Lecture 8:
Image Segmentation

彭　超 Peng Chao

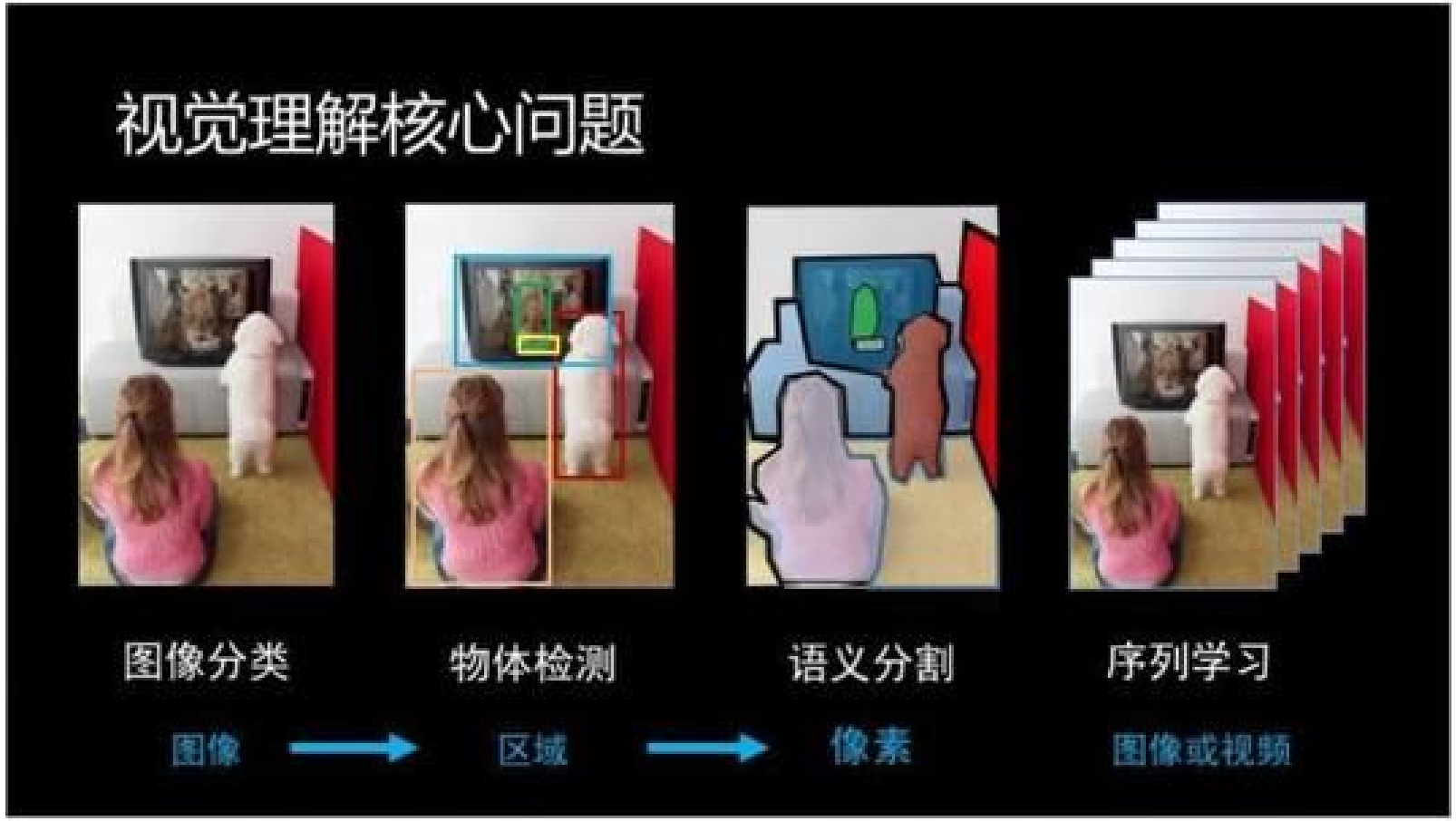
Face++ Researcher

pengchao@megvii.com

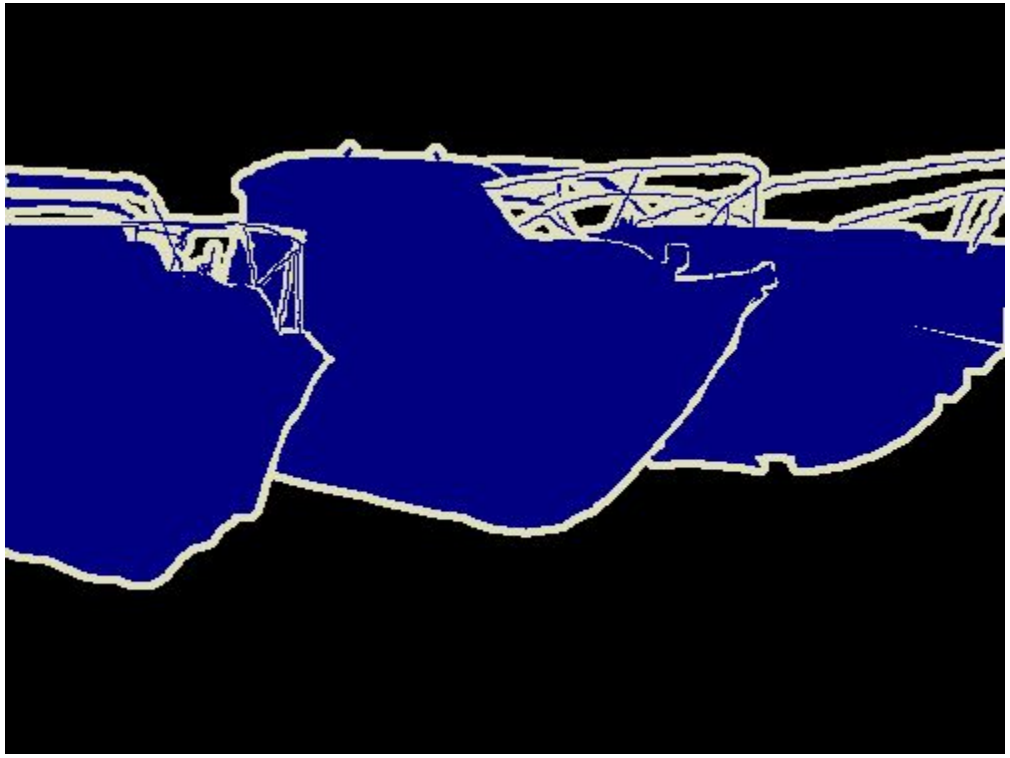
Nov. 2017

Face++ 旷视

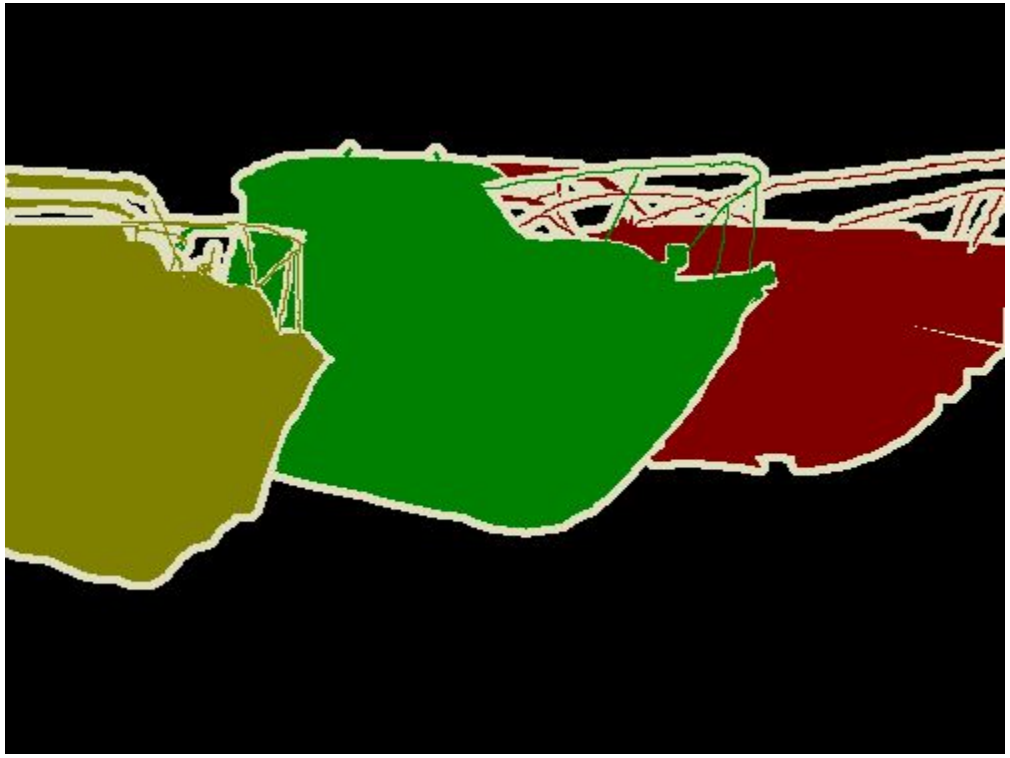
Image Segmentation



Semantic Segmentation



Instance Segmentation



Face++ 旷视

Scene Parsing

Face++ 旷视

Human Parsing

Face++ 旷视

Stuff Segmentation

- New Track in COCO 2017
- Stuff: mountain, grass, wall, sky

  ……
- Stuff covers about 66% of the pixels in COCO

Face++ 旷视

UlrtraSound Segmentation

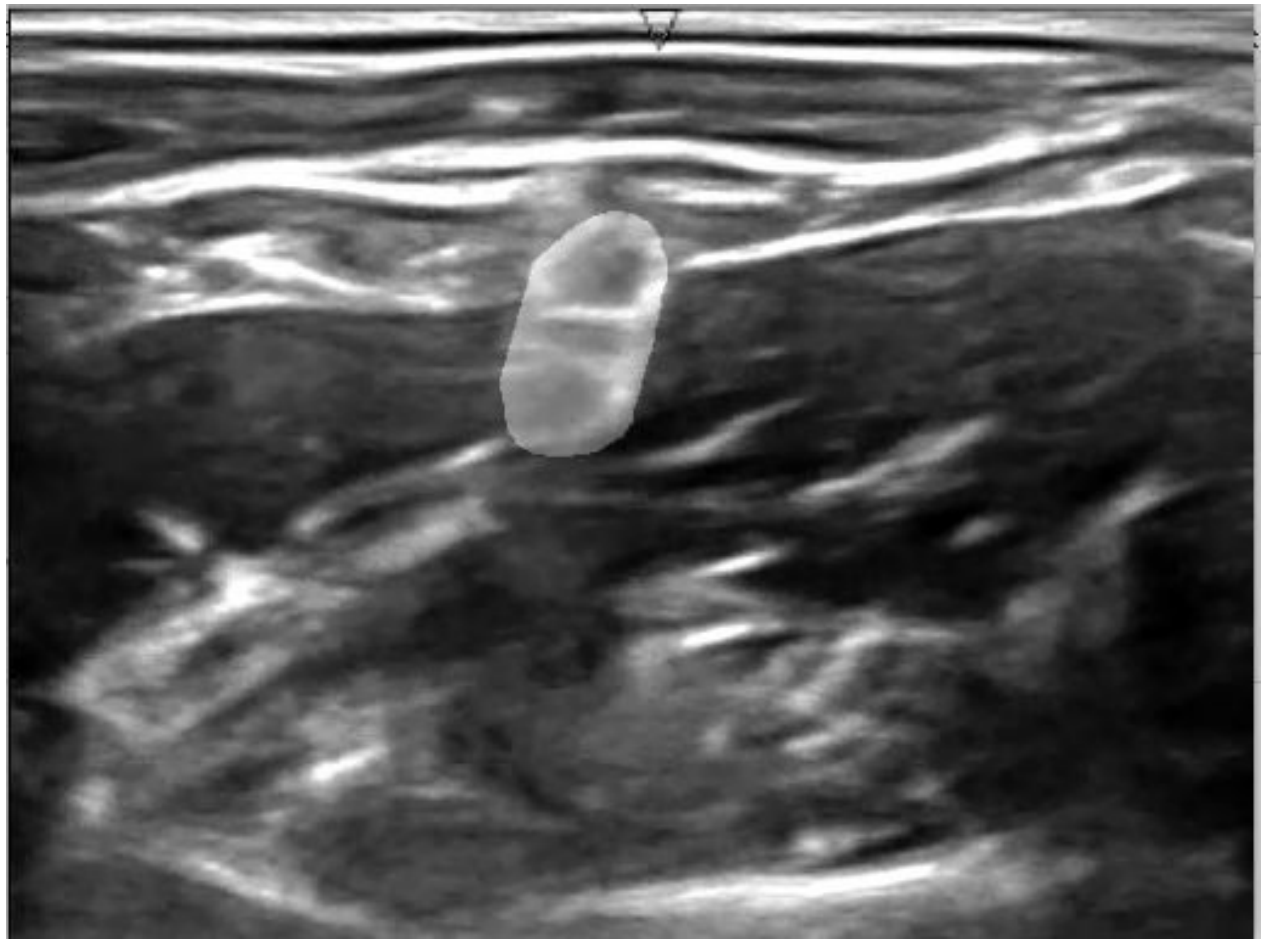


Figure credit: Ultrasound Nerve Segmentation on Kaggle

Selfie Segmentation

Face++ 旷视

Evaluation

$$Accuracy(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{i=0}^{n} \frac{I[y_i = \hat{y}_i]}{n}$$

$$mean\ IOU(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\sum_c^m \sum_i^n I[y_i = c, \hat{y}_i = c]}{\sum_c^m \sum_i^n I[y_i = c\ \mathbf{or}\ \hat{y} = c]}$$

Normally, we use mean IOU to judge the results!

