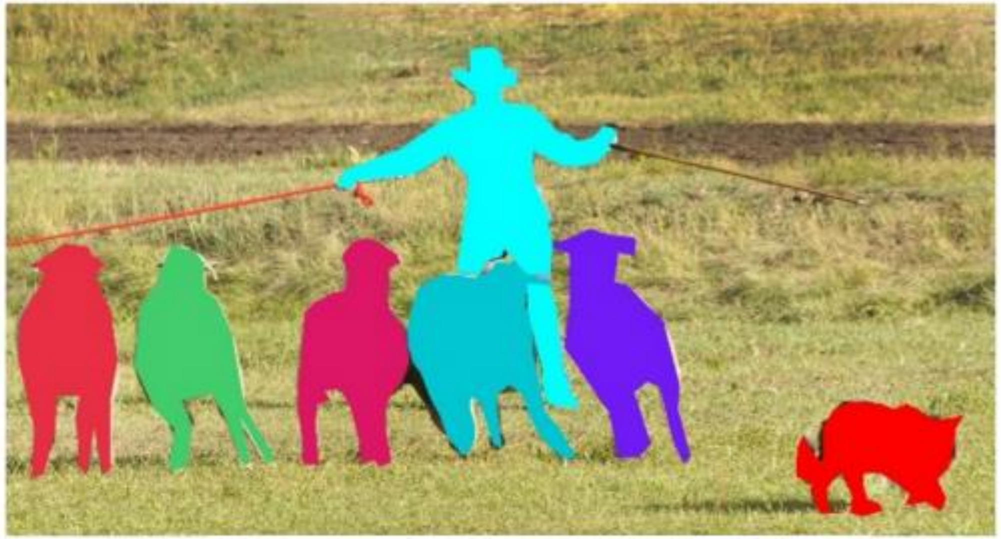
Image classification



Object detection



Semantic segmentation

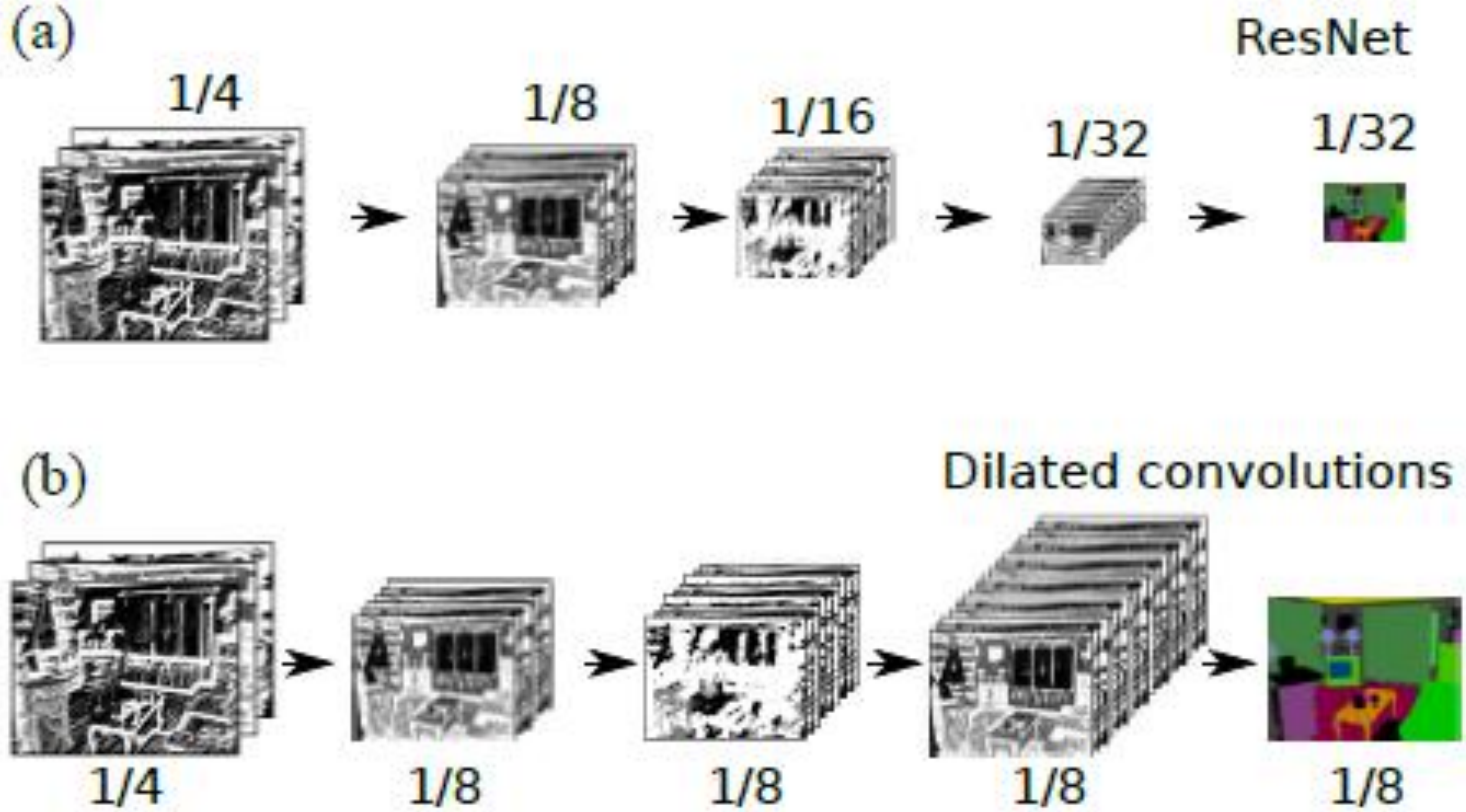


Instance segmentation

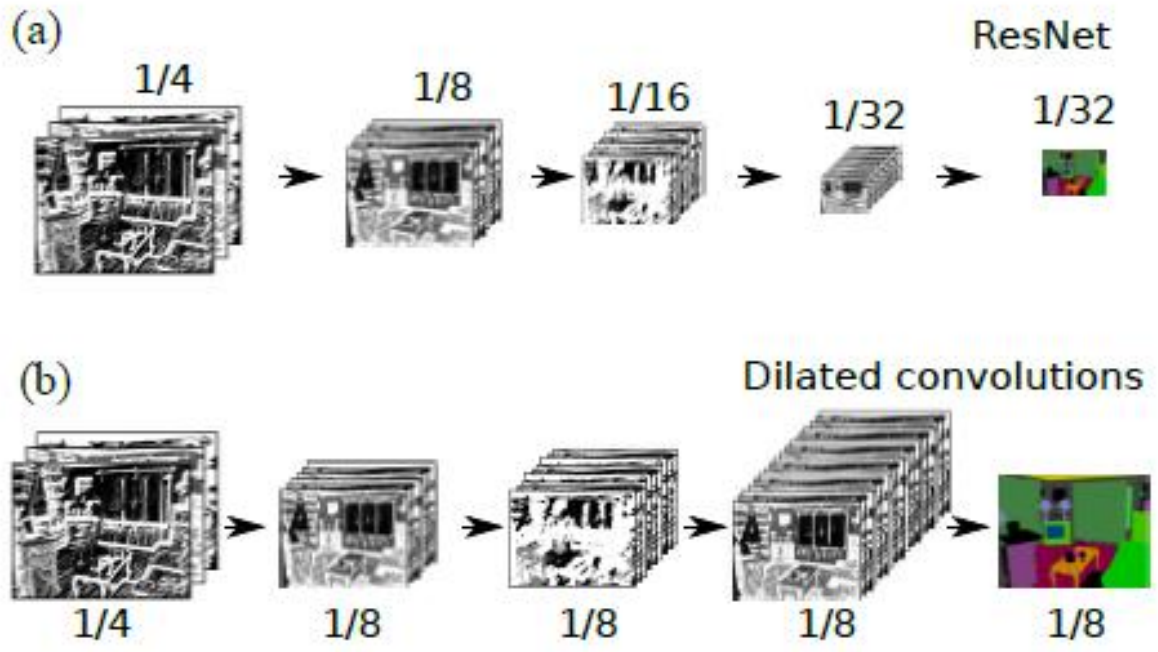
Instance-level Object Understanding Today

He, Gkioxari, Dollár, Girshick. Mask R-CNN. In ICCV 2017

Problems of ResNet and Dilated Convolution



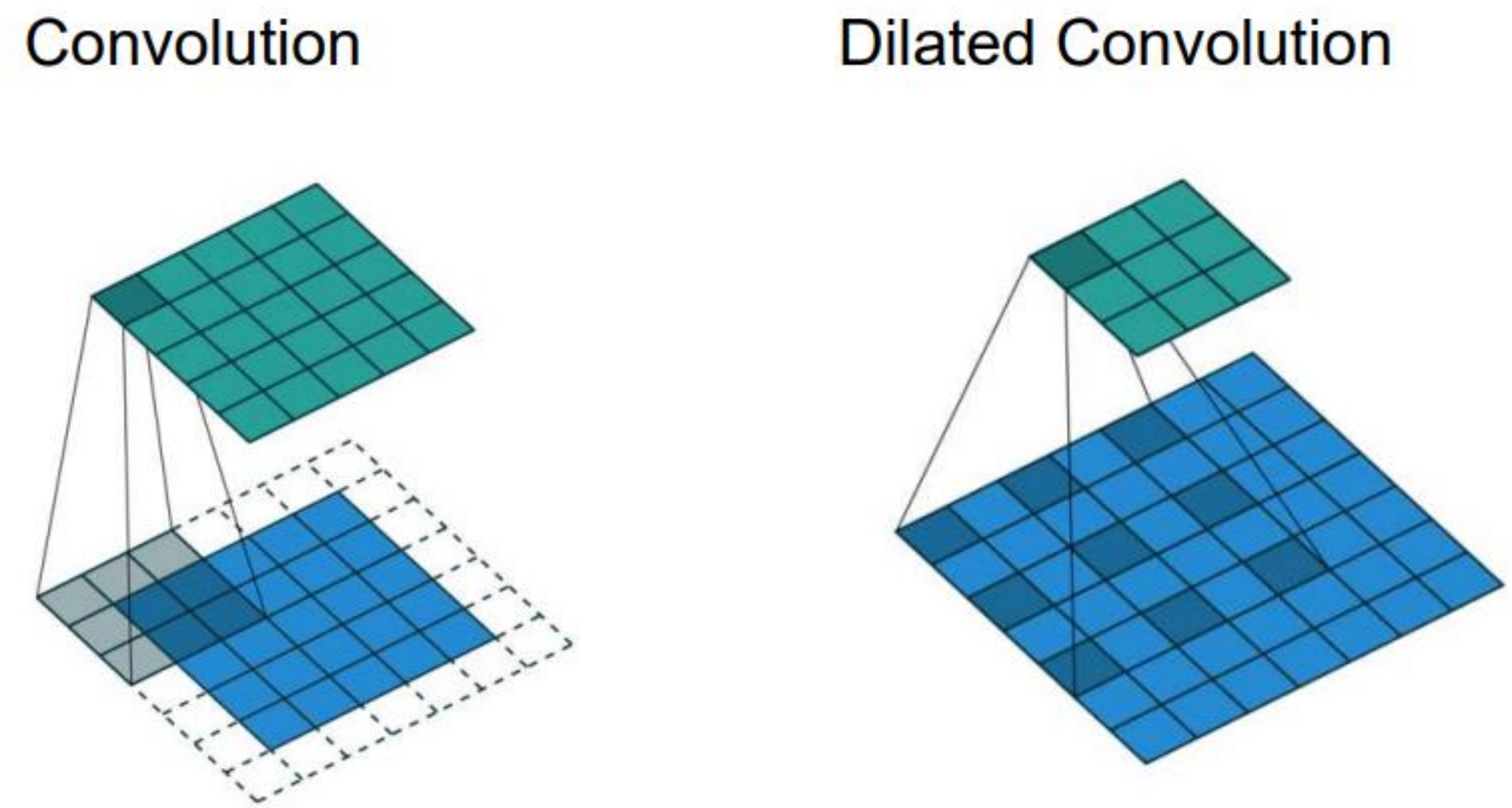
Problems of ResNet and Dilated Convolution



ResNet:
It **suffers from downscaling of the feature maps** which is not good for semantic segmentation.