Face++ 旷视

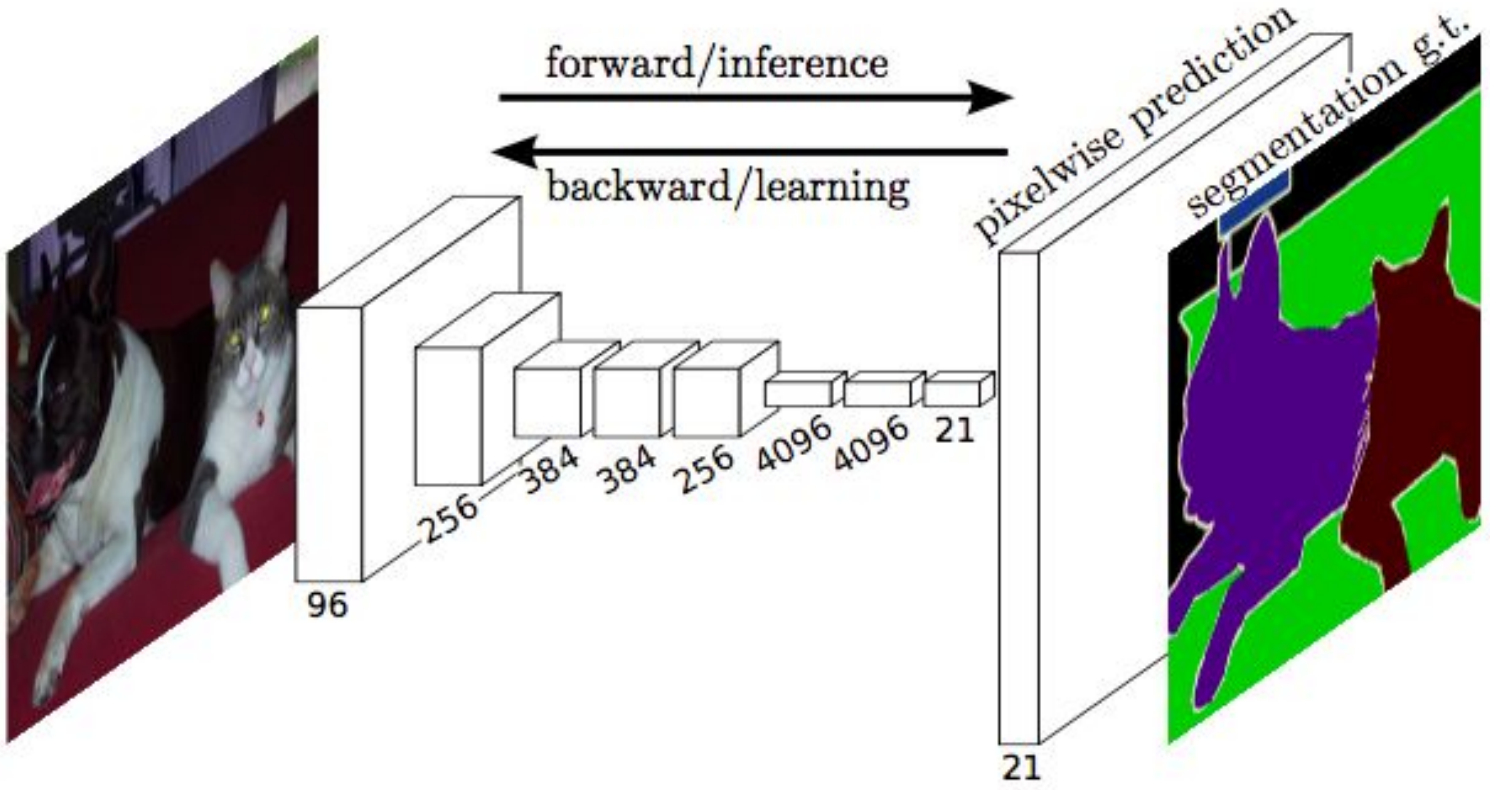
Outline

- Semantic Segmentation
- Instance Segmentation

Outline

- Semantic Segmentation
- Instance Segmentation

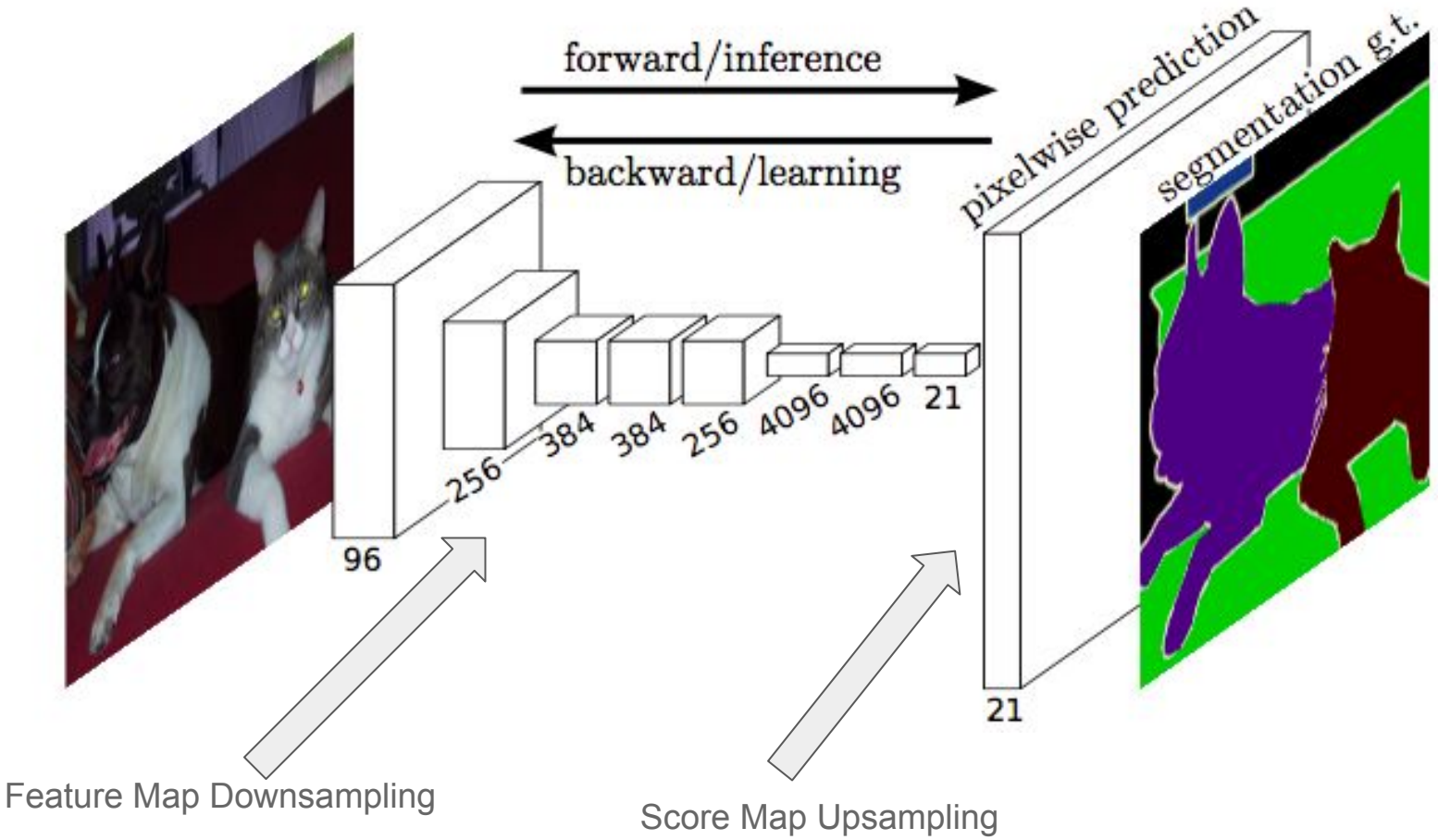
Fully Convolutional Network



Face++ 旷视

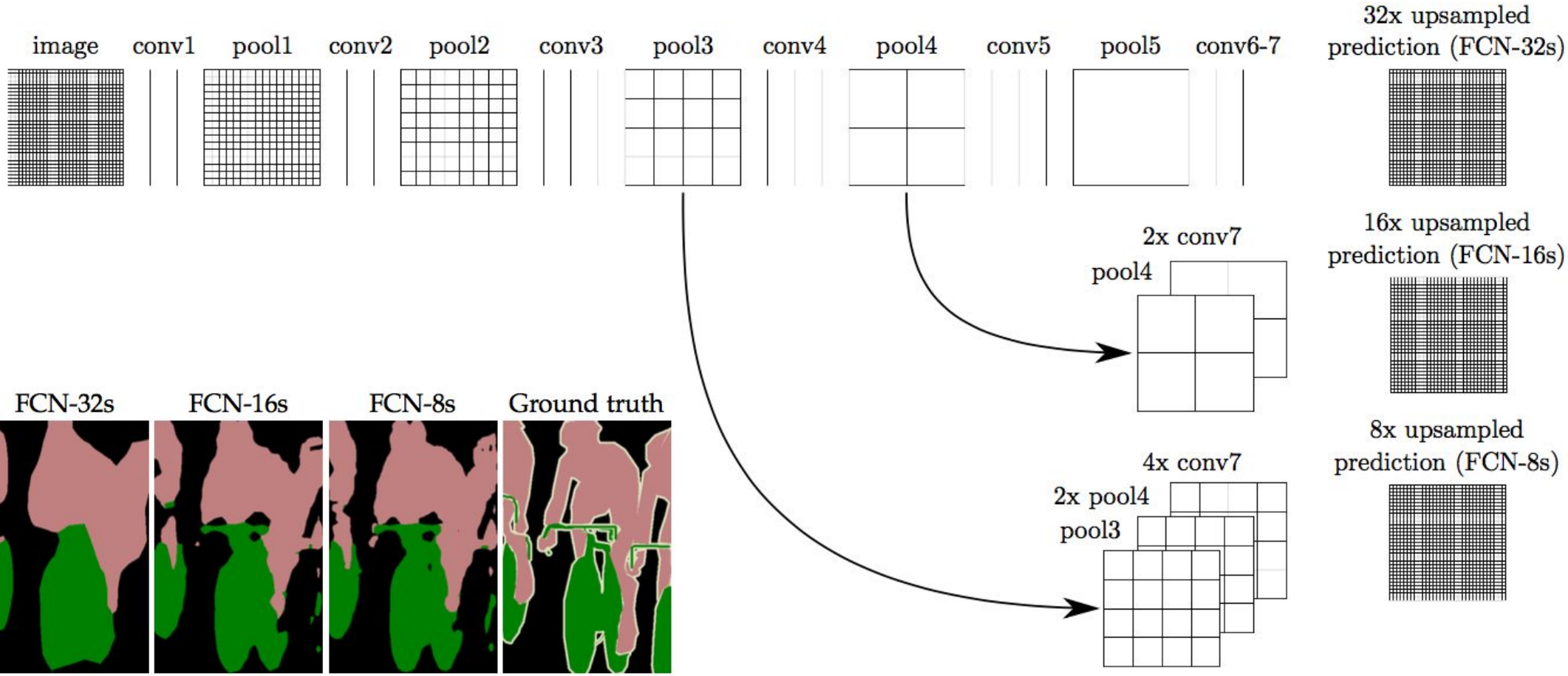
Long, Shelhamer, and Darrell, “Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015

Fully Convolutional Network



Long, Shelhamer, and Darrell, “Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015

Fully Convolutional Network



Face++ 旷视

Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015

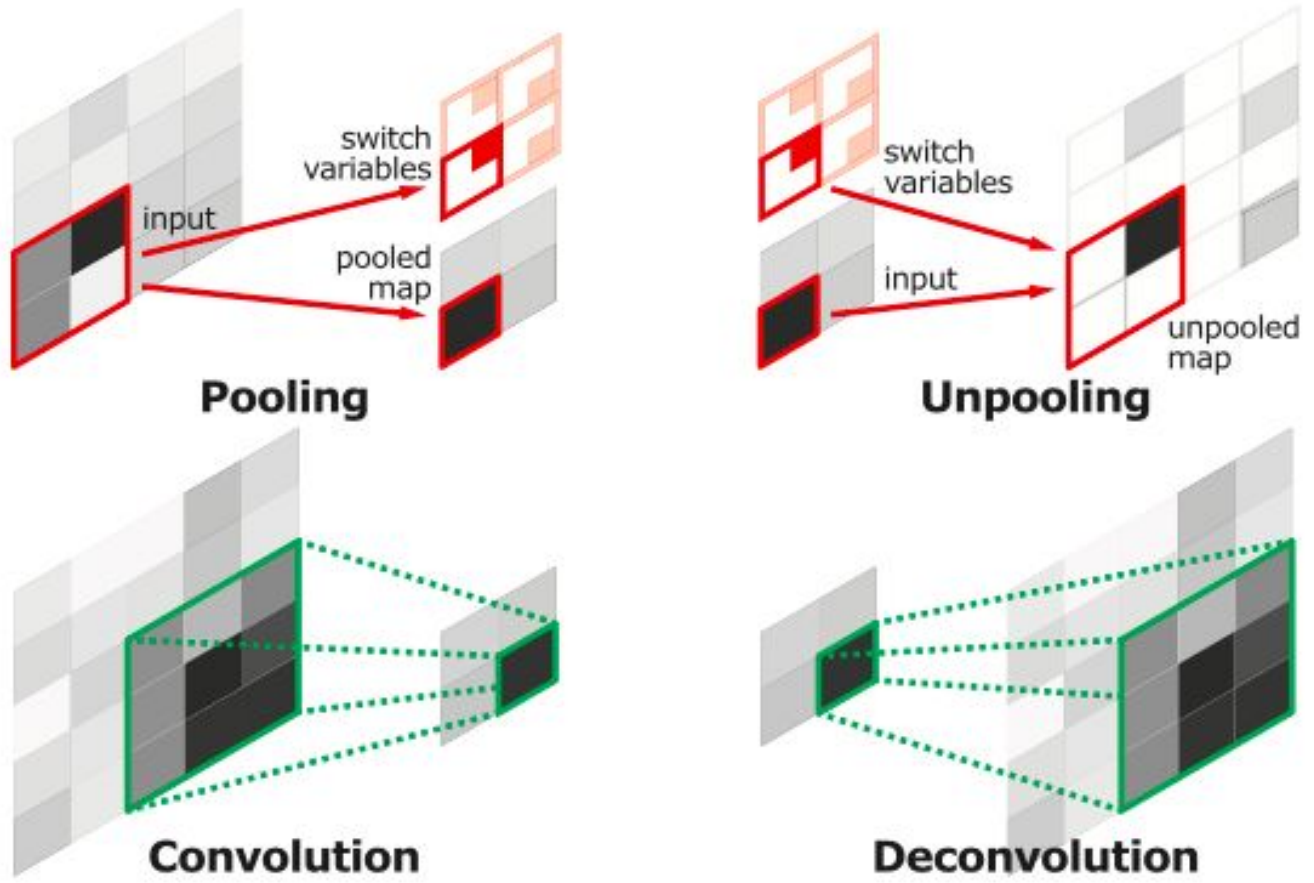
Fully Convolutional Network

- First work using CNN to solve the semantic segmentation
- Introducing skip-net framework
- Large Improvement! (60 vs 30)

Face++ 旷视

Long, Shelhamer, and Darrell, “Fully Convolutional Networks for Semantic Segmentation”, CVPR 2015

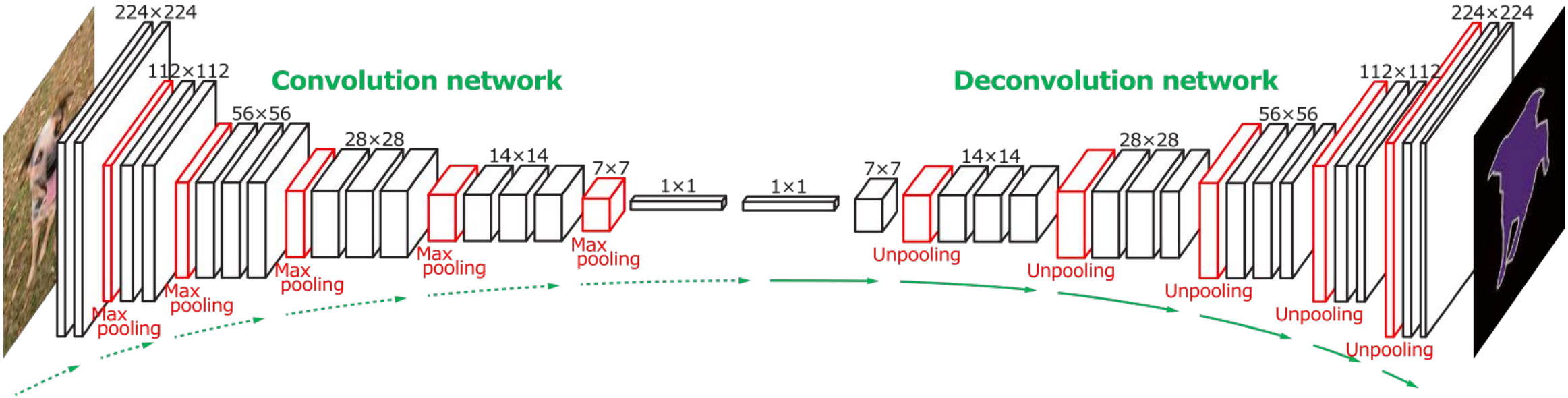
Learning Deconvolution Network for Semantic Segmentation



Face++ 旷视

Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning Deconvolution Network for Semantic Segmentation." ICCV 2015

Learning Deconvolution Network for Semantic Segmentation



Face++ 旷视

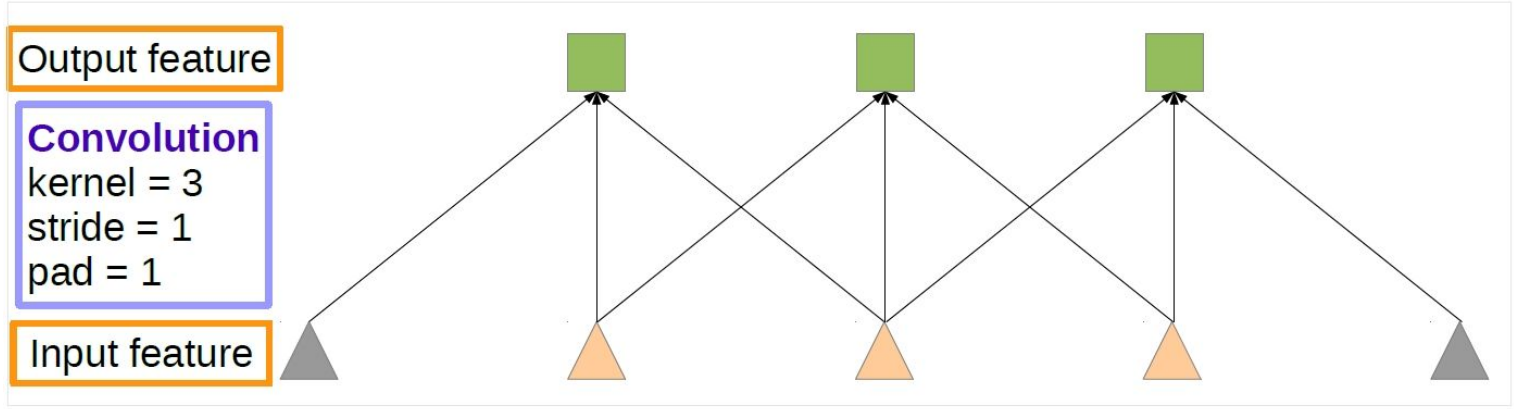
Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning Deconvolution Network for Semantic Segmentation." ICCV 2015

Learning Deconvolution Network for Semantic Segmentation

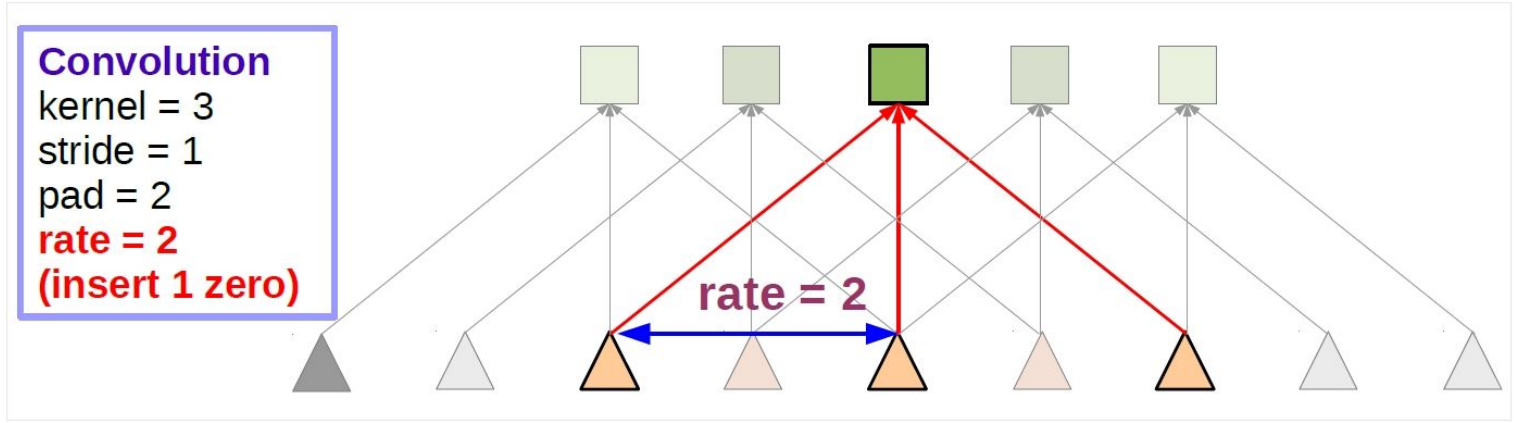
- Introducing un-pool and de-convolution operations.
- Introducing hourglass-like framework.

Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning Deconvolution Network for Semantic Segmentation." ICCV 2015

DeepLab



(a) Sparse feature extraction

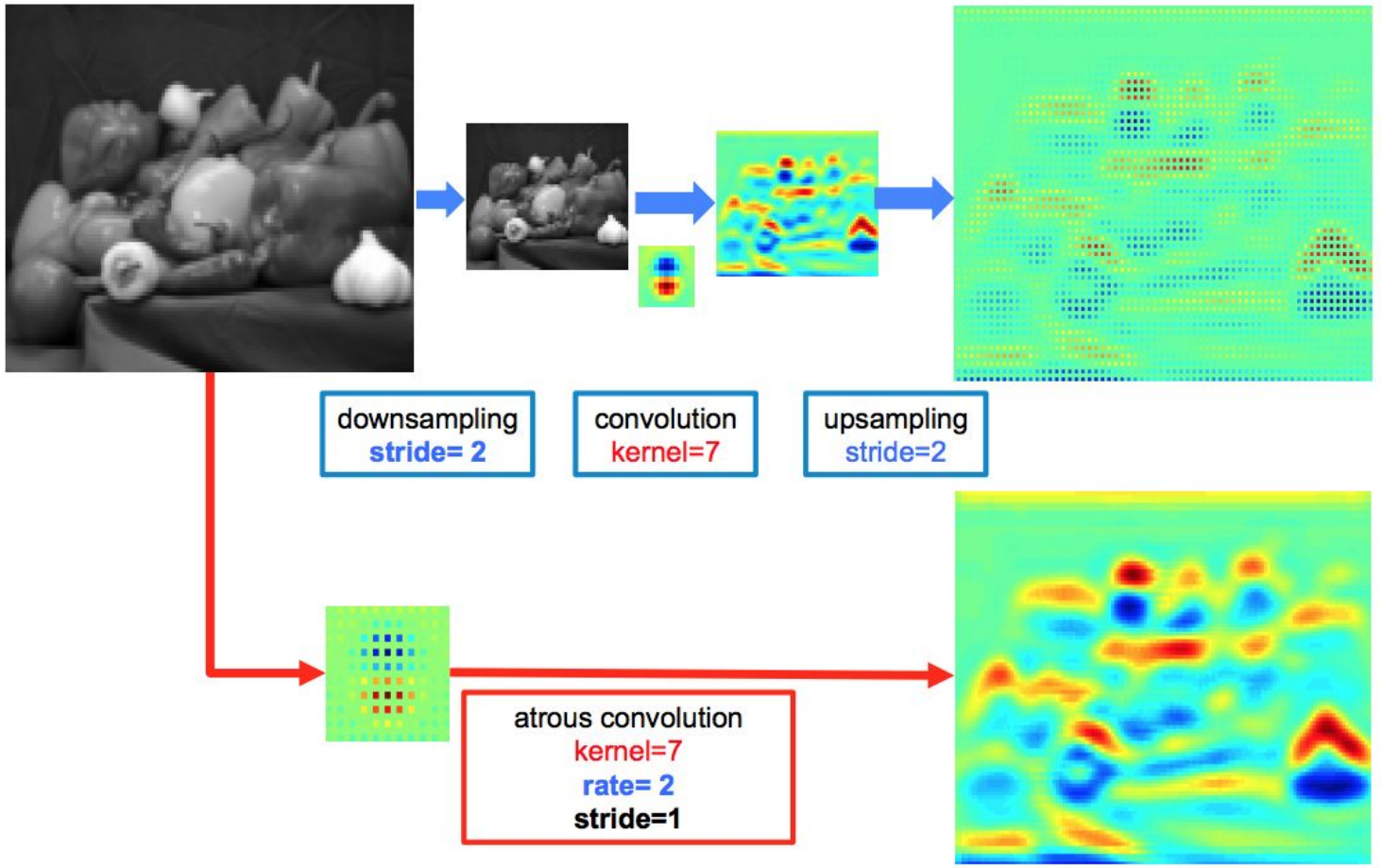


(b) Dense feature extraction

Face++ 旷视

Liang-Chieh Chen*, George Papandreou*, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille (*equal contribution), arXiv preprint, 2016

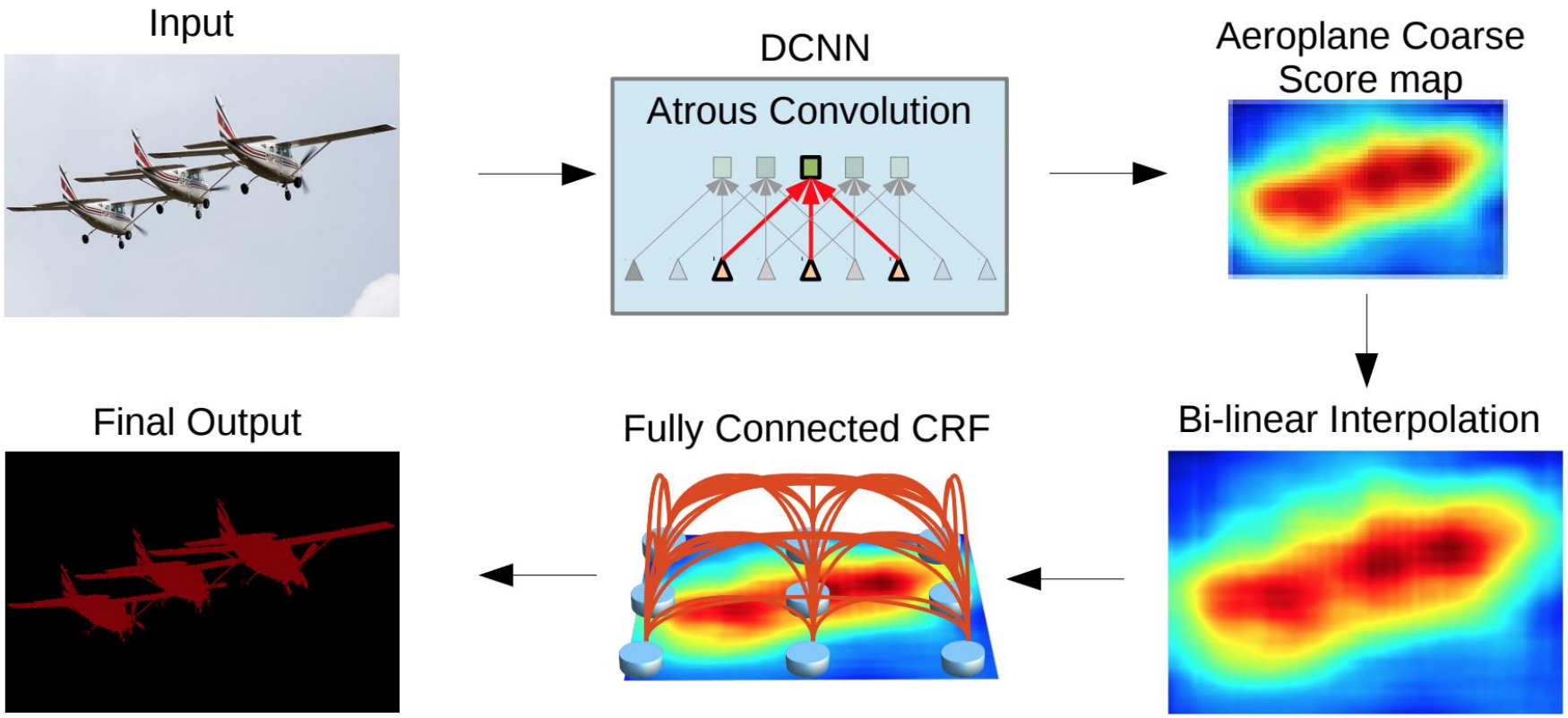
DeepLab



Face++ 旷视

Liang-Chieh Chen*, George Papandreou*, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille (*equal contribution), arXiv preprint, 2016

DeepLab



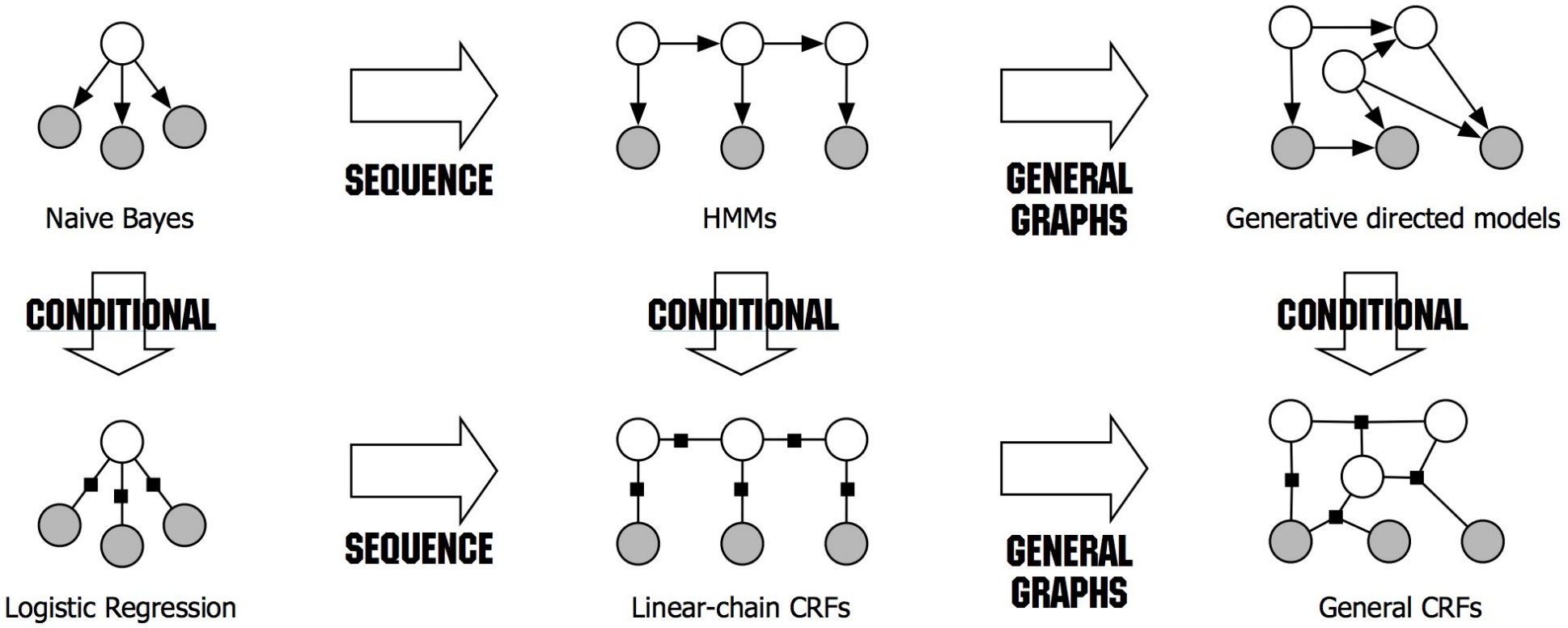
Face++ 旷视

Liang-Chieh Chen*, George Papandreou*, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille (*equal contribution), arXiv preprint, 2016

DeepLab

- Introducing dilated-convolution
- Combining traditional method (post processing): DenseCRF

Conditional Random Field



Face++ 旷视

Sutton, Charles A., and Andrew Mccallum. "An Introduction to Conditional Random Fields." arXiv: Machine Learning 4.4 (2012)

Conditional Random Field

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \exp[-E(\boldsymbol{y}, \boldsymbol{x})]$$

y is the label, **x** is the image

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \exp[-E(\boldsymbol{y}, \boldsymbol{x})]$$

$$E(\boldsymbol{y}, \boldsymbol{x}) = \sum_{U \in \mathcal{U}} \sum_{p \in \mathcal{N}_U} U(y_p, \boldsymbol{x}_p) + \sum_{V \in \mathcal{V}} \sum_{(p,q) \in \mathcal{S}_V} V(y_p, y_q, \boldsymbol{x}_{pq})$$

U: Unary relation;
V: pairwise relation

Face++ 旷视

Sutton, Charles A., and Andrew Mccallum. "An Introduction to Conditional Random Fields." arXiv: Machine Learning 4.4 (2012)

CRF Inference

$$\forall p \in \mathcal{N}: \quad P(y_p|\boldsymbol{x}) = \sum_{\boldsymbol{y} \setminus y_p} P(\boldsymbol{y}|\boldsymbol{x}).$$

- However, for **loopy graph**, the above problem is NP-hard. (The nodes relations are complex, making computing marginal probability harder)
- Approximated methods:
  - MCMC (Gibbs Sampling)
  - Loopy Belief propagation
  - **Mean Field**

Face++ 旷视

Sutton, Charles A., and Andrew Mccallum. "An Introduction to Conditional Random Fields." arXiv: Machine Learning 4.4 (2012)

CRF Inference: Mean field

$$min \sum_{\mathbf{y}} Q(\mathbf{y}) log(\frac{Q(\mathbf{y})}{P(\mathbf{y}|\mathbf{x})})$$

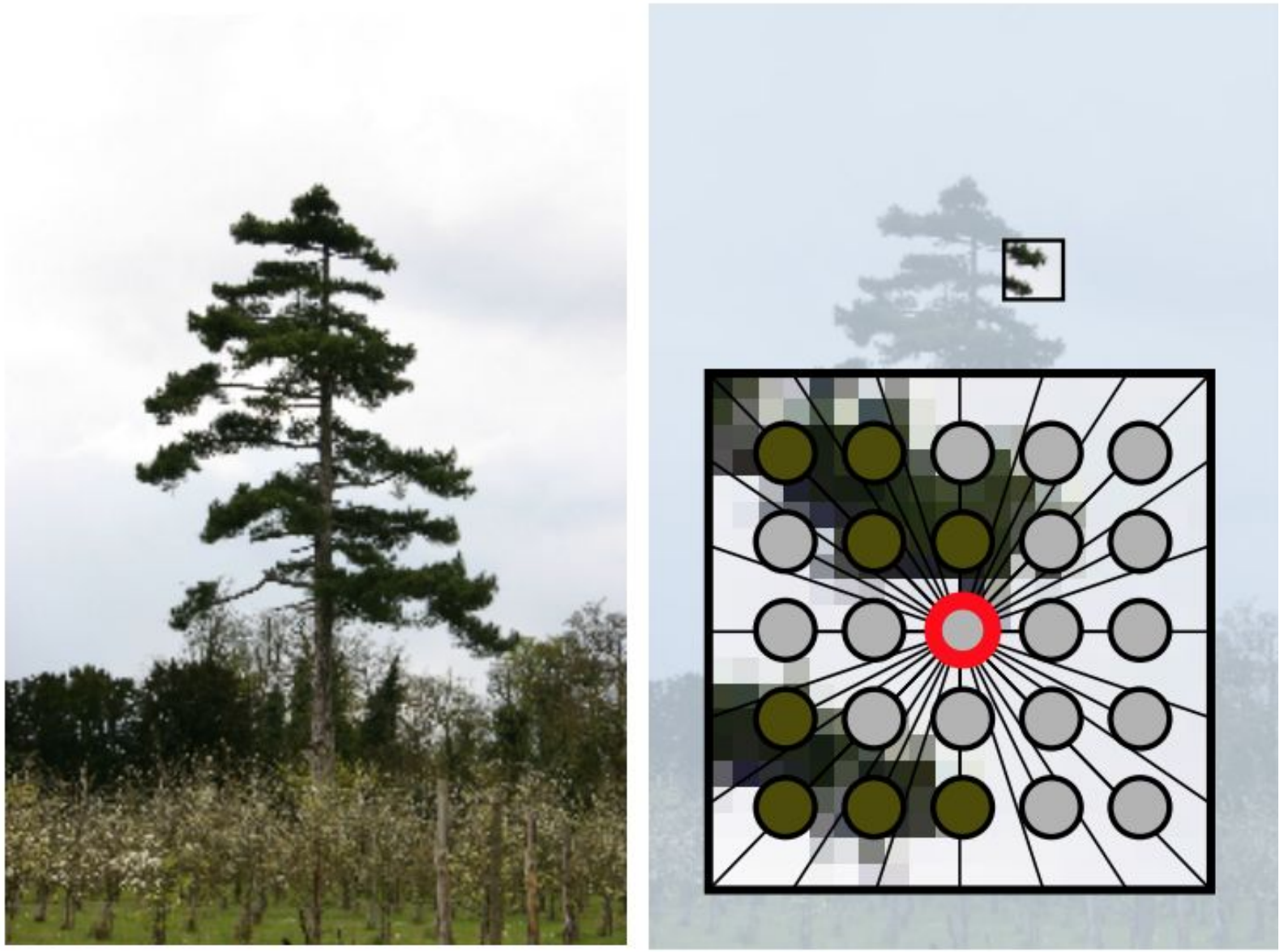
$$s.t.\ Q(\mathbf{y}) = \prod_{i} Q(y_i)$$

$$\sum_{\mathbf{y}} Q(\mathbf{y}) = 1$$

Face++ 旷视

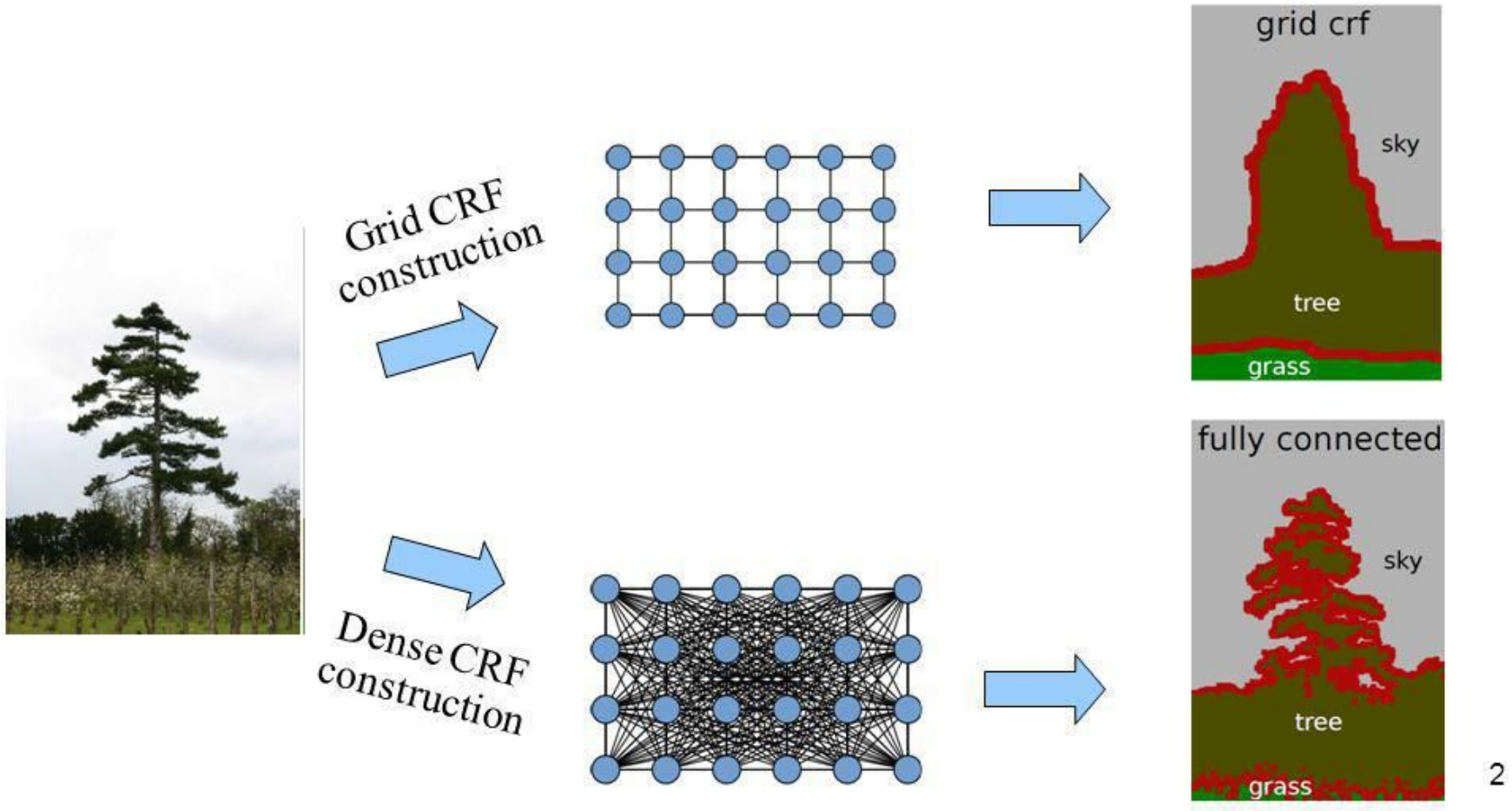
Sutton, Charles A., and Andrew Mccallum. "An Introduction to Conditional Random Fields." arXiv: Machine Learning 4.4 (2012)