Dilated (Atrous) Convolution:
It can help to keep the resolution of output feature maps larger, atrous filters are **computationally expensive to train** and quickly reach memory limits even on modern GPUs.

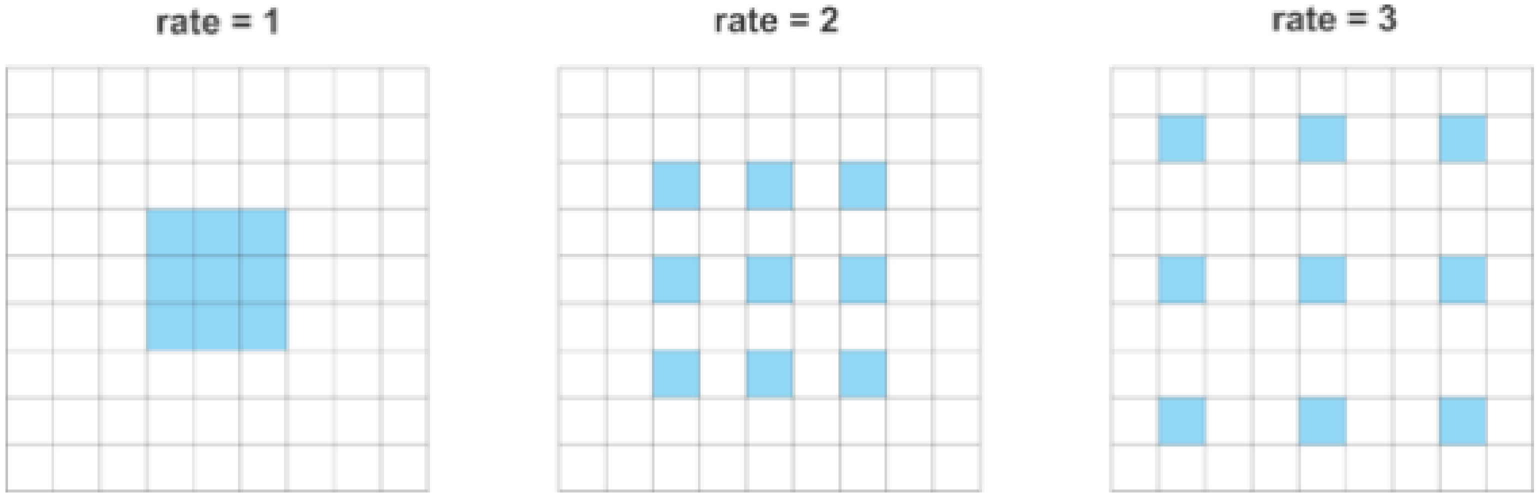
Problems of ResNet and Dilated Convolution



Problems of ResNet and Dilated Convolution

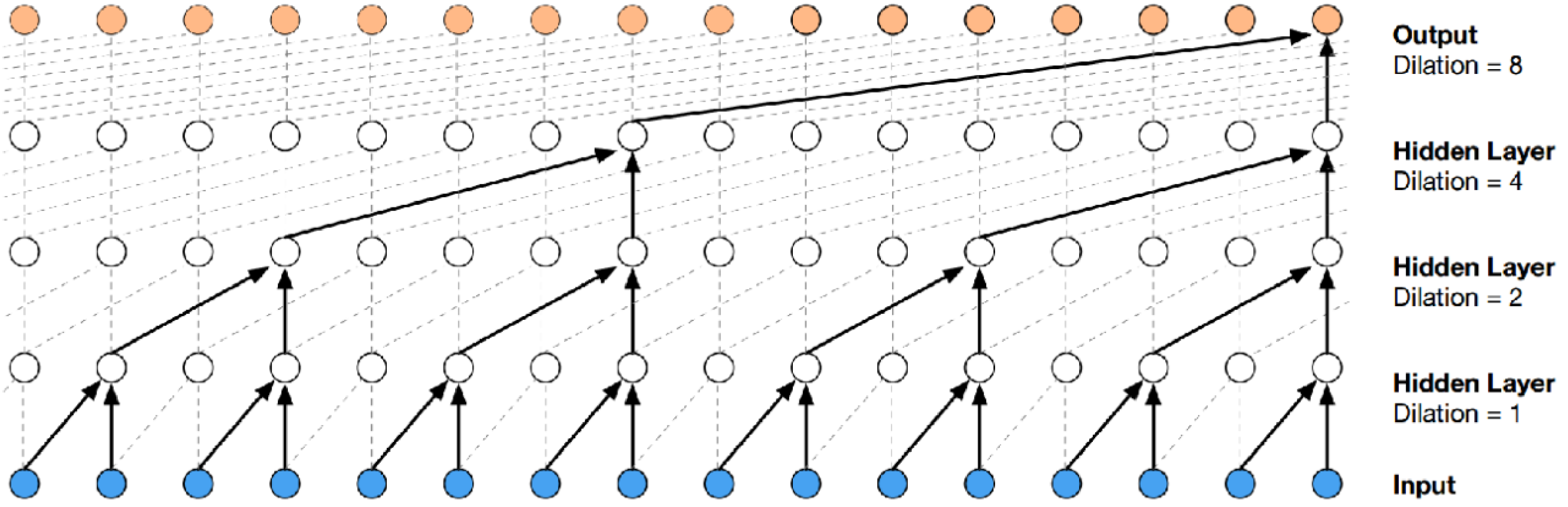
Atrous Convolution

- Small field of view cause accurate localization
- Large field of view cause to context assimilation



Dilated convolutions

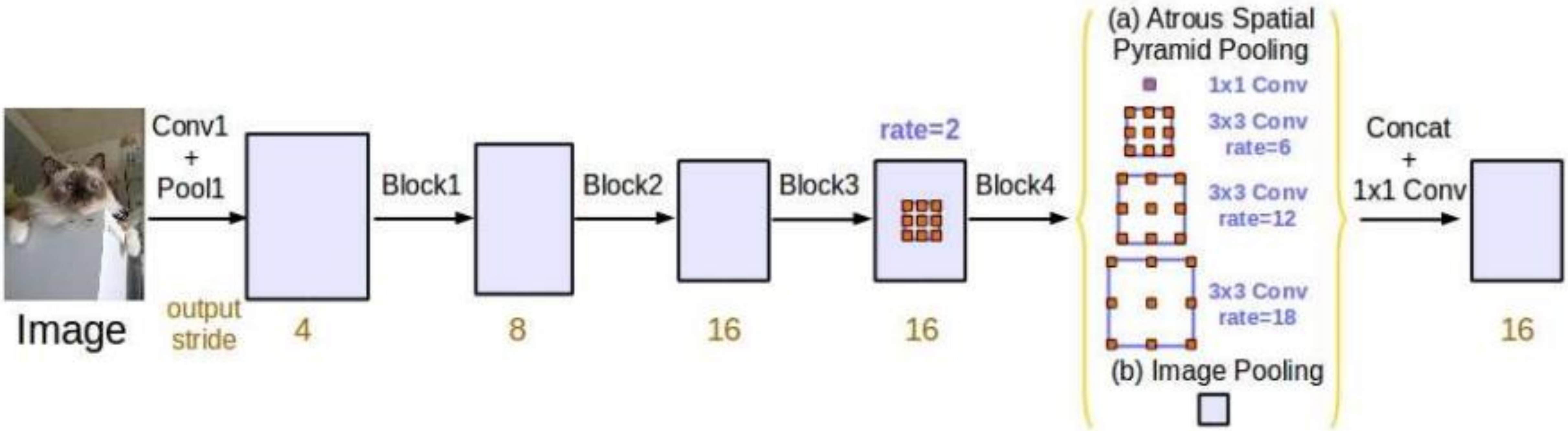
- "Multi-Scale Context Aggregation by Dilated Convolutions", Fisher Yu, Vladlen Koltun, **23 Nov, 2015**
- a.k.a stroud convolution, convolution with holes
- Enlarge the size of receptive field without losing resolution



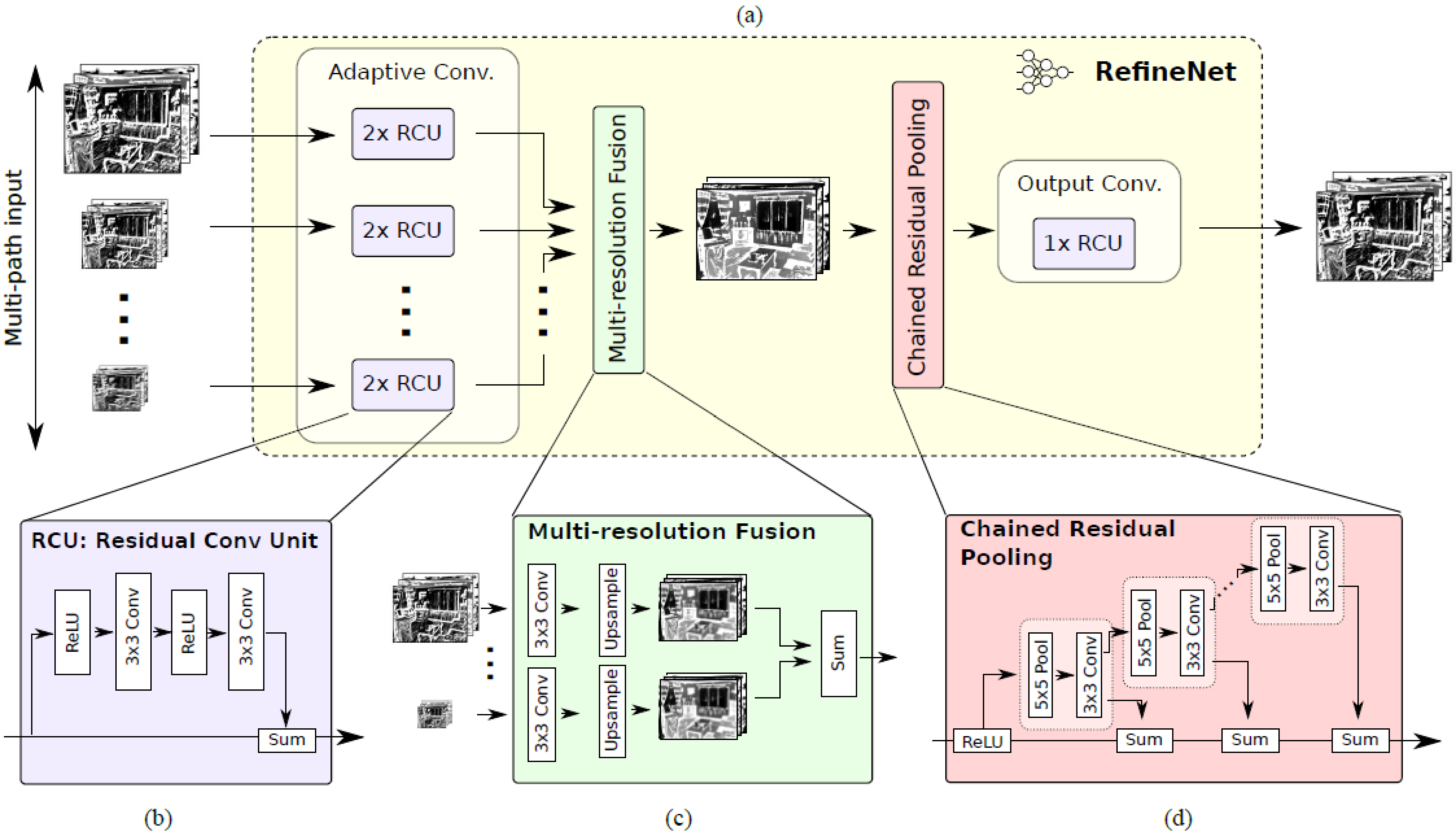
The figure is from "WaveNet: A Generative Model for Raw Audio"

Problems of ResNet and Dilated Convolution

DeepLab v3 Architecture

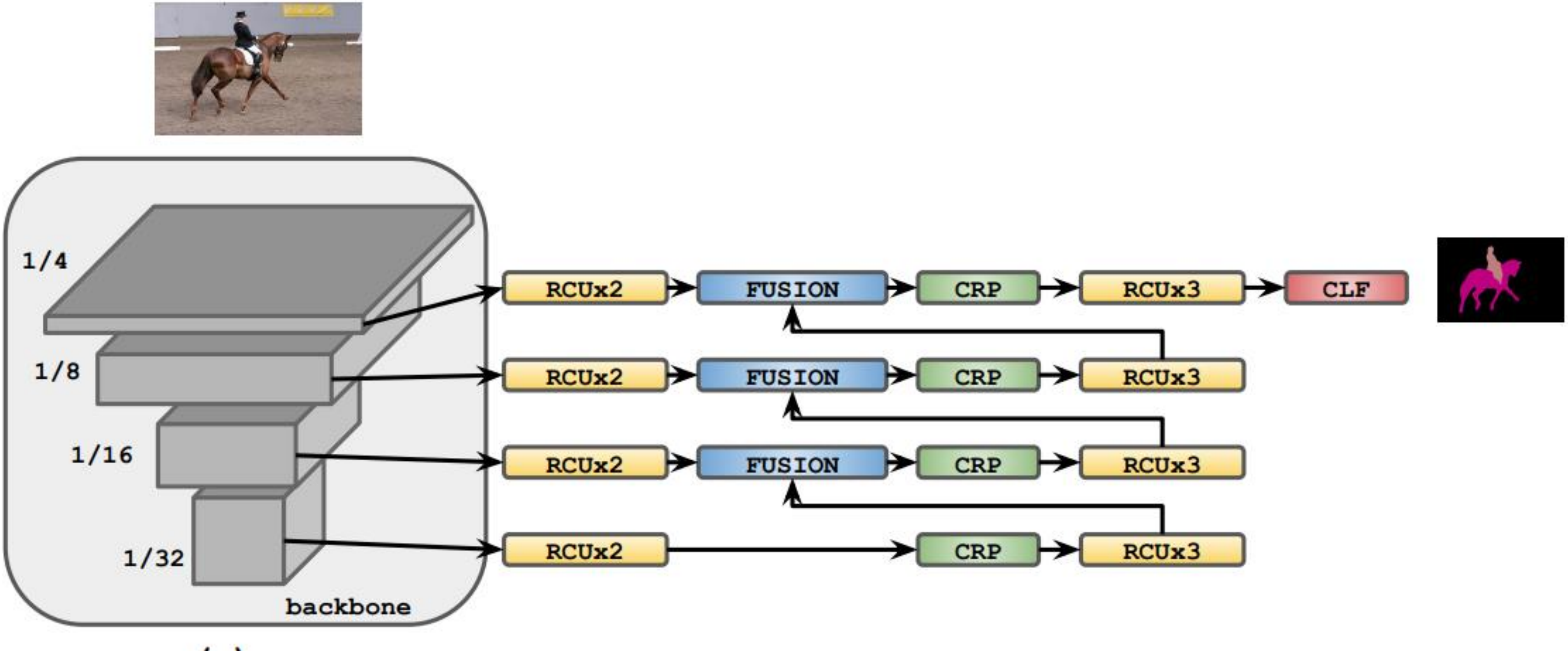


RefineNet



(a)
RefineNet
Multi-path input
Adaptive Conv.
2x RCU
2x RCU
2x RCU
Multi-resolution Fusion
Chained Residual Pooling
Output Conv.
1x RCU
RCU: Residual Conv Unit
ReLU
3x3 Conv
ReLU
3x3 Conv
Sum
Multi-resolution Fusion
3x3 Conv
Upsample
3x3 Conv
Upsample
Sum
Chained Residual Pooling
5x5 Pool
3x3 Conv
5x5 Pool
3x3 Conv
5x5 Pool
3x3 Conv
ReLU
Sum
Sum
Sum
(b)
(c)
(d)

RefineNet


1/4
1/8
1/16
1/32
backbone
RCUx2
FUSION
CRP
RCUx3
CLF
RCUx2
FUSION
CRP
RCUx3
RCUx2
FUSION
CRP
RCUx3
RCUx2
CRP
RCUx3

Overall Architecture of RefineNet

RefineNet uses the ResNet as the backbone. Along the ResNet, different resolutions of feature maps go through **Residual Conv Unit** (RCU). Pre-Activation ResNet is used.

Overall Architecture of RefineNet

RefineNet uses the ResNet as the backbone. Along the ResNet, different resolutions of feature maps go through **Residual Conv Unit** (RCU). Pre-Activation ResNet is used.

RCU: Residual block is used but with batch normalization removed.

Overall Architecture of RefineNet

RefineNet uses the ResNet as the backbone. Along the ResNet, different resolutions of feature maps go through **Residual Conv Unit** (RCU). Pre-Activation ResNet is used.

RCU: Residual block is used but with batch normalization removed.

Fusion: Then multi-resolution fusion is used to merge the feature maps using element-wise summation.

Overall Architecture of RefineNet

RefineNet uses the ResNet as the backbone. Along the ResNet, different resolutions of feature maps go through **Residual Conv Unit** (RCU). Pre-Activation ResNet is used.

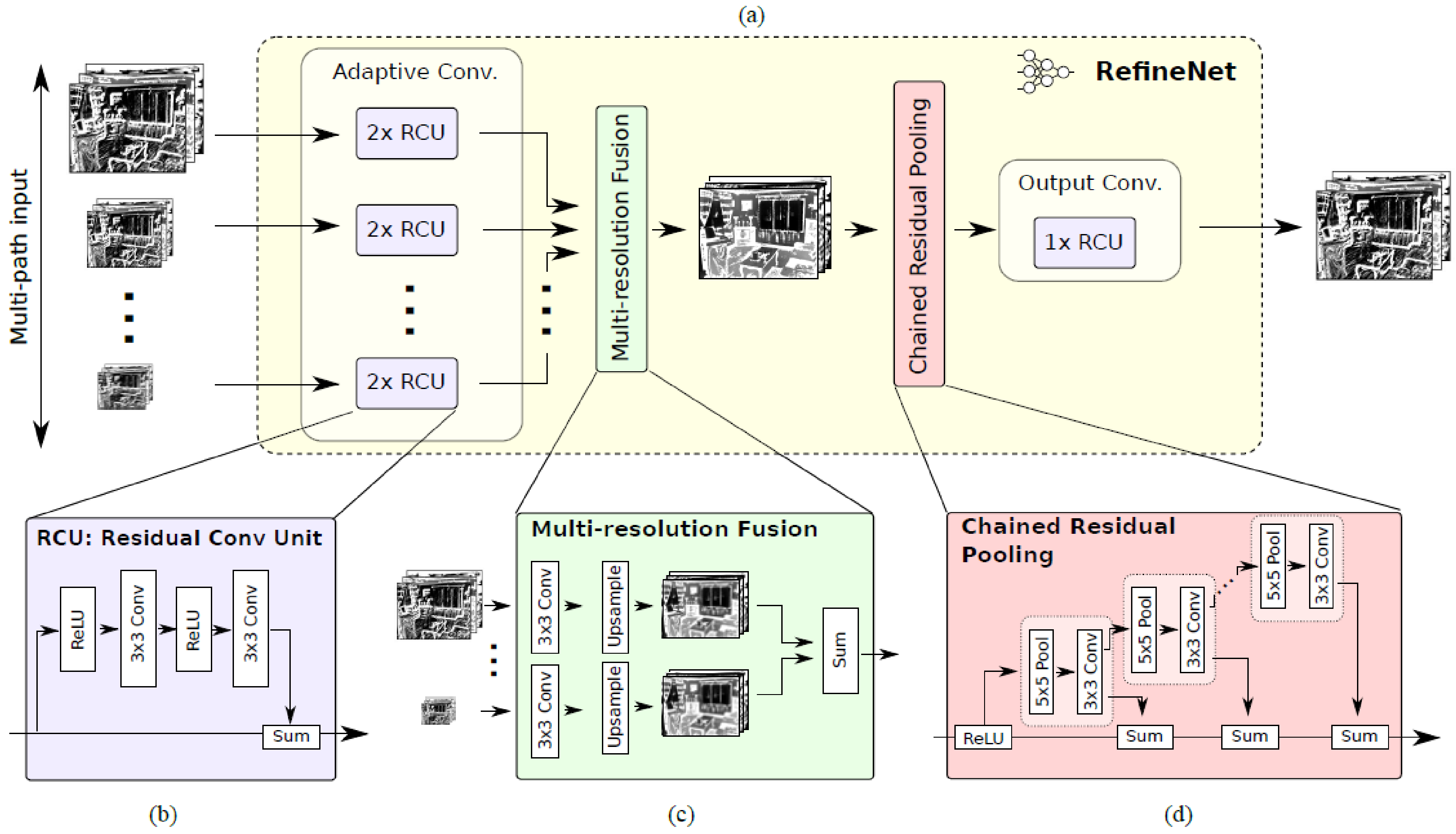
RCU: Residual block is used but with batch normalization removed.