DenseCRF



Krahenbuhl, Philipp, and Vladlen Koltun. "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials." NIPS 2011

DenseCRF



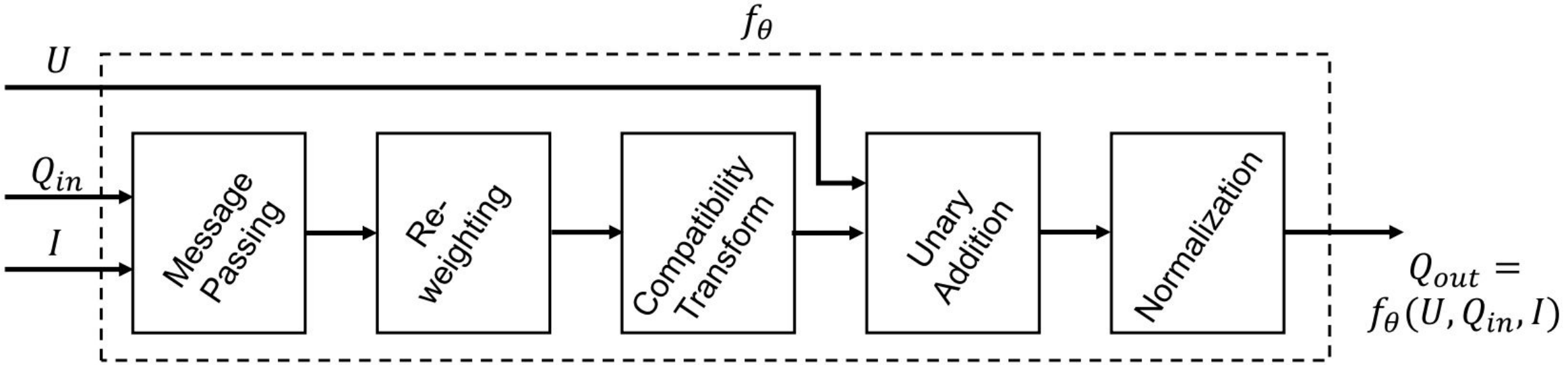
Face++ 旷视

Krahenbuhl, Philipp, and Vladlen Koltun. "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials." NIPS 2011

DenseCRF

- Best Traditional Method!
- Poor accuracy on segmentation! (poor feature)

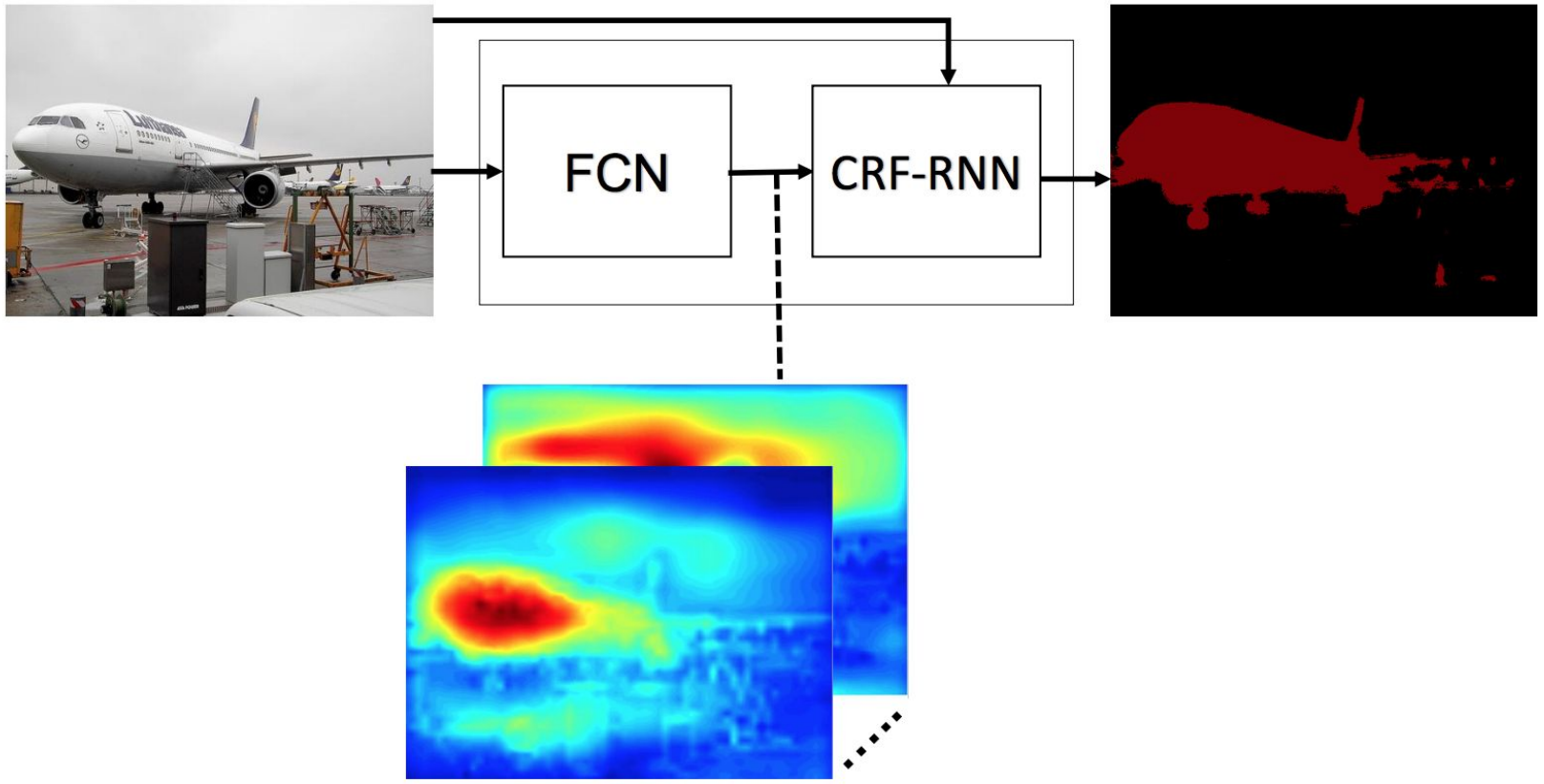
CRF AS RNN



Face++ 旷视

Zheng, Shuai, et al. "Conditional Random Fields as Recurrent Neural Networks." ICCV 2015

CRF AS RNN



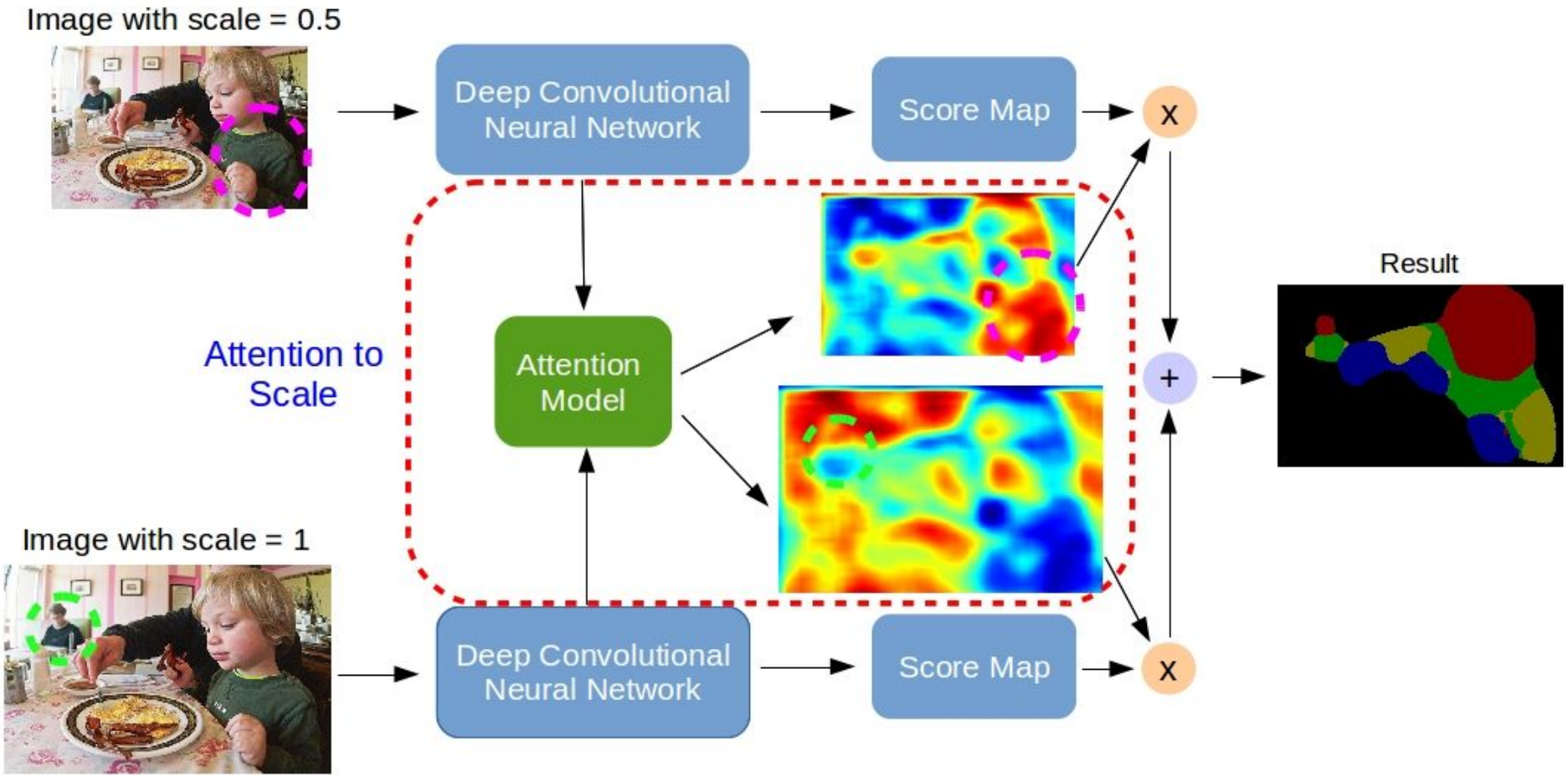
Face++ 旷视

Krahenbuhl, Philipp, and Vladlen Koltun. "Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials." NIPS 2011

CRF AS RNN

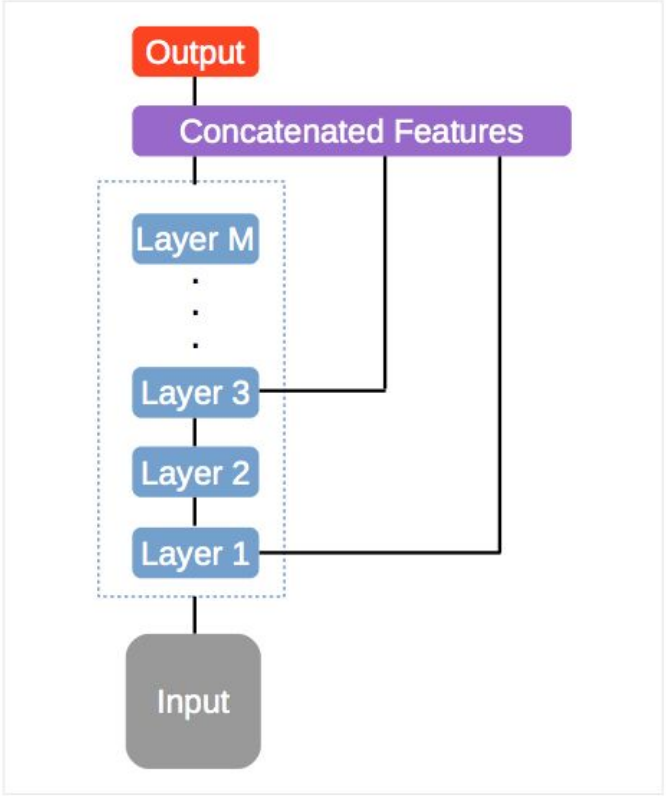
- Encoding the DenseCRF into the CNN framework!
- Better results than DenseCRF in post-processing (Deeplab)