Fusion: Then multi-resolution fusion is used to merge the feature maps using element-wise summation.

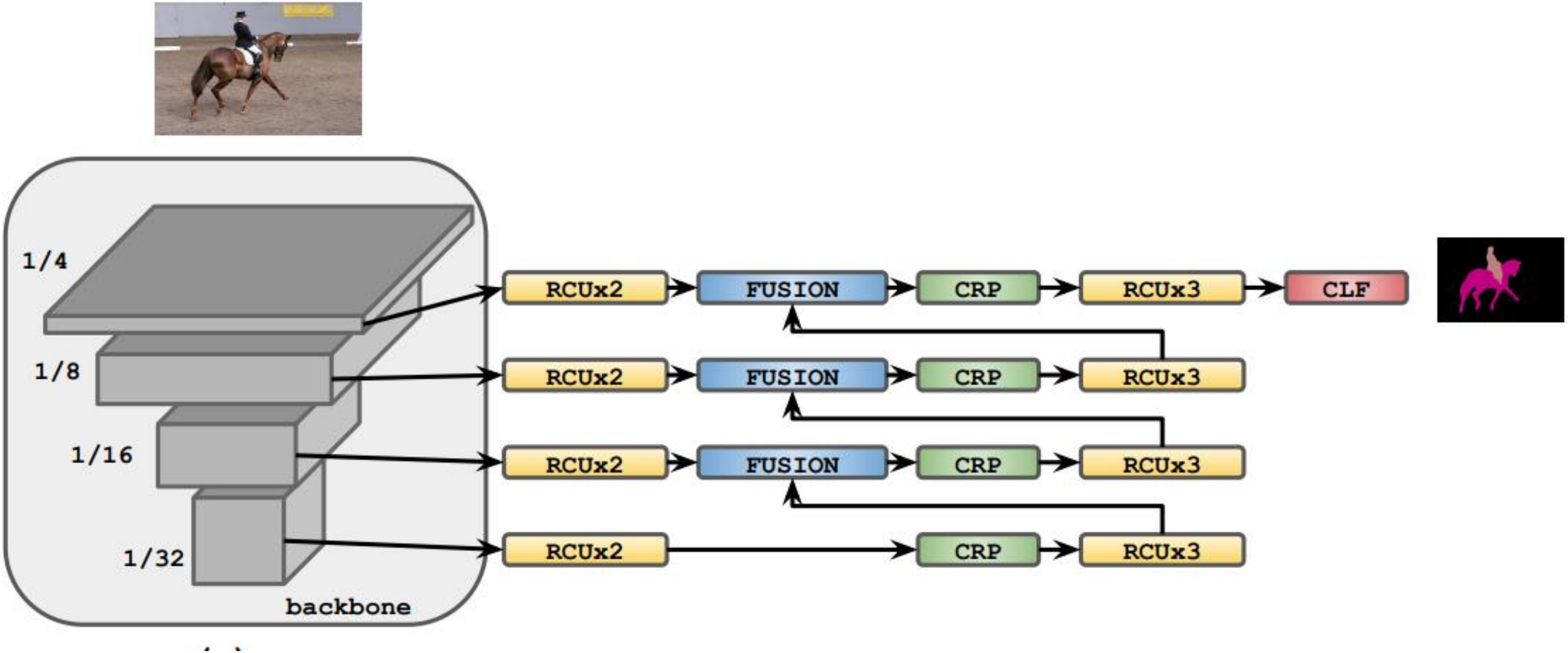
Chained Residual Pooling: The output feature maps of all pooling blocks are fused together with the input feature map through summation of residual connections. It **aims to capture background context from a large image region.**

Overall Architecture of RefineNet

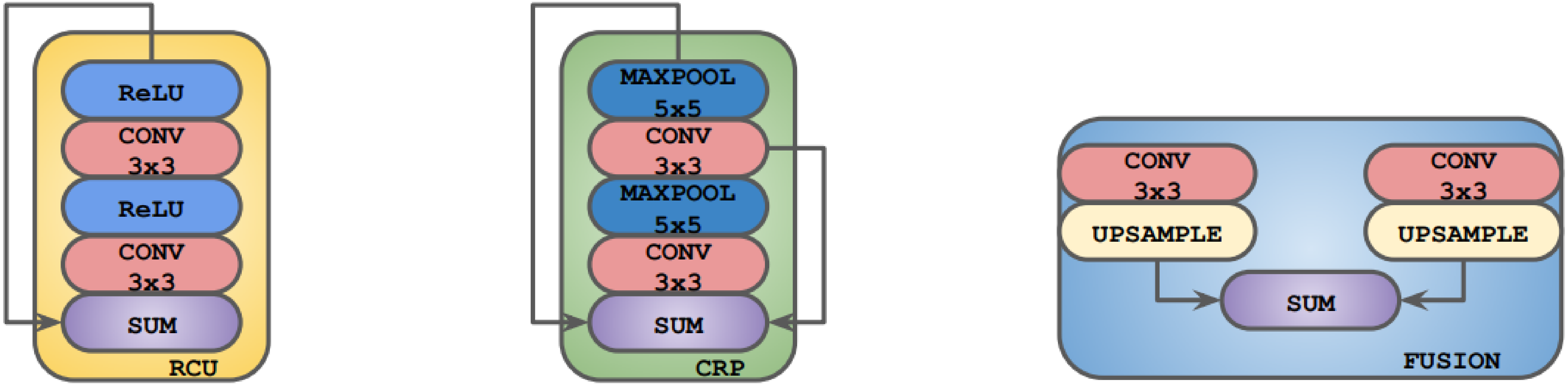
Output Conv: Another RCU is placed to employ non-linearity operations on the multi-path fused feature maps to generate features for further processing or for final prediction.



Overall Architecture of RefineNet



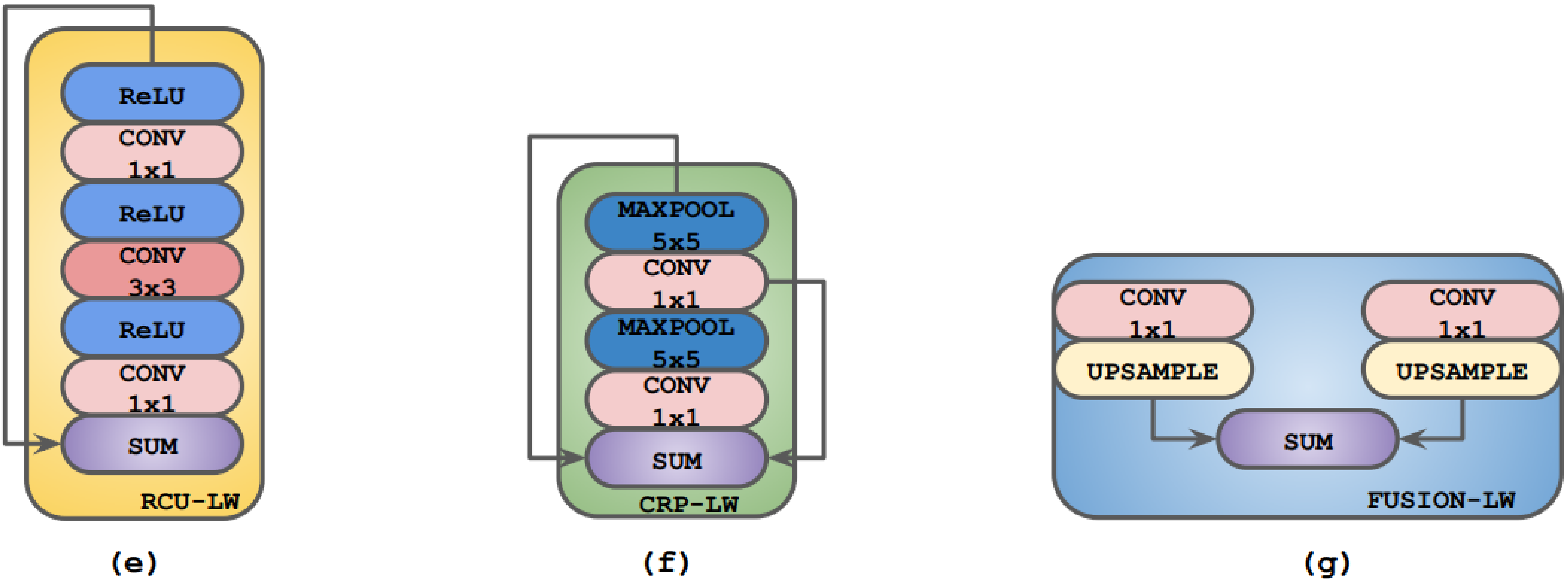
Overall Architecture of RefineNet



(b) (c) (d)

(b)-(d) general outline of original RCU, CRP and fusion blocks

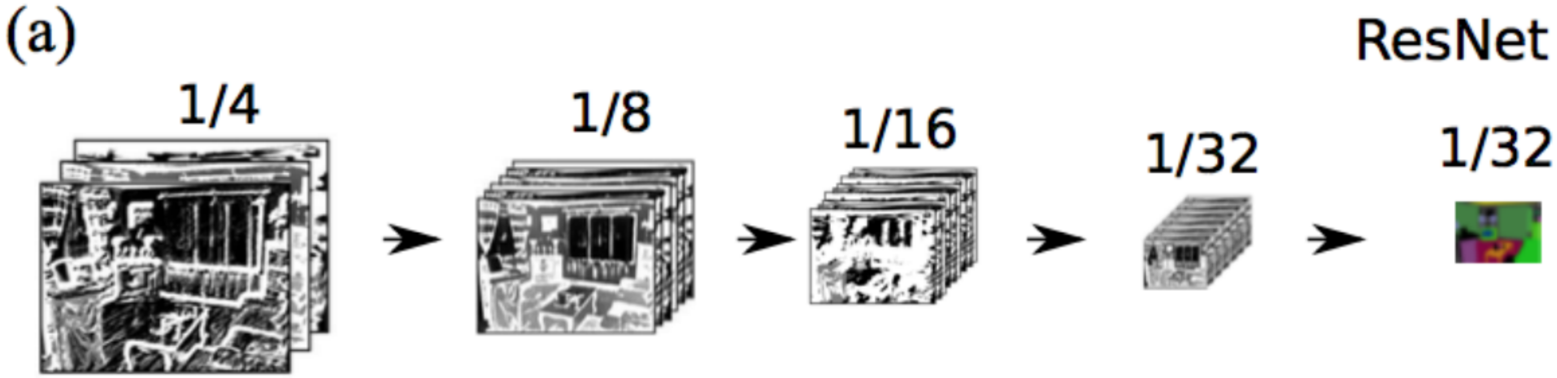
Overall Architecture of RefineNet

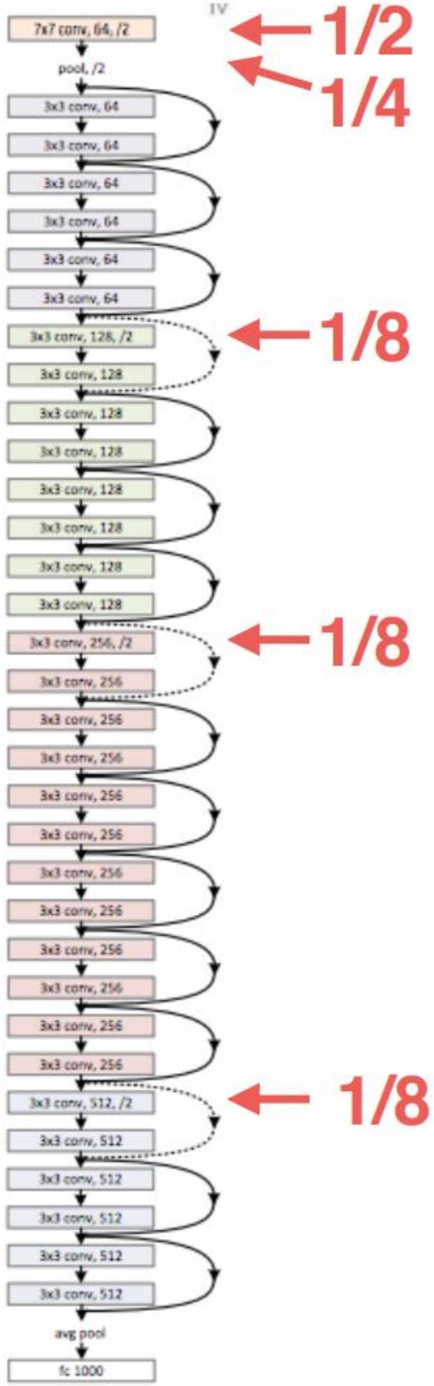


(e)-(g) light-weight RCU, CRP and fusion blocks. In the interests of brevity, we only visualize 2 convolutional layers for the CRP blocks (instead of 4 used in the original architecture).

Dilated convolutions

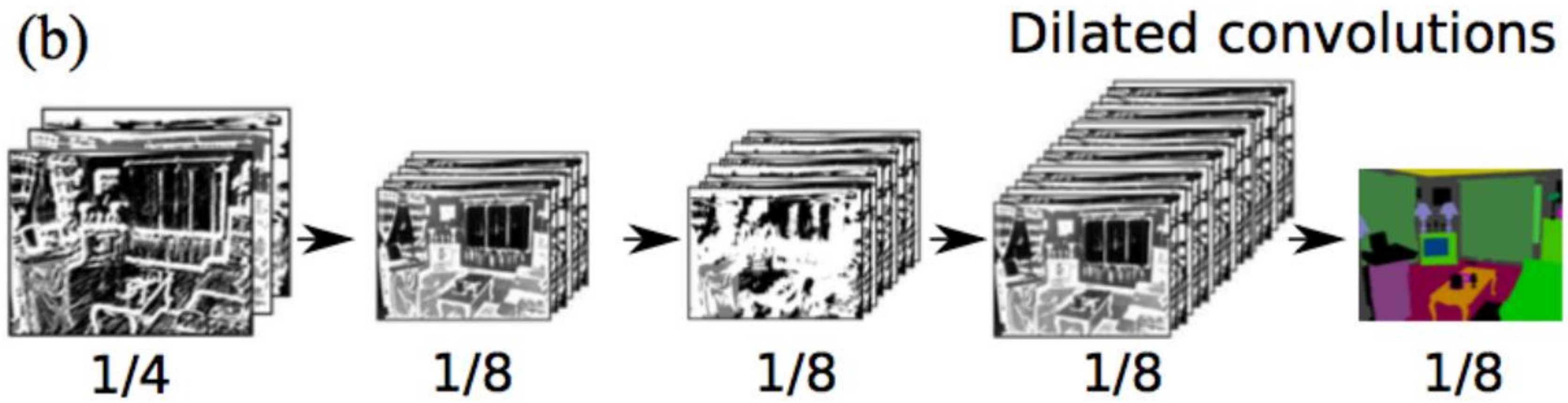
- For example, the feature maps of ResNet are downsampled 5 times, and 4 times in the 5 are done by convolutions with stride of 2 (only the first one is by pooling with stride of 2)

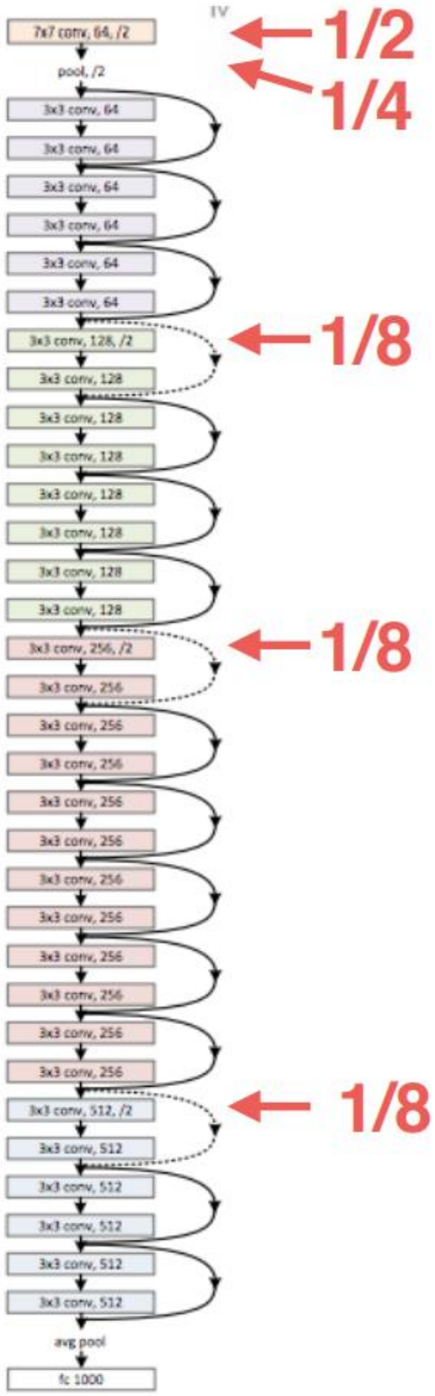




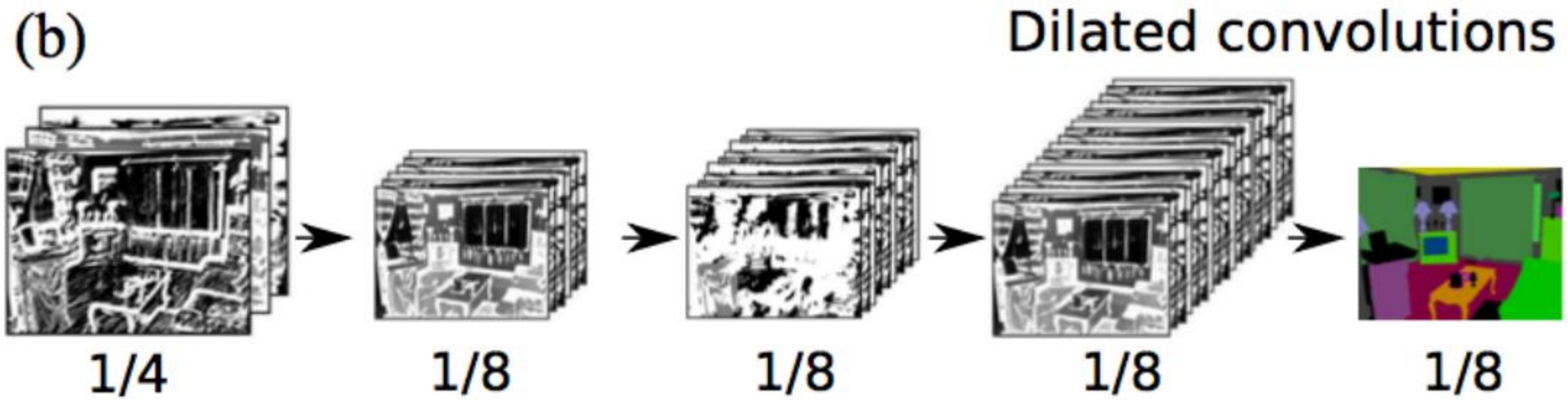
Dilated convolutions

- By using dilated convolutions instead of vanilla convolutions, the resolution after the first pooling can be kept as the same to the end