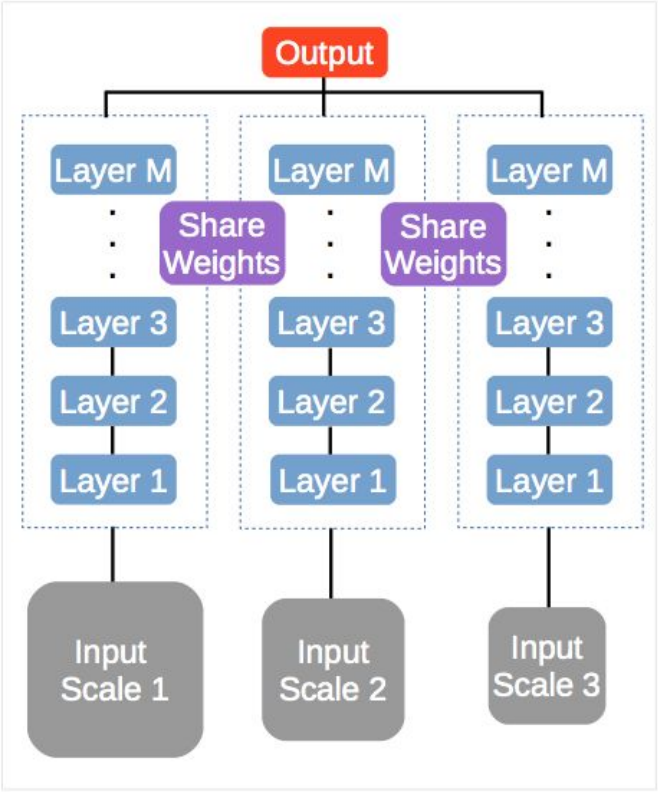
Deeplab Attention



Face++ 旷视

Deeplab Attention

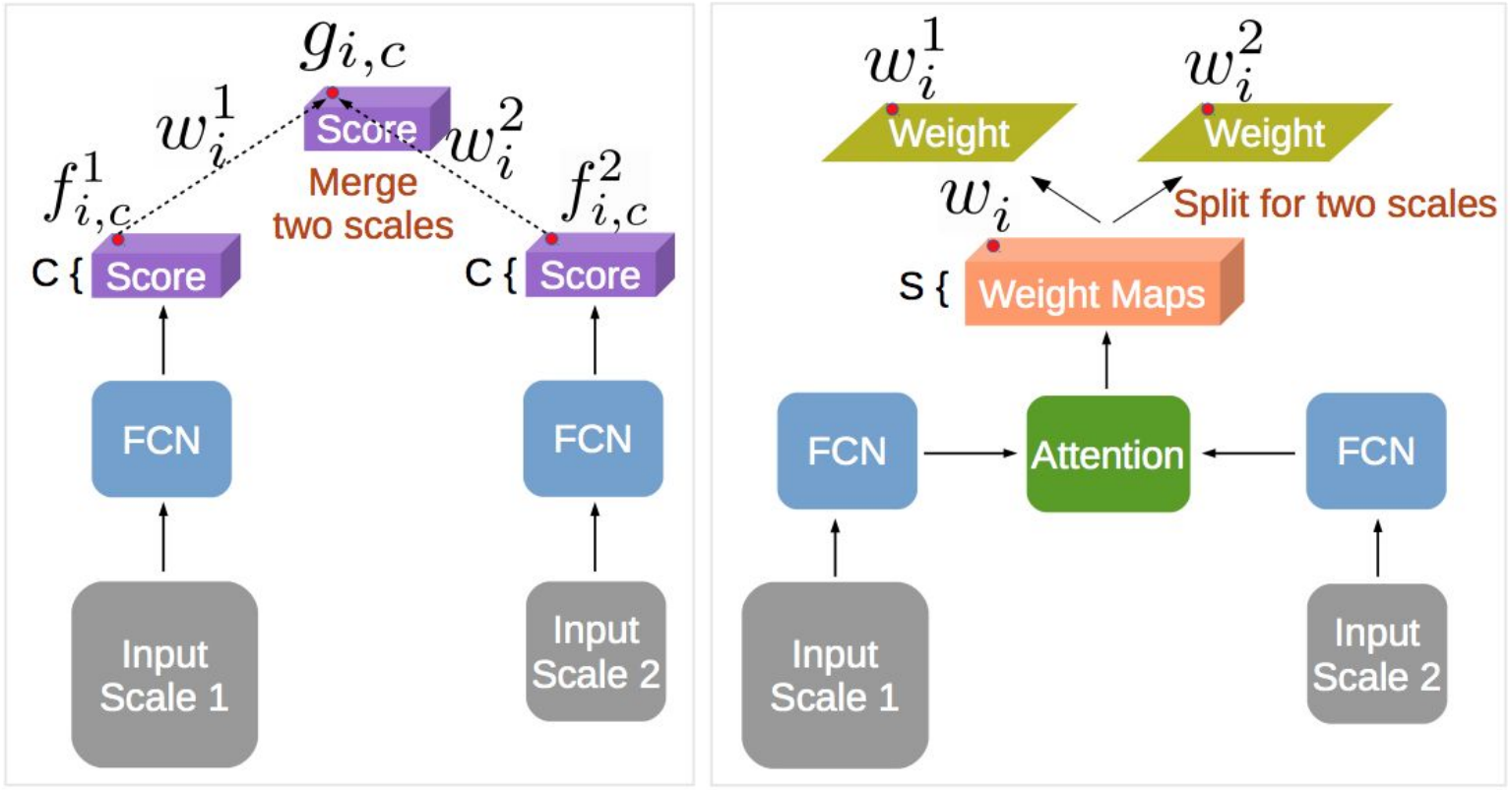




Face++ 旷视

Attention to Scale: Scale-aware Semantic Image Segmentation CVPR 2016

Deeplab Attention



Face++ 旷视

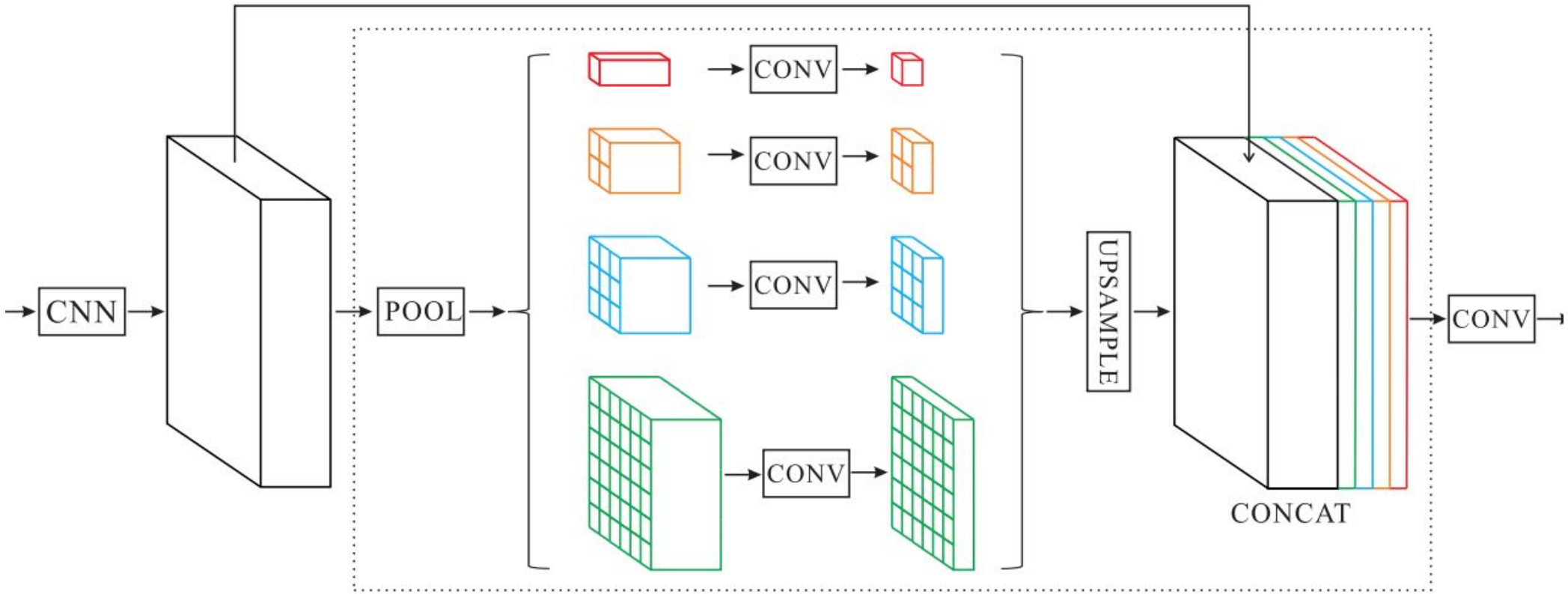
Attention to Scale: Scale-aware Semantic Image Segmentation CVPR 2016

Deeplab Attention

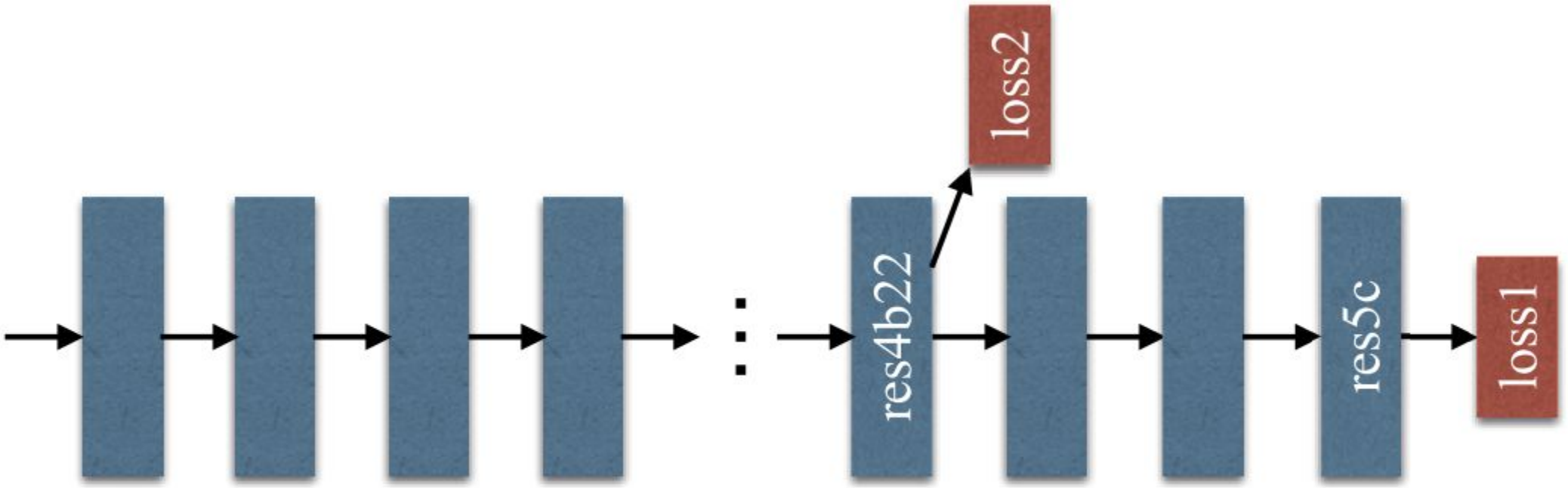
- Fusion framework for multi-scale training and inference!
- Combining Attention model into Segmentation framework!

Face++ 旷视

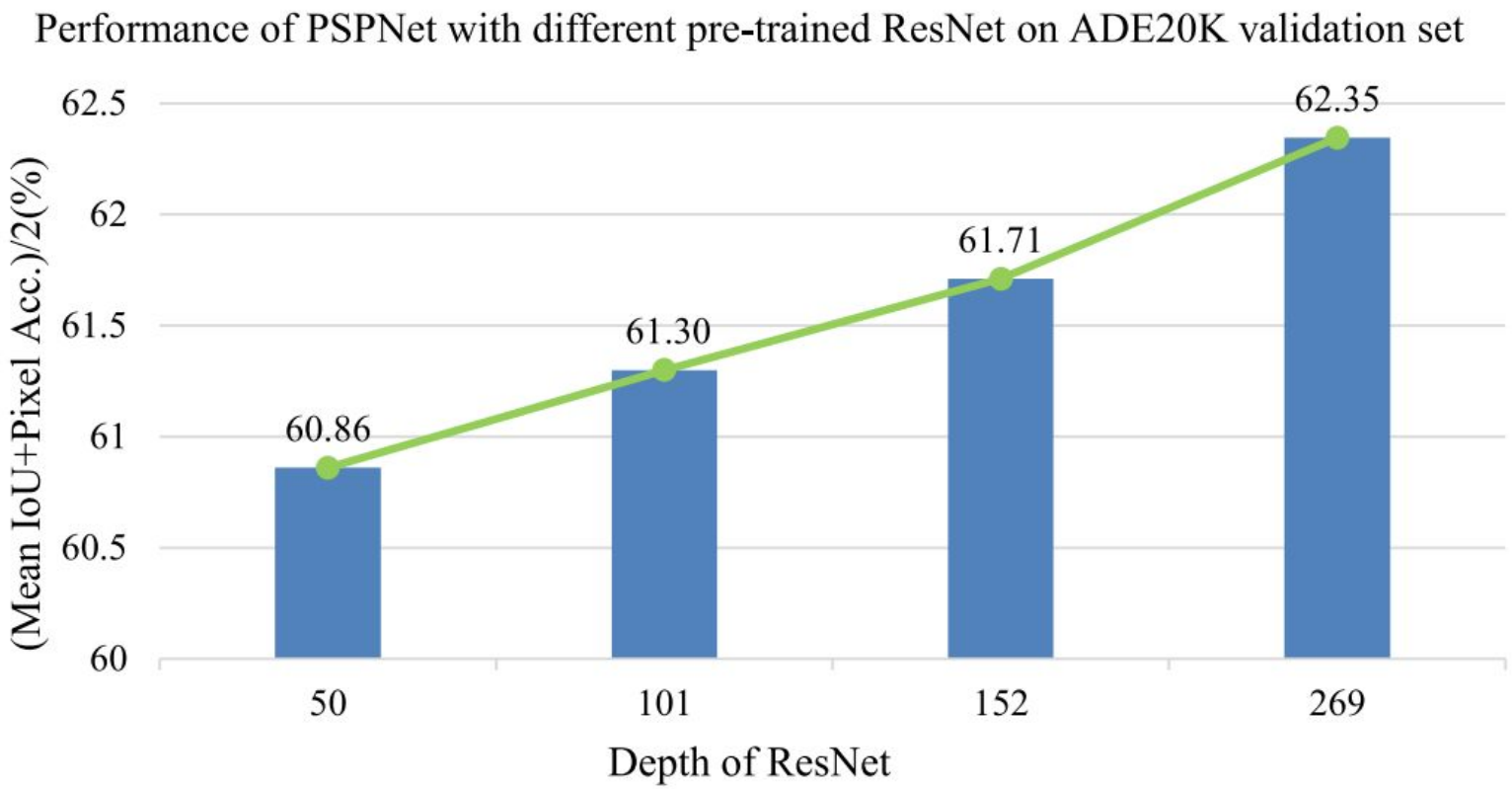
PSPNet



PSPNet



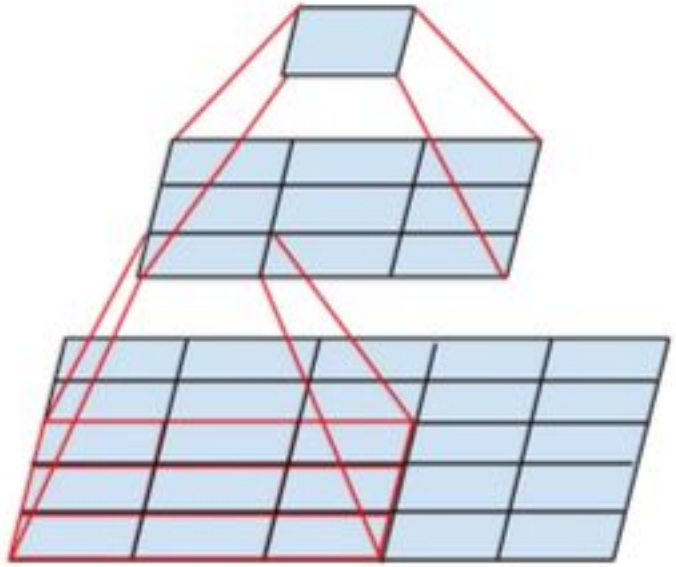
PSPNet

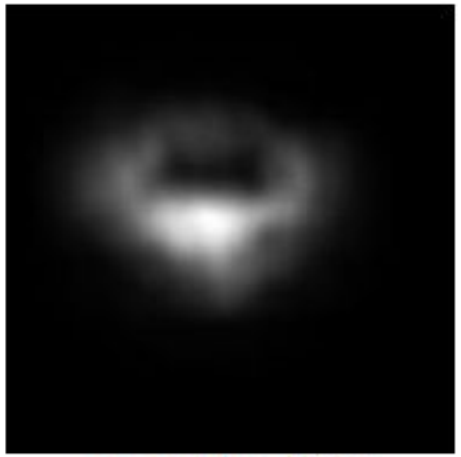

Performance of PSPNet with different pre-trained ResNet on ADE20K validation set
(Mean IoU+Pixel Acc.)/2(%)
62.5
62
61.5
61
60.5
60
60.86
61.30
61.71
62.35
50
101
152
269
Depth of ResNet

PSPNet

- Propose the Pyramid Pooling Module!
- Hard to reproduce!

Global Convolutional Network





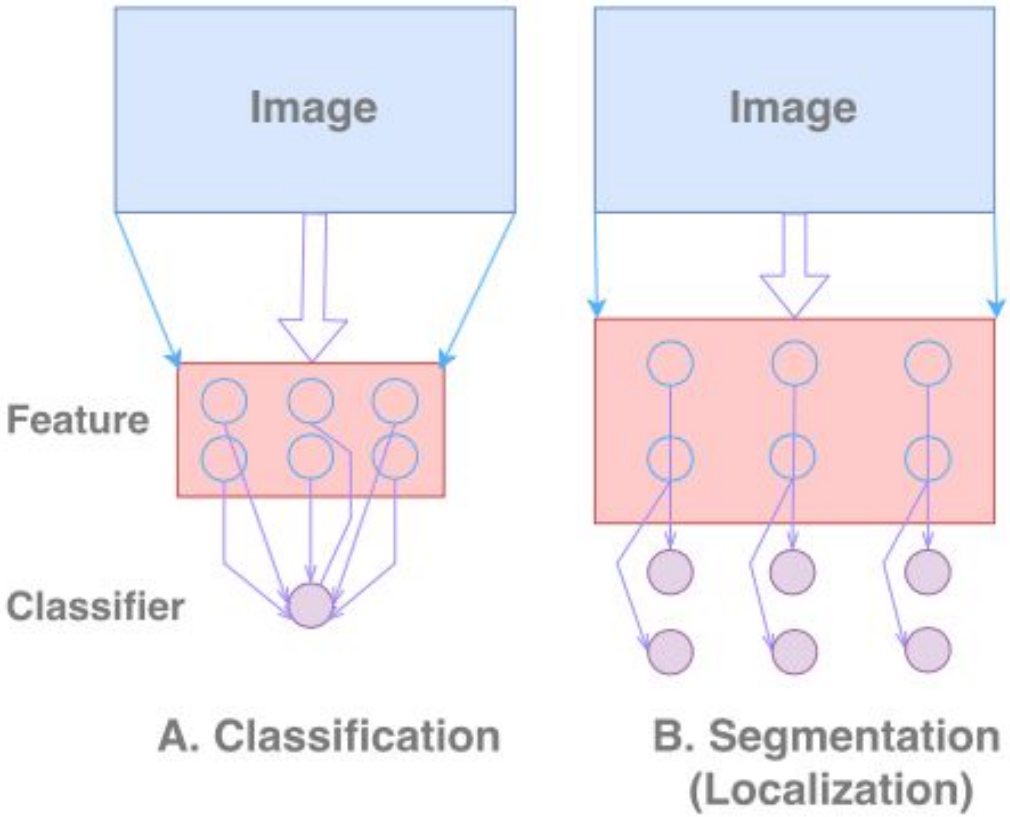
TRF of two 3x3 conv is: 5

However the VRF maybe different！

Figure credit: Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[J]. Computer Science, 2016.
Zhou B, Khosla A, Lapedriza A, et al. Object Detectors Emerge in Deep Scene CNNs[J]. Computer Science, 2015.

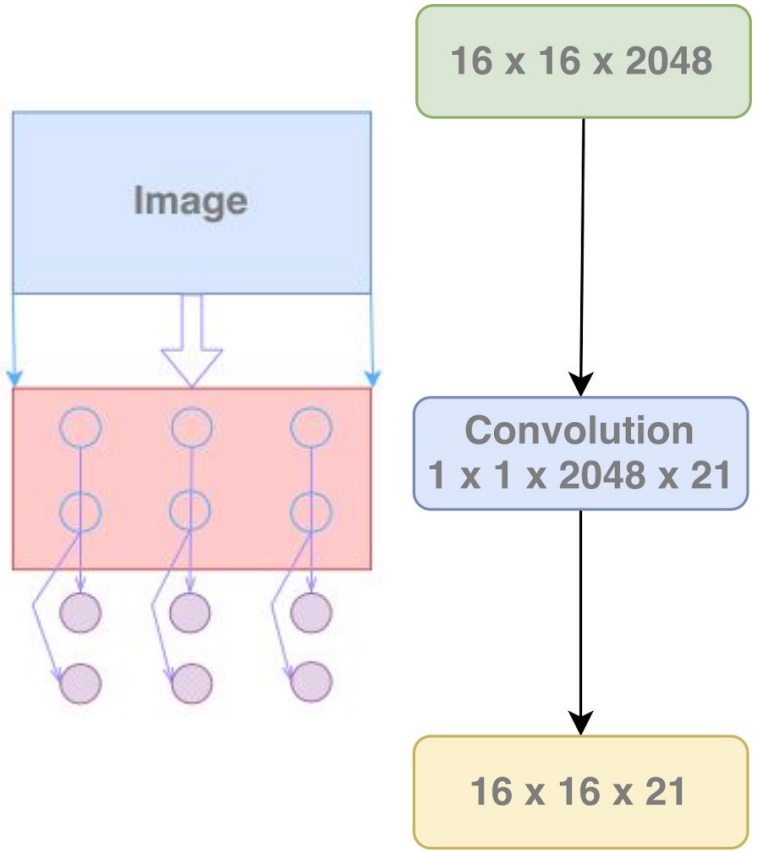
Face++ 旷视

Global Convolutional Network



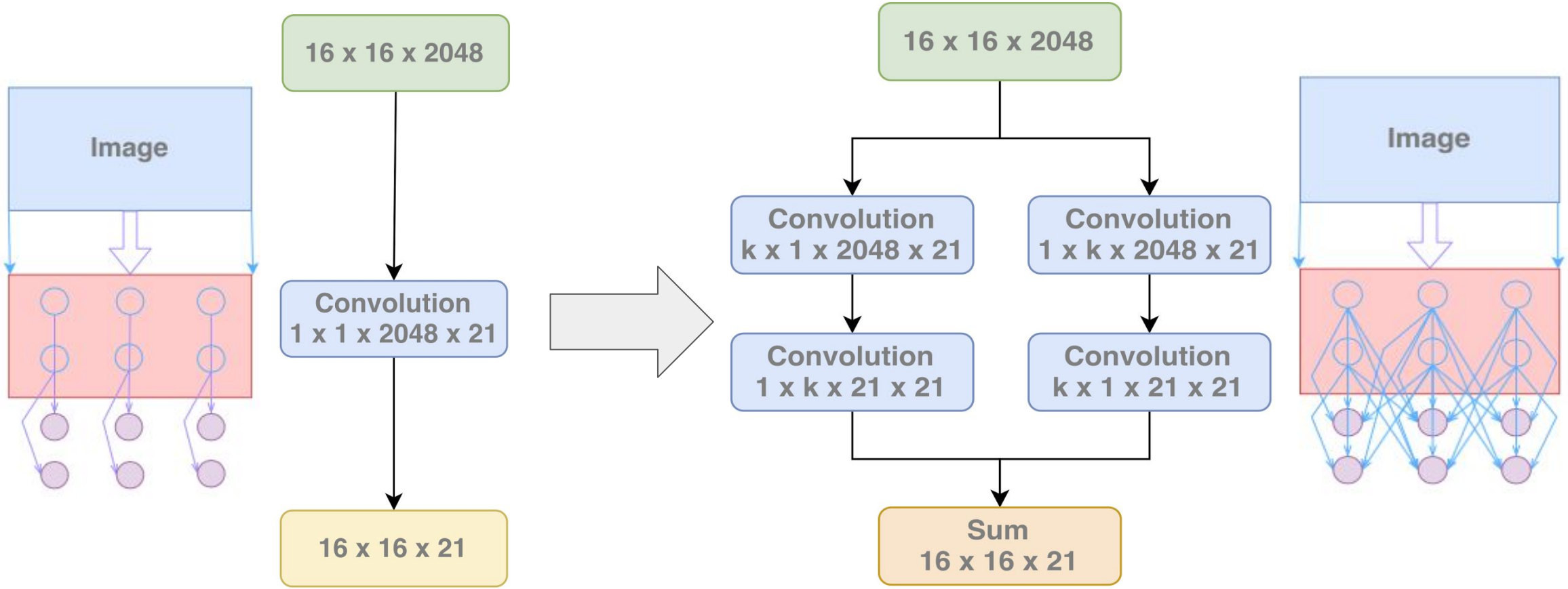
Global Convolutional Network

Global Convolutional Network



Baseline (FCN)

Global Convolutional Network

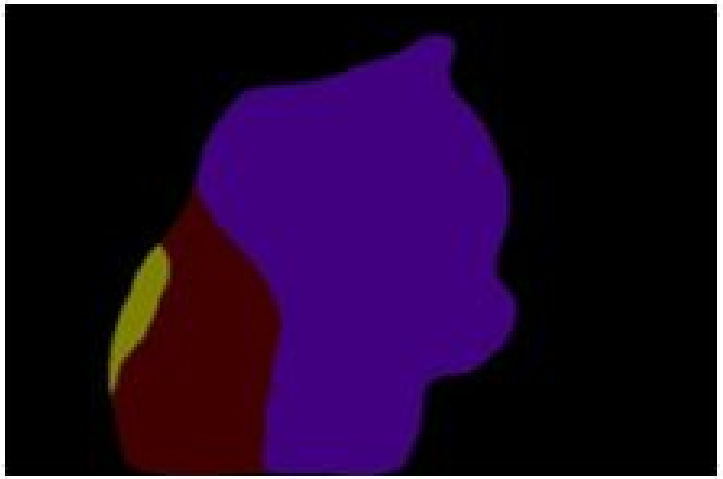


Baseline (FCN)

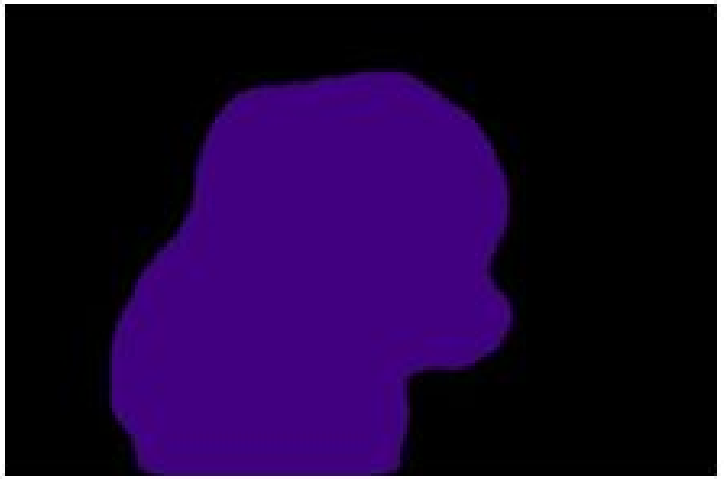
Gloabl Convolutional Network (GCN)

Global Convolutional Network

Image



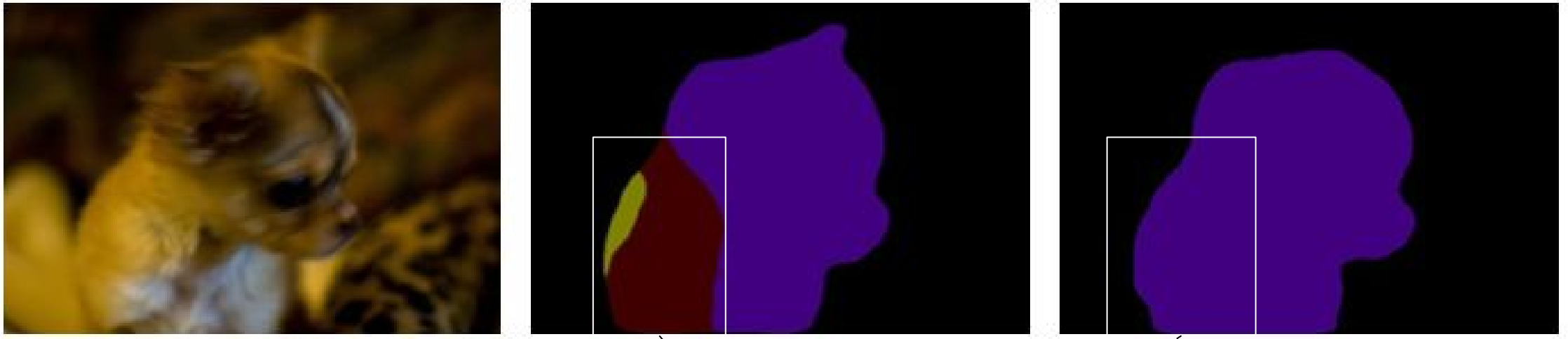
Baseline (FCN)



GCN

Face++ 旷视

Global Convolutional Network



Image

Baseline (FCN)

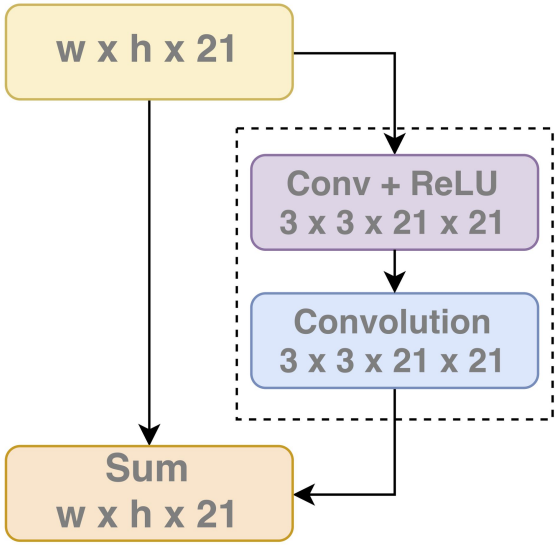
GCN

Region Mis-Classifications are corrected!

Face++ 旷视

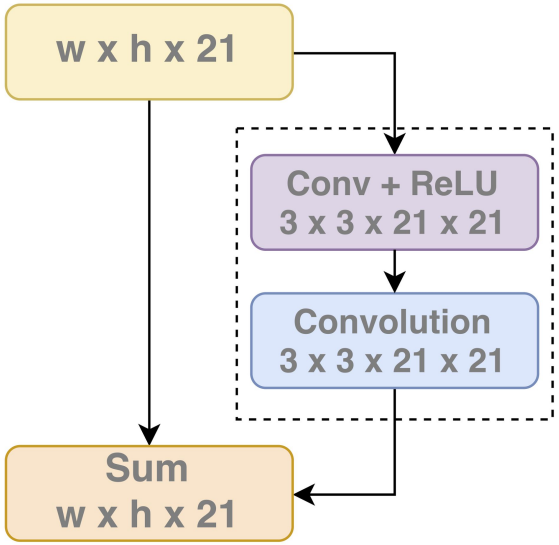
Global Convolutional Network

Face++ 旷视

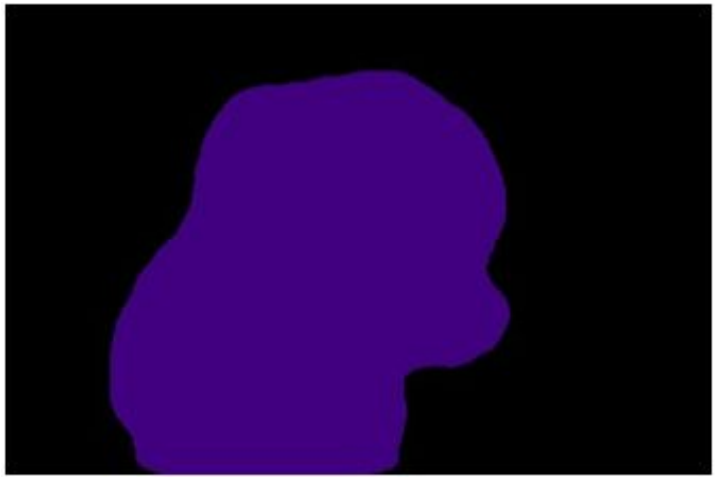


Boundary Refinement (BR)

Face++ 旷视



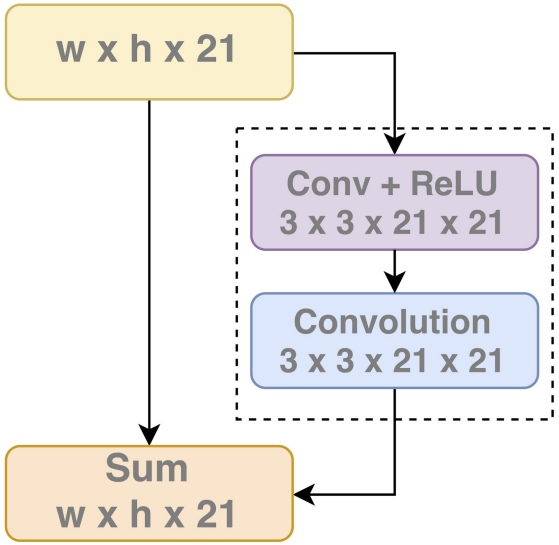
Boundary Refinement (BR)



GCN

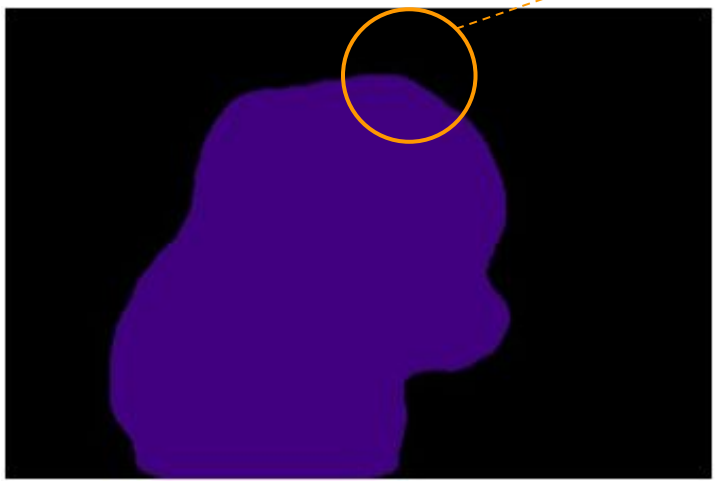
GCN + BR

Face++ 旷视

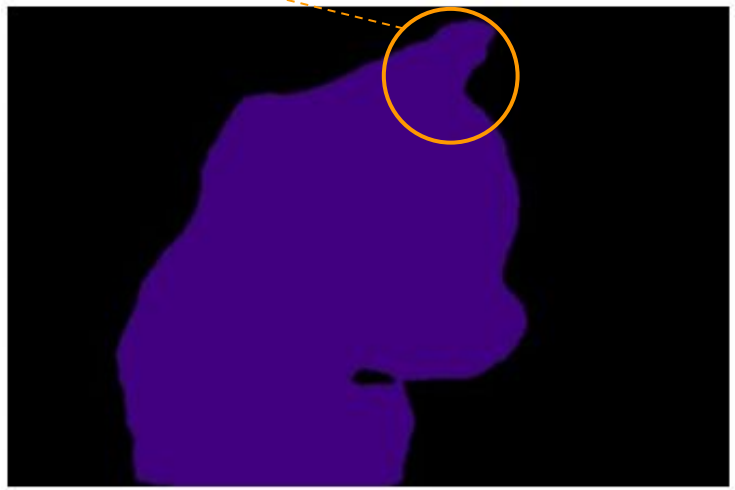


Boundary Refinement (BR)

The Details are recoved!

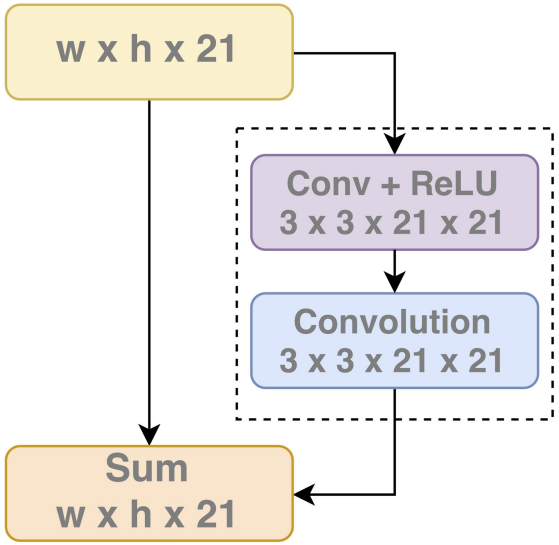


GCN



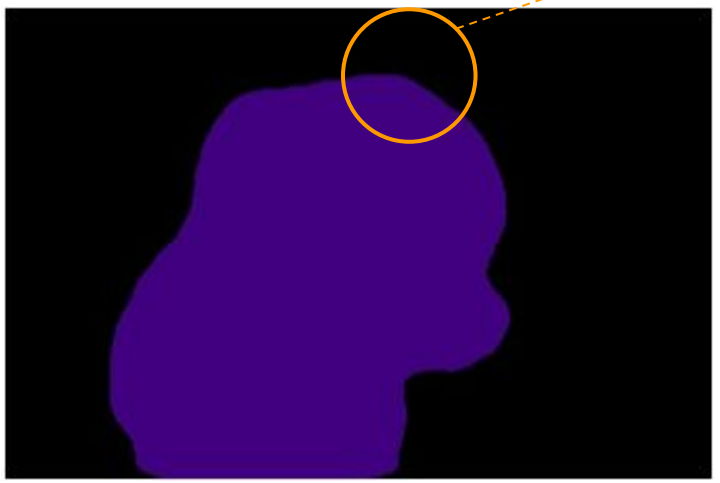
GCN + BR

Face++ 旷视

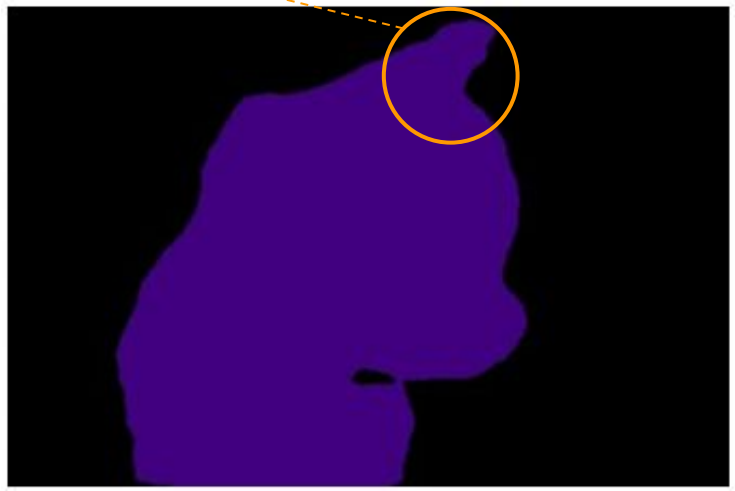


Boundary Refinement (BR)

The Details are recoved!