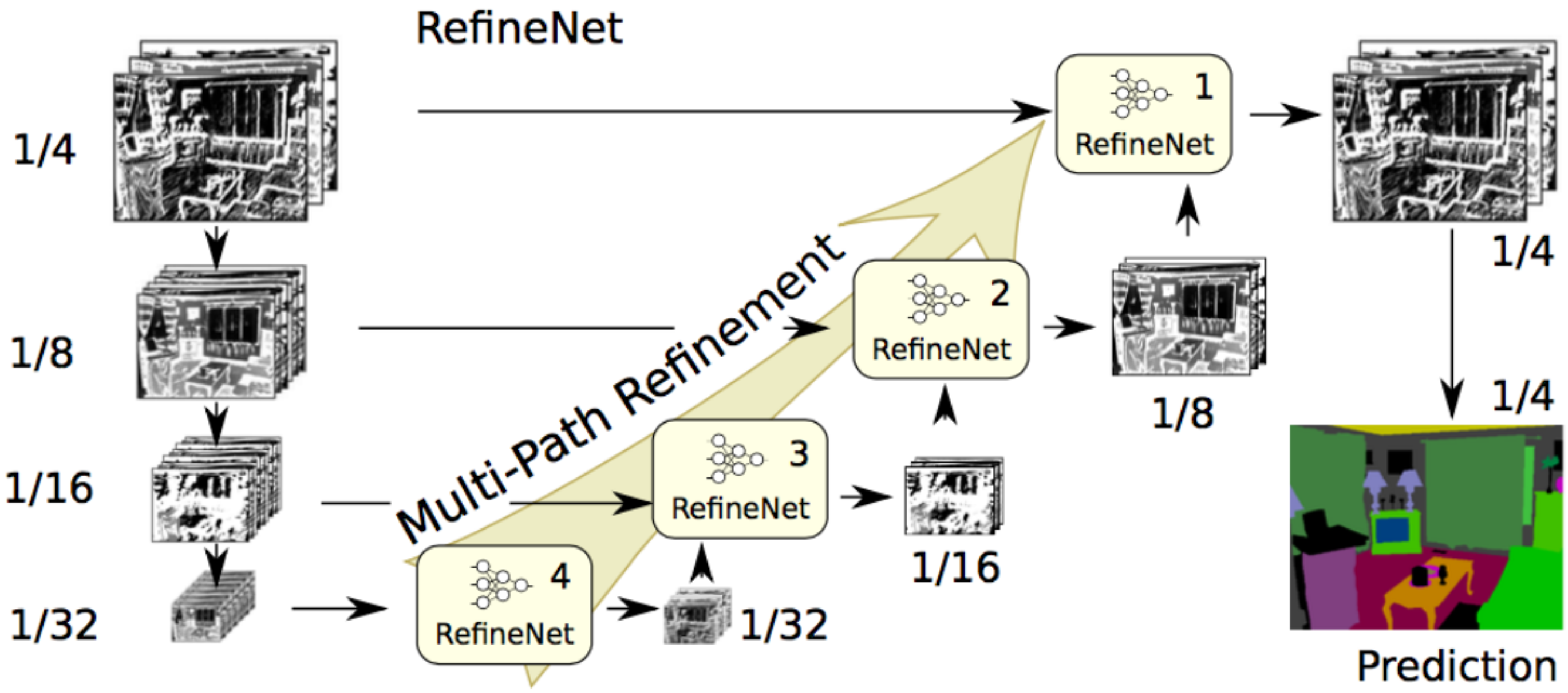
Dilated convolutions



But, it is still 1/8…

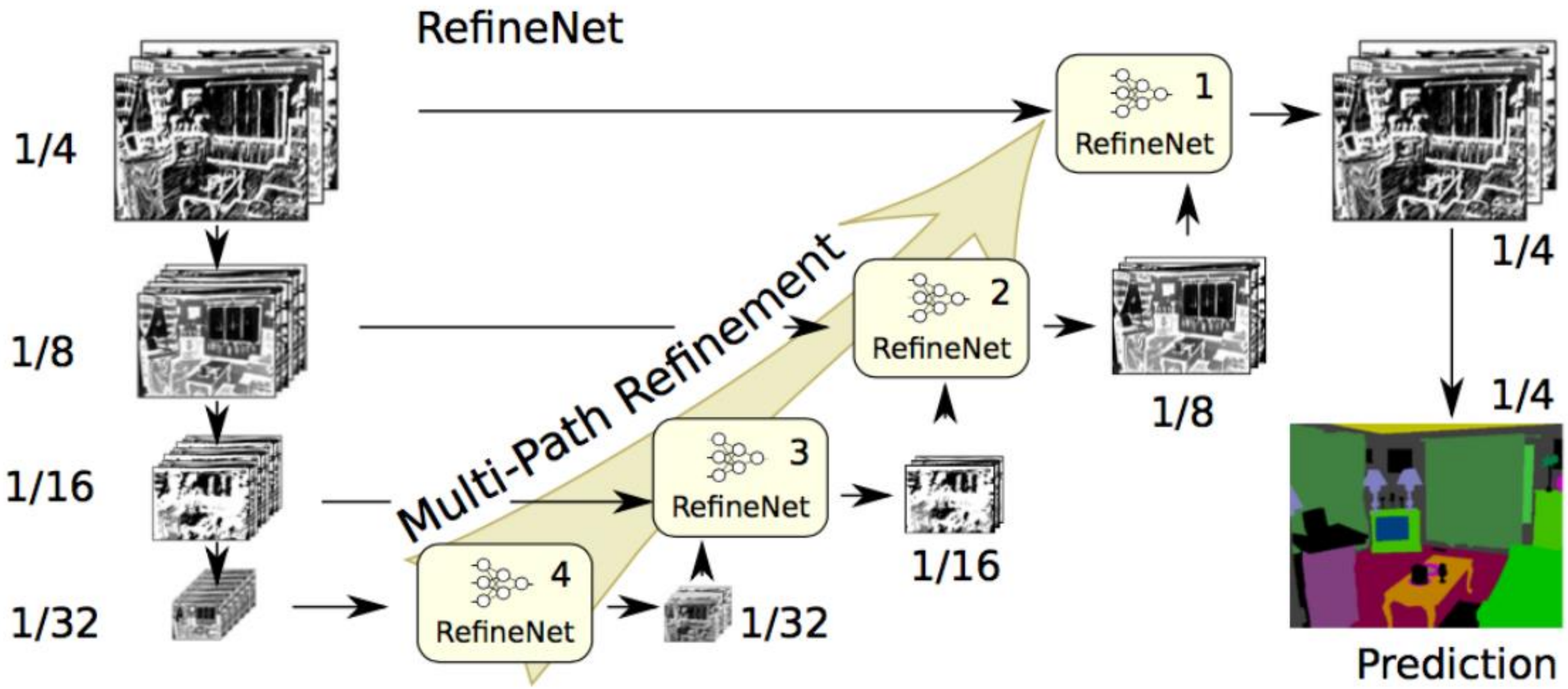
RefineNet

- “RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation”, Guosheng Lin, Anton Milan, Chunhua Shen, Ian Reid, **20 Nov. 2016**

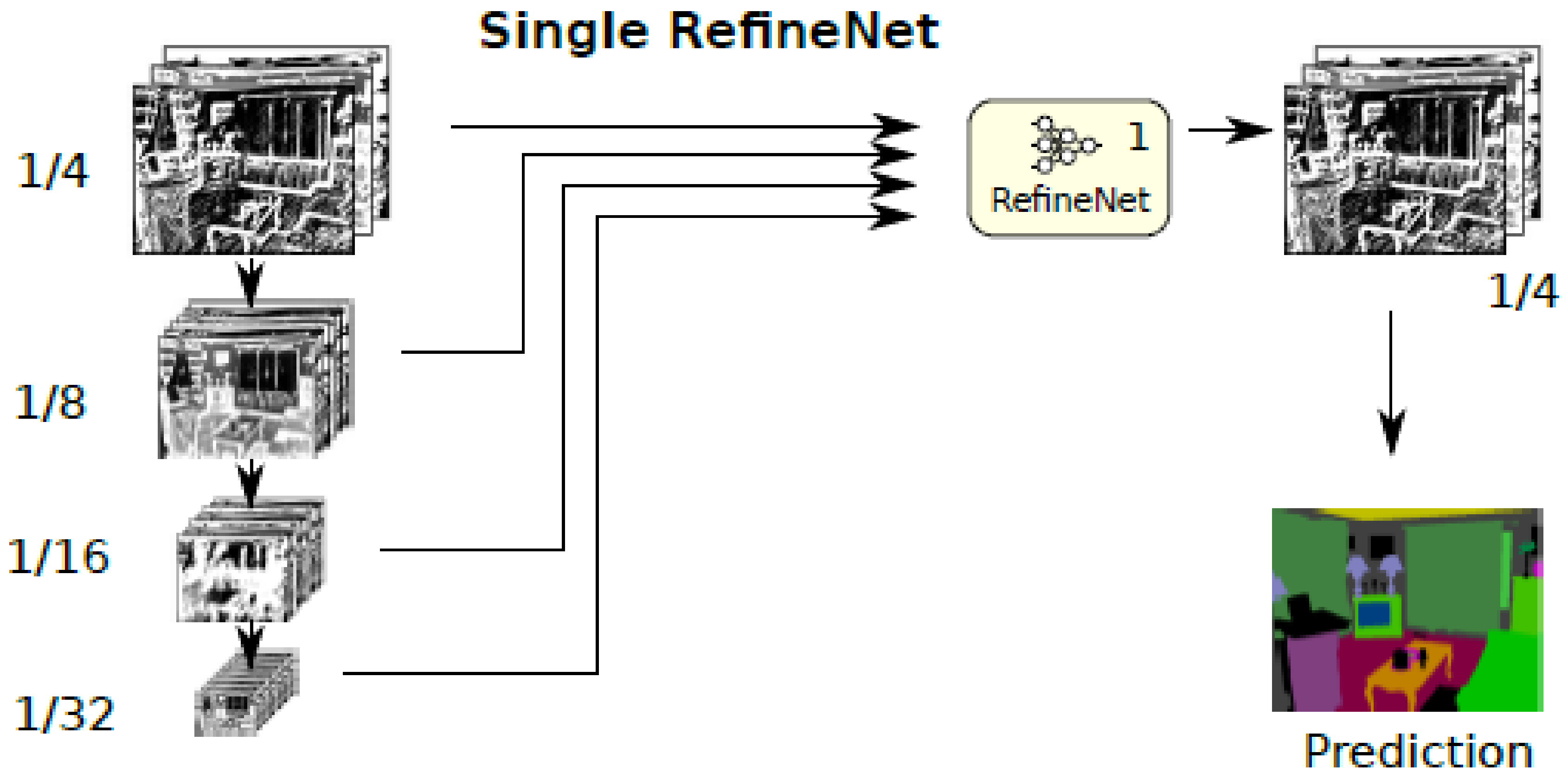


RefineNet

- Each intermediate feature map is refined through "RefineNet module"