GCN



GCN + BR

Ground-Truth

Face++ 旷视

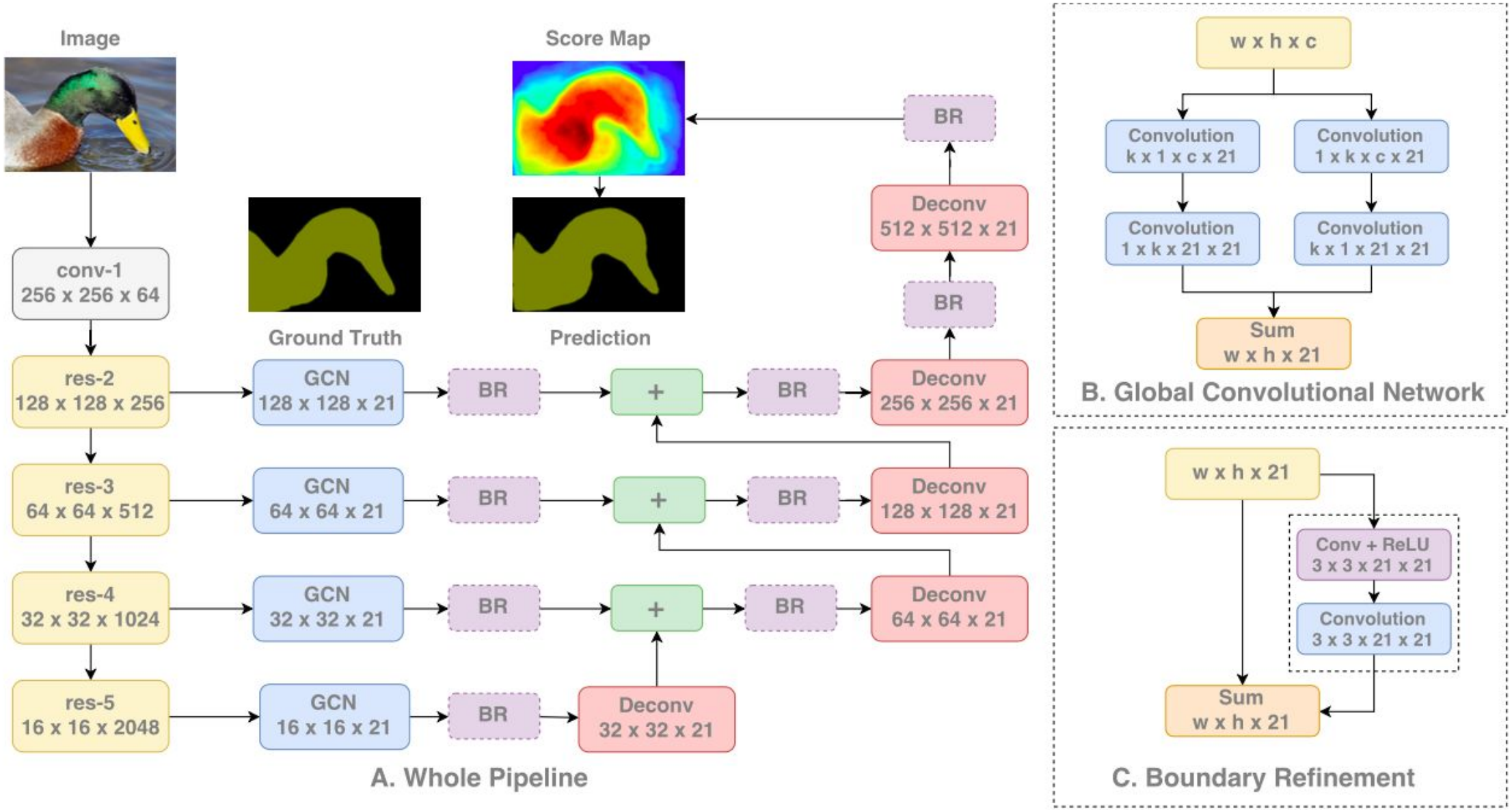


Image
Score Map
conv-1
256 x 256 x 64
Ground Truth
Prediction
res-2
128 x 128 x 256
GCN
128 x 128 x 21
BR
+
BR
Deconv
256 x 256 x 21
BR
Deconv
512 x 512 x 21
BR
res-3
64 x 64 x 512
GCN
64 x 64 x 21
BR
+
BR
Deconv
128 x 128 x 21
res-4
32 x 32 x 1024
GCN
32 x 32 x 21
BR
+
BR
Deconv
64 x 64 x 21
res-5
16 x 16 x 2048
GCN
16 x 16 x 21
BR
Deconv
32 x 32 x 21
A. Whole Pipeline
w x h x c
Convolution
k x 1 x c x 21
Convolution
1 x k x c x 21
Convolution
1 x k x 21 x 21
Convolution
k x 1 x 21 x 21
Sum
w x h x 21
B. Global Convolutional Network
w x h x 21
Conv + ReLU
3 x 3 x 21 x 21
Convolution
3 x 3 x 21 x 21
Sum
w x h x 21
C. Boundary Refinement

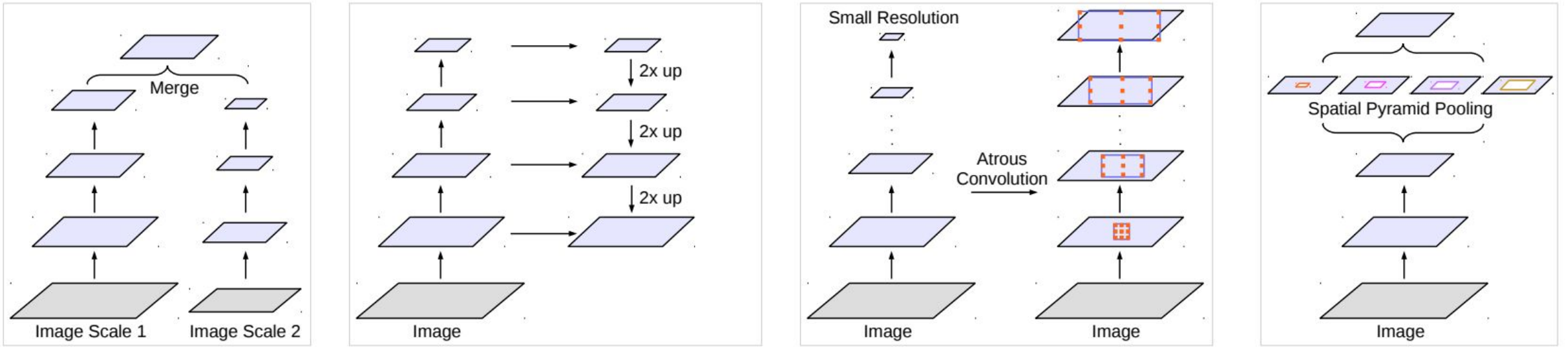
Face++ 旷视

Global Convolutional Network

- Extend the FCN framework!
- Partially Solve the Receptive Field Problem!
- Two key components! (GCN and BRN)

Face++ 旷视

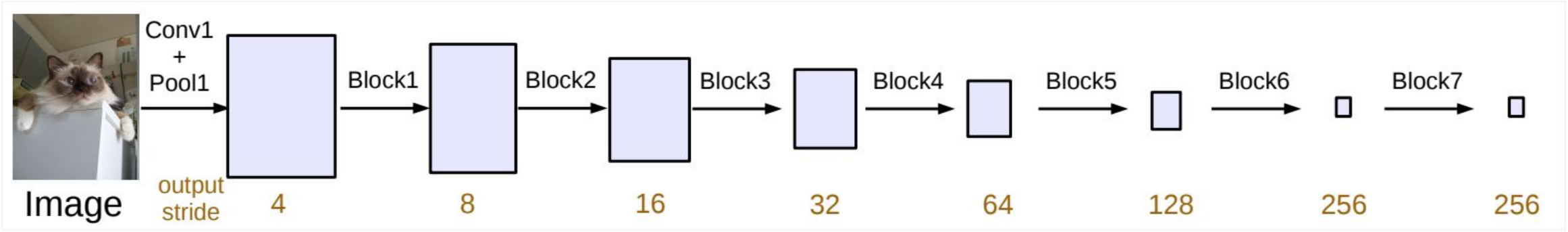
Deeplab V3



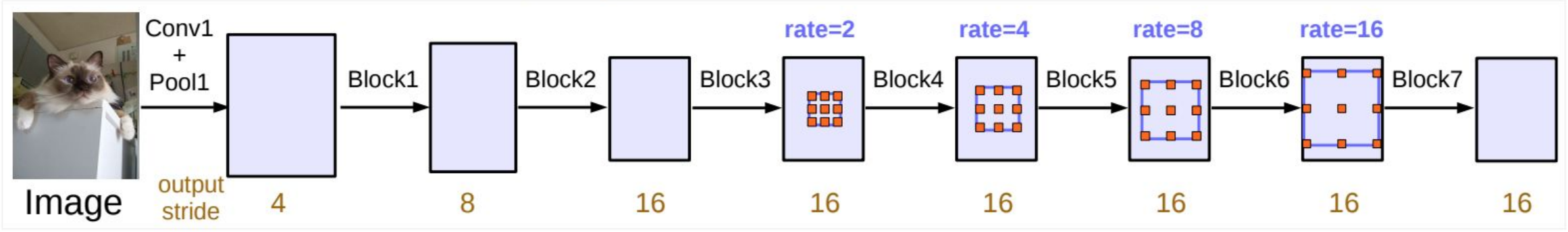
(a) Image Pyramid (b) Encoder-Decoder (c) Deeper w. Atrous Convolution (d) Spatial Pyramid Pooling

Figure 2. Alternative architectures to capture multi-scale context.

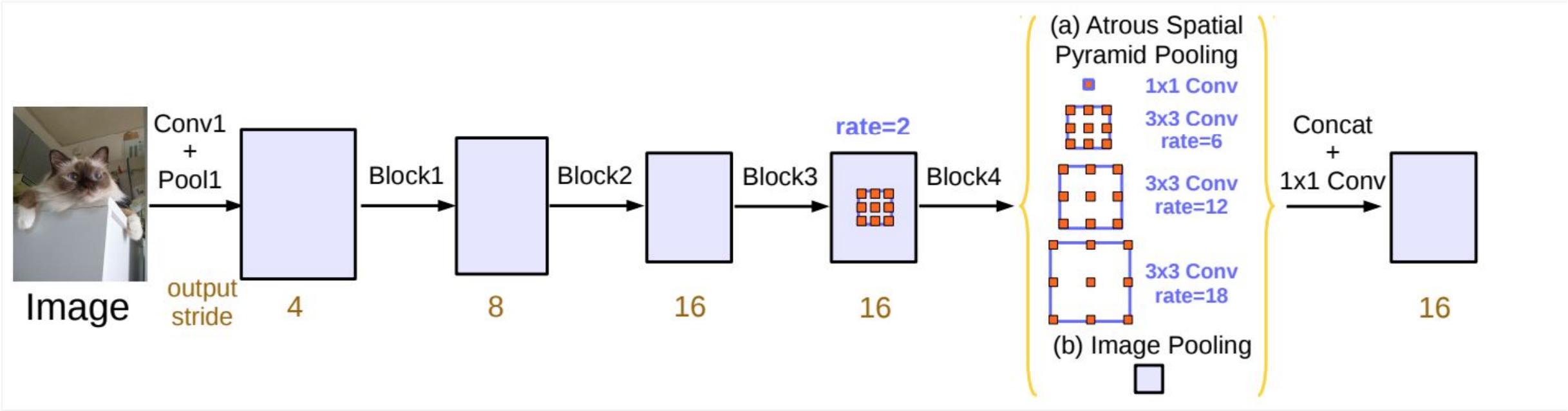
Deeplab V3



(a) Going deeper without atrous convolution.



Deeplab V3



Rethinking Atrous Convolution for Semantic Image Segmentation, arxiv

Deeplab V3

	mean	aero plane	bicycle	bird
DeepLabv3-JFT [?]	86.9	96.9	73.2	95.5
DIS [?]	86.8	94.0	73.3	93.5
CASIA_IVA_SDN [?]	86.6	96.9	78.6	96.0
IDW-CNN [?]	86.3	94.8	67.3	93.4
DeepLabv3 [?]	85.7	96.4	76.6	92.7
PSPNet [?]	85.4	95.8	72.7	95.0
** [?]	84.9	96.2	75.2	95.4

Face++ 旷视

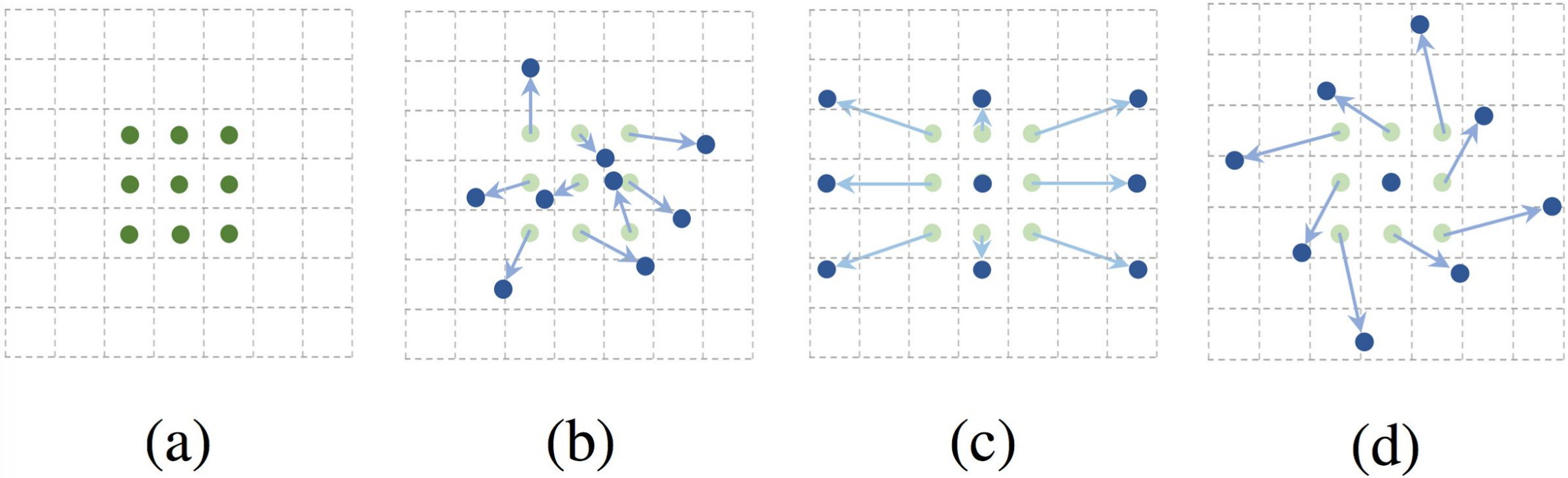
Rethinking Atrous Convolution for Semantic Image Segmentation, arxiv

Deeplab V3

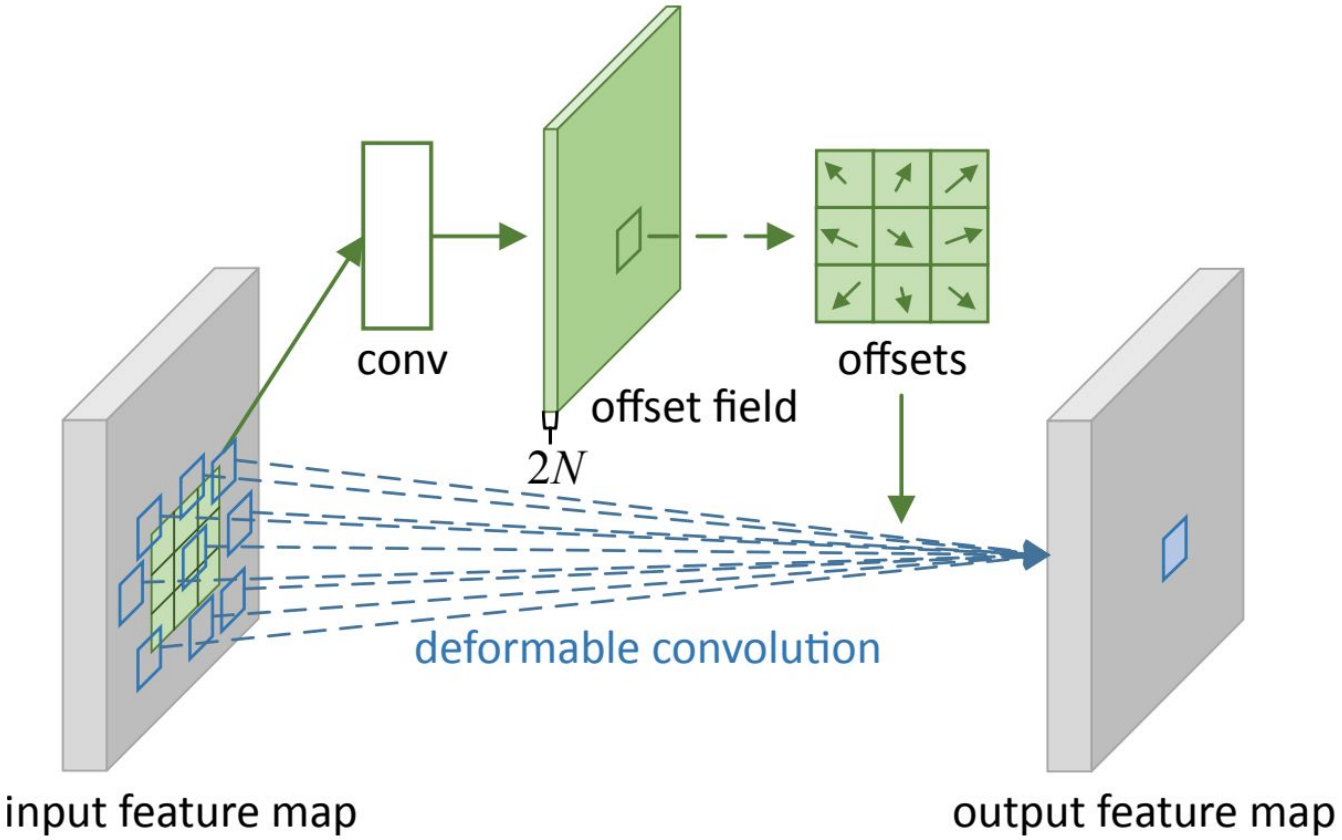
- Currently State-Of-Art on PASCAL VOC 2012
- Conclude the dilate-convolution technique on segmentation

Rethinking Atrous Convolution for Semantic Image Segmentation, arxiv

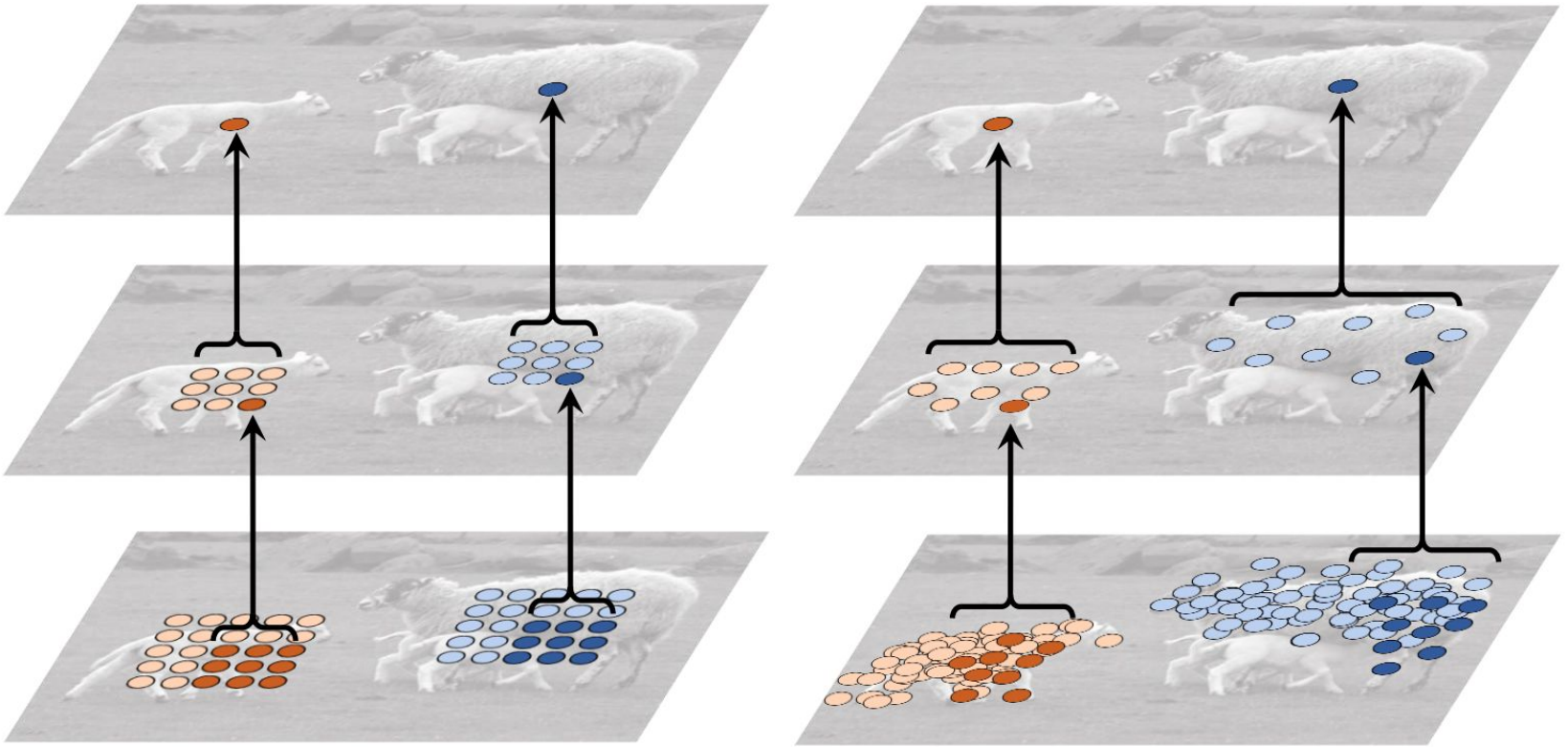
Deformable Convolution



Deformable Convolution



Deformable Convolution



Face++ 旷视

Deformable Convolutional Networks, arxiv

Deformable Convolution

Deformable Convolution

- Solve the receptive field problem using learned offsets!
- Also valid for detection!