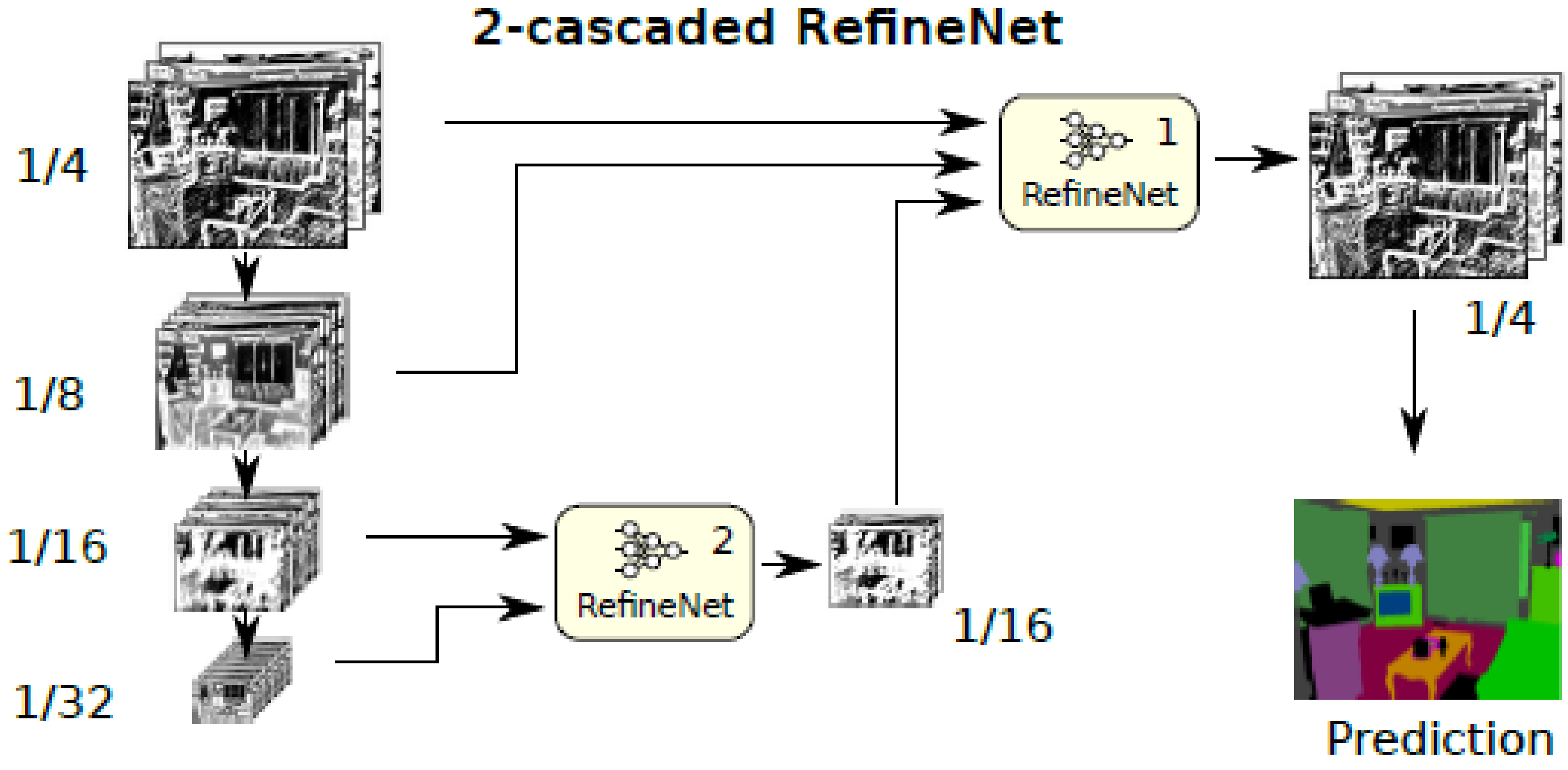
Different RefineNet Variants



(a)

Single RefineNet model: It takes all four inputs from the four blocks of ResNet and fuses all-resolution feature maps in a single process.

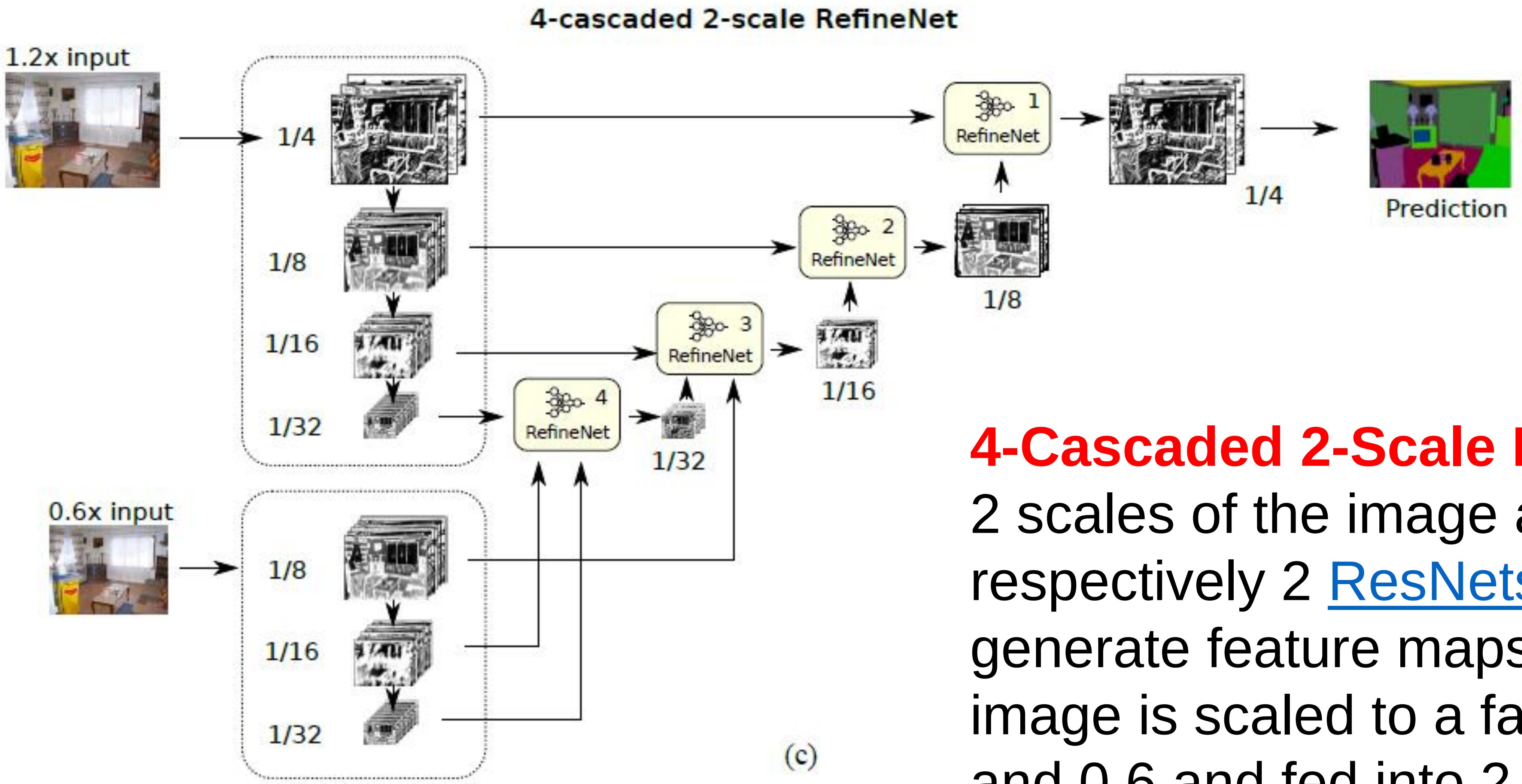
Different RefineNet Variants



(b)

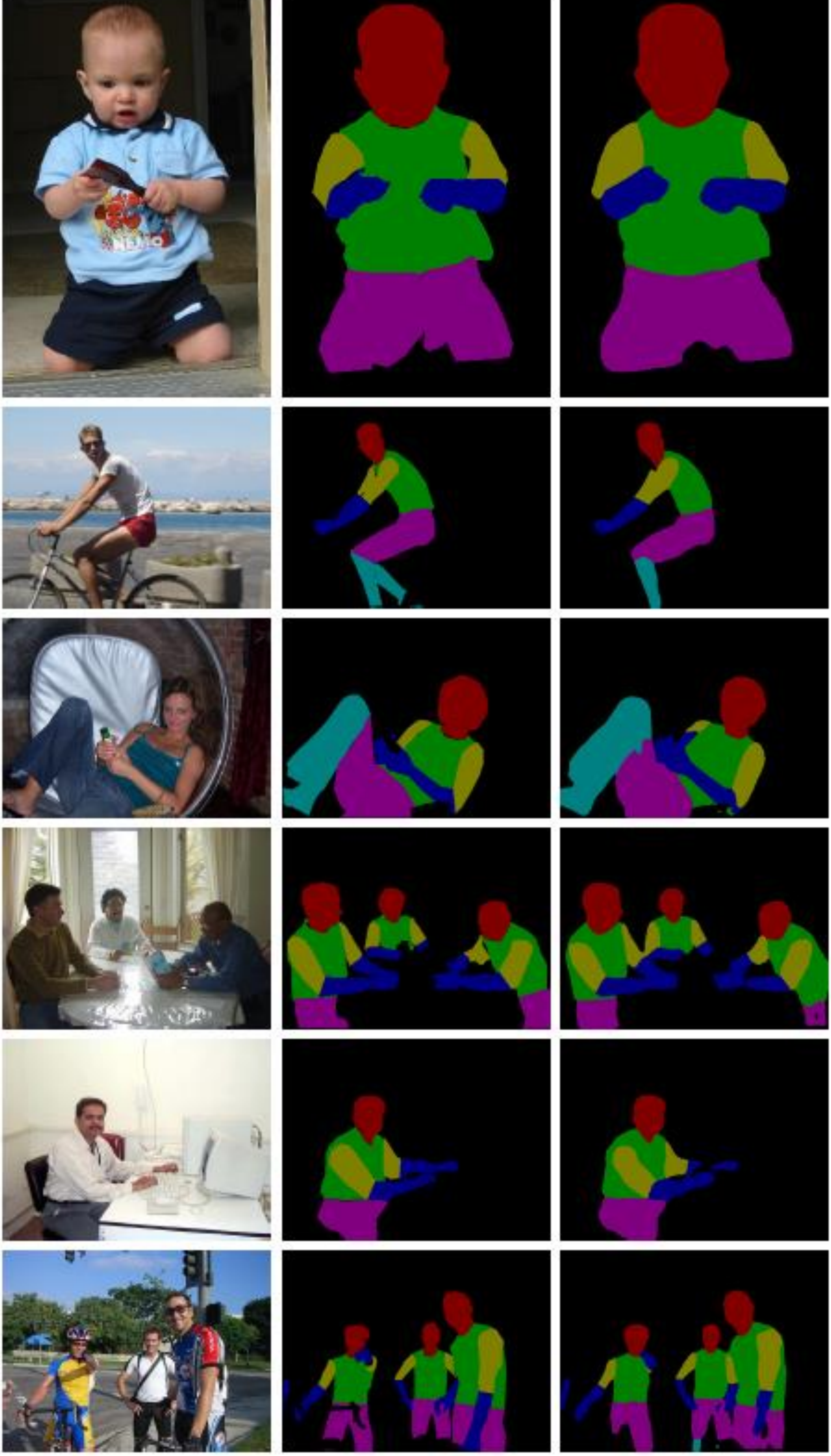
2-Cascaded RefineNet: It employs only two RefineNet modules instead of four. The bottom one, RefineNet-2, has two inputs from ResNet blocks 3 and 4, and the other one has three inputs, two coming from the remaining ResNet blocks and one from RefineNet-2.

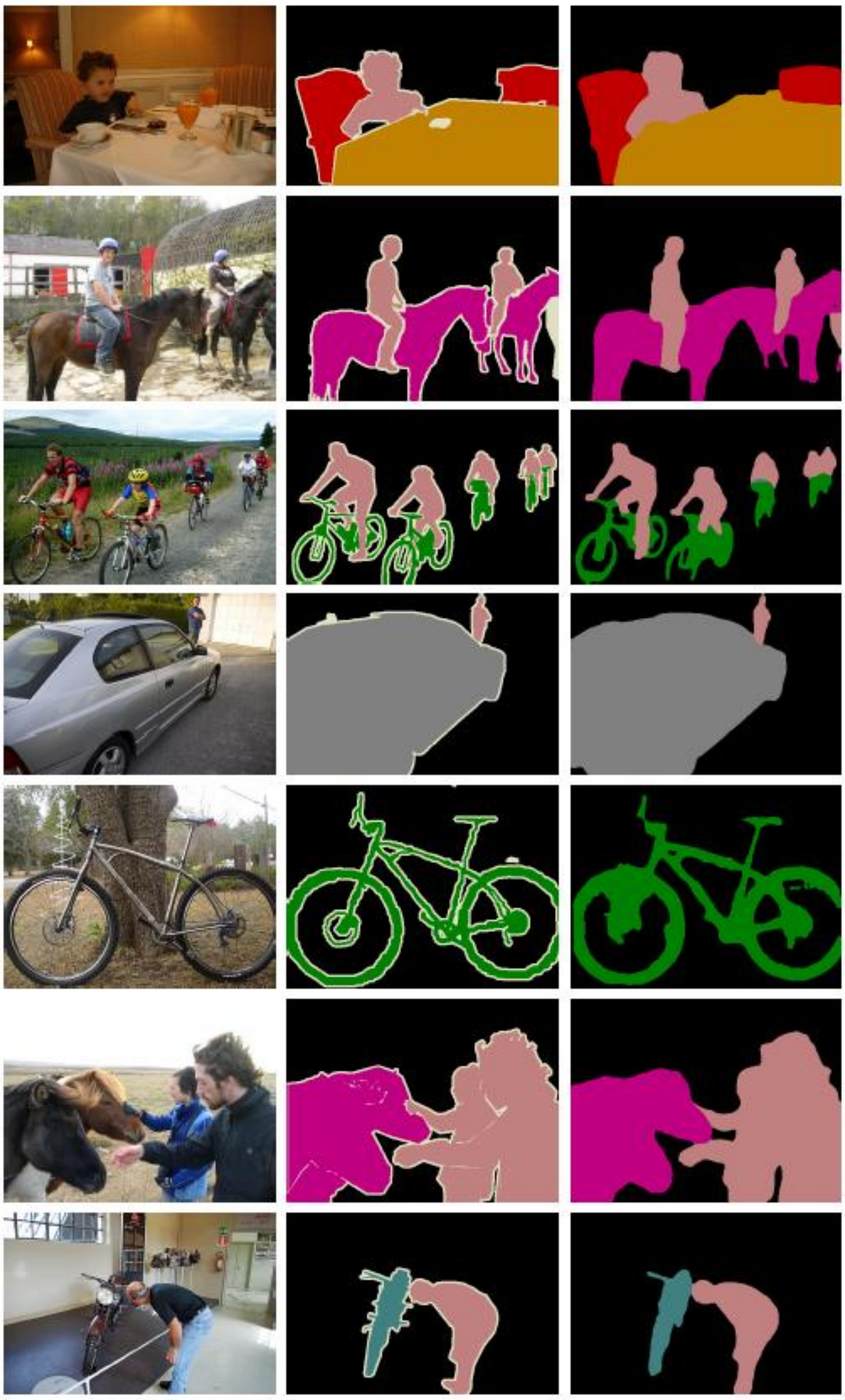
Different RefineNet Variants



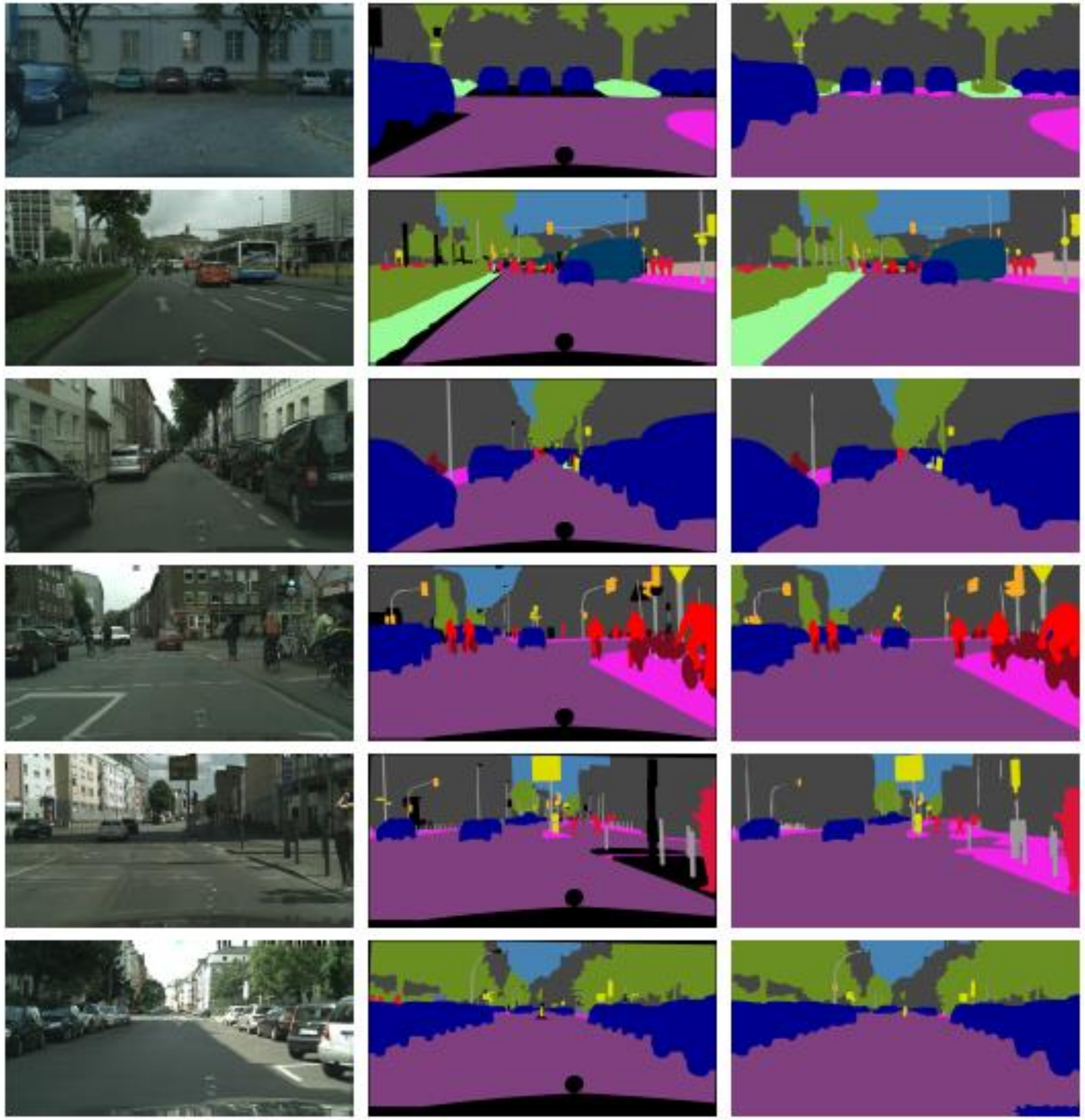
4-Cascaded 2-Scale RefineNet has the best results due to the larger capacity of the network, but it also results in longer training times.

4-Cascaded 2-Scale RefineNet: 2 scales of the image as input and respectively 2 ResNets to generate feature maps. The input image is scaled to a factor of 1.2 and 0.6 and fed into 2 independent ResNets.


(a) Test Image
(b) Ground Truth
(c) Prediction



(a) Test Image (b) Ground Truth (c) Prediction



(a) Test Image (b) Ground Truth (c) Prediction

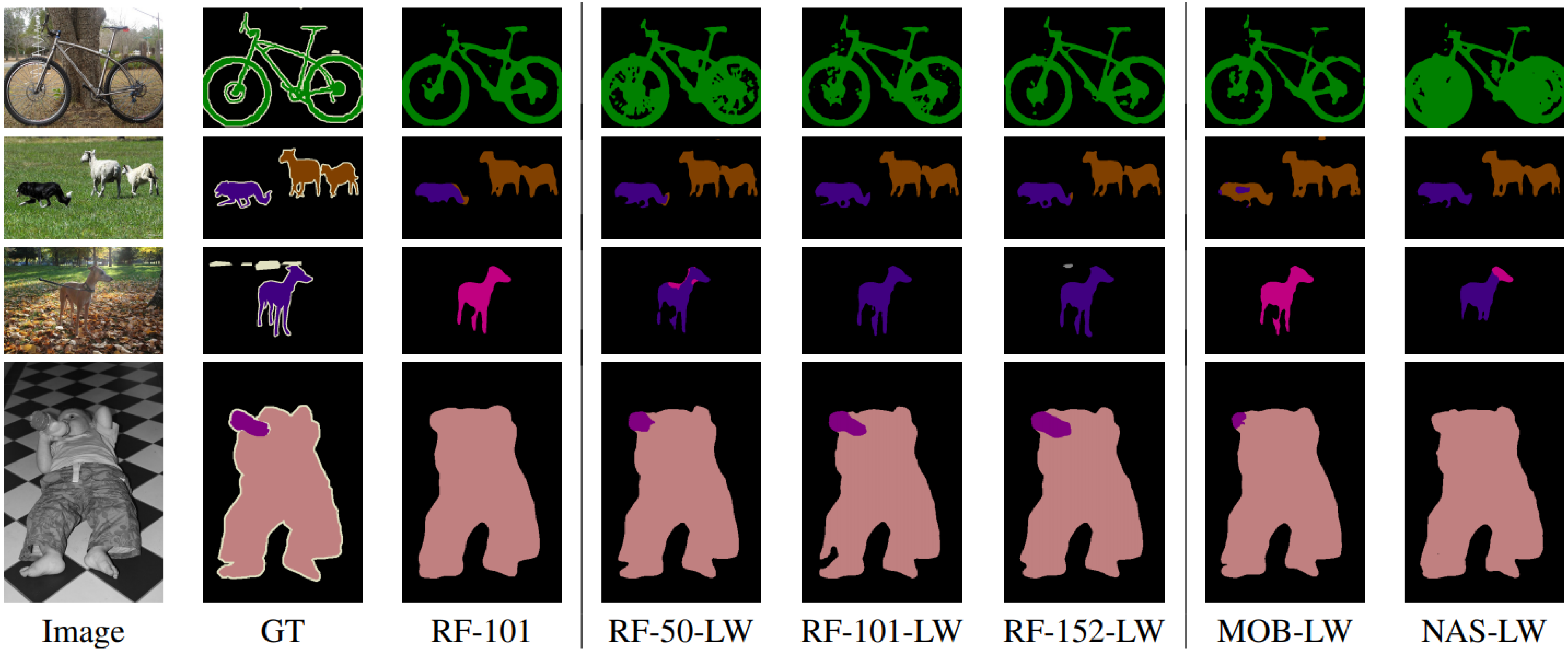


Image
GT
RF-101
RF-50-LW
RF-101-LW
RF-152-LW
MOB-LW
NAS-LW