Re-Cap

- Segmentation with CNN: FCN, Deeplab, GCN ...
- Segmentation with CRF: DenseCRF, CRFAsRNN, ...
- Different Convolutions: Dilated Conv, Global Conv, Deformable, ...

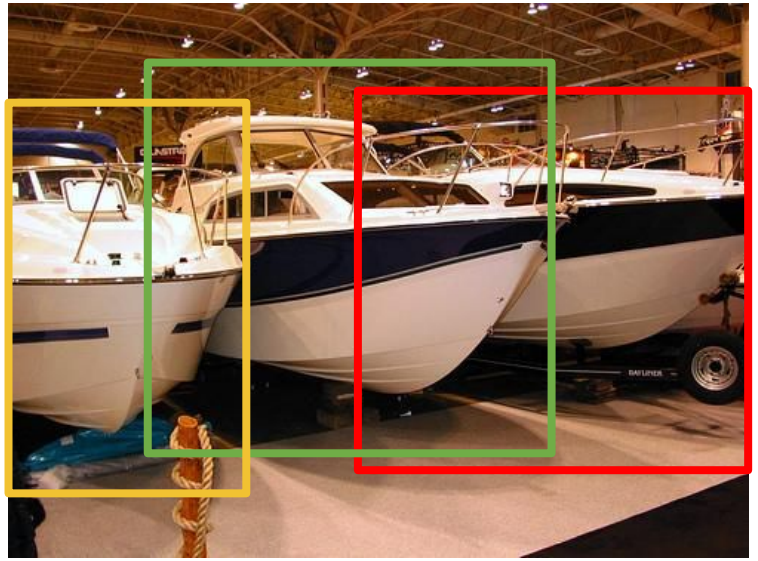
Outline

- Semantic Segmentation
- Instance Segmentation

Top-Down Pipeline

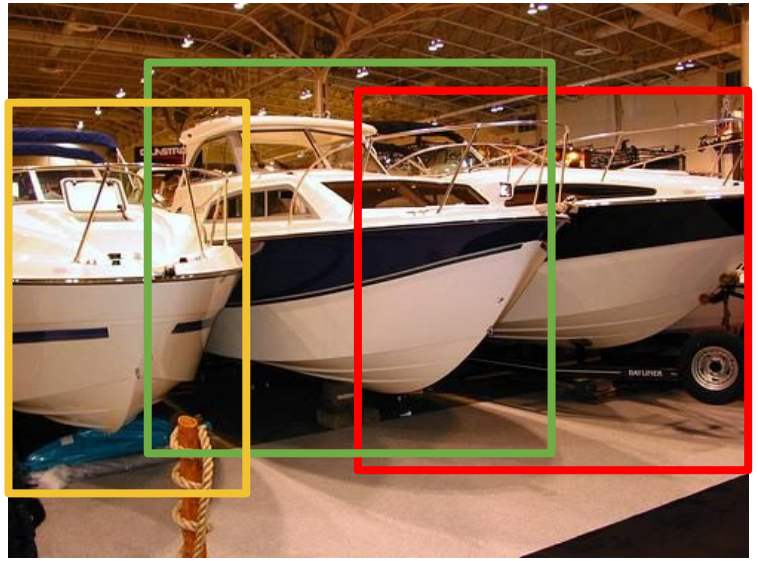
Face++ 旷视

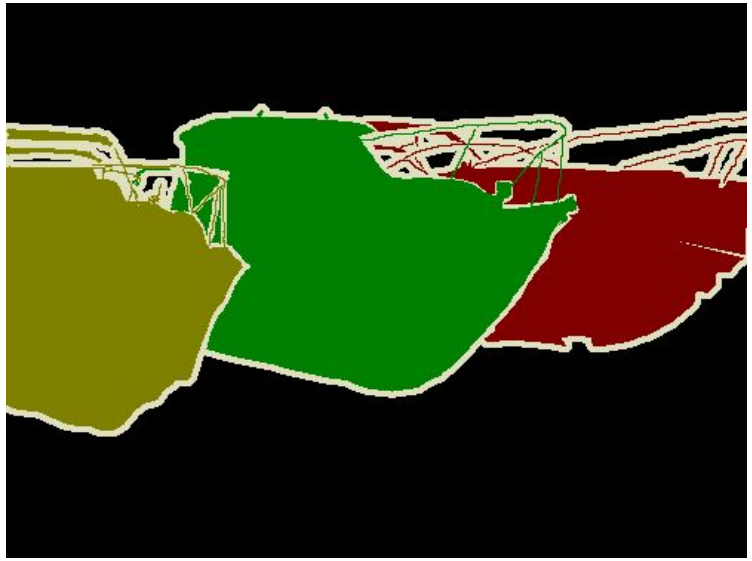
Top-Down Pipeline



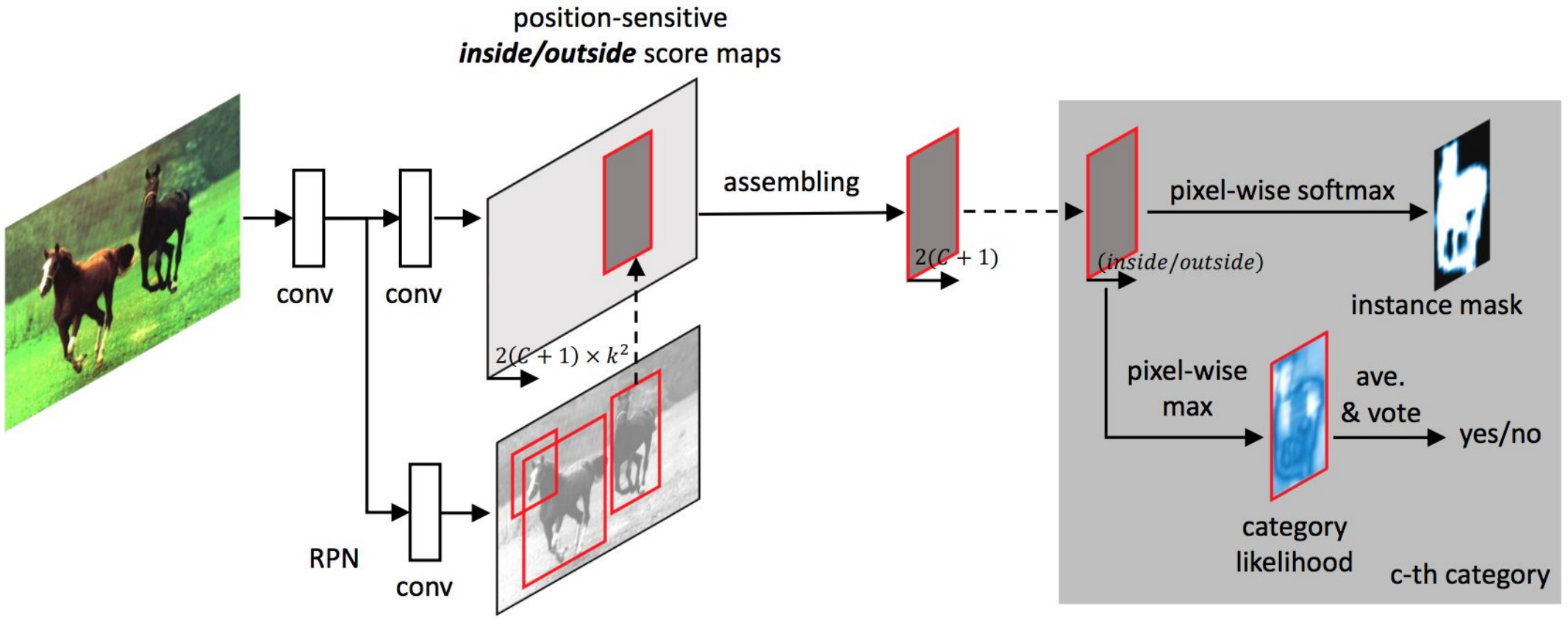
Face++ 旷视

Top-Down Pipeline

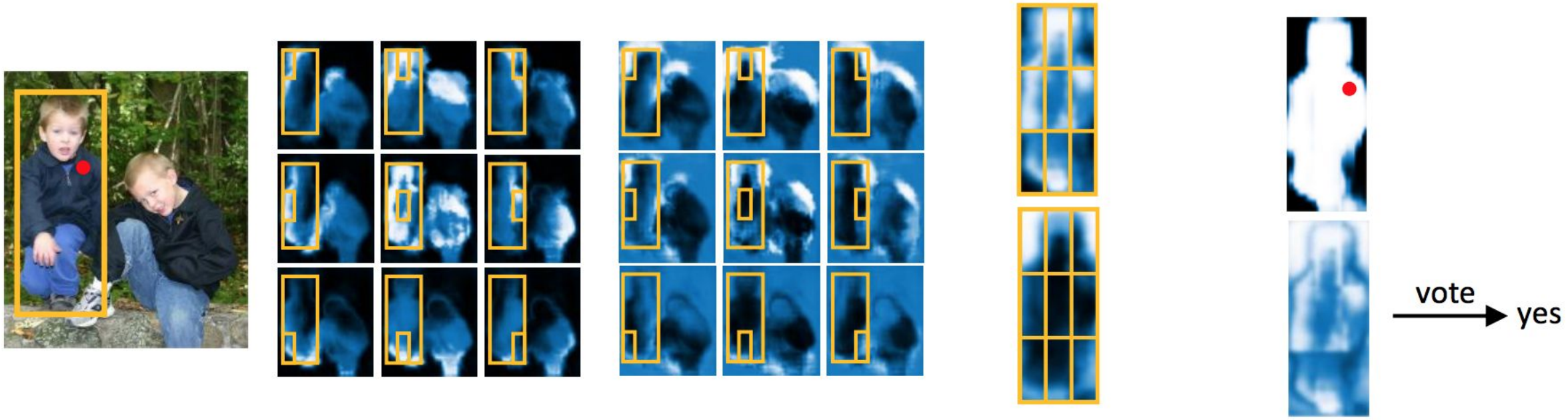




FCIS

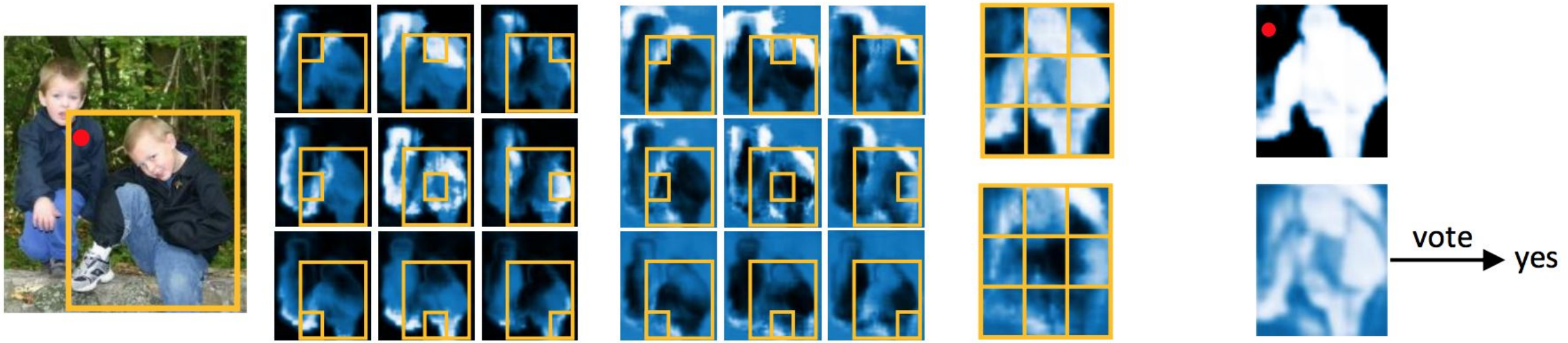

position-sensitive
inside/outside score maps
conv
conv
RPN
conv
$2(C+1)\times k^2$
assembling
$2(C+1)$
pixel-wise softmax
$(inside/outside)$
instance mask
pixel-wise
max
ave.
& vote
yes/no
category
likelihood
c-th category

FCIS

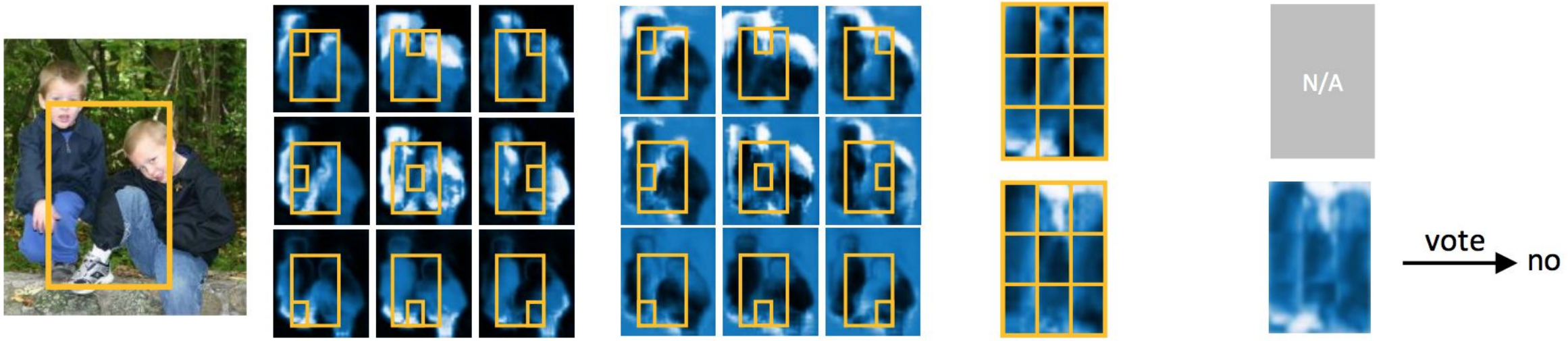


Fully Convolutional Instance-aware Semantic Segmentation, CVPR 2017

FCIS



FCIS

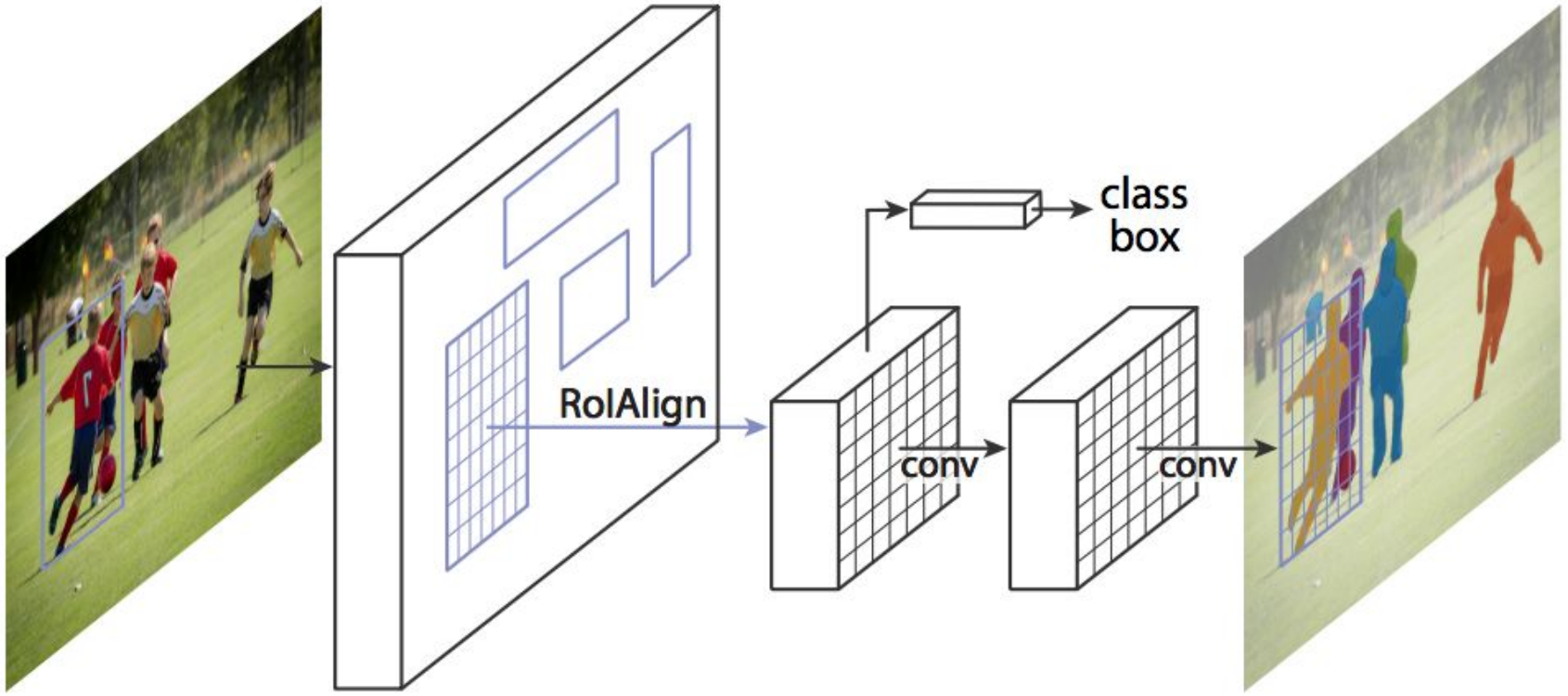


Fully Convolutional Instance-aware Semantic Segmentation, CVPR 2017

FCIS

- Bases on R-FCN framework
- Propose the inside-outside scoring technique!
- Winner of COCO InstanceSeg Challenge 2016!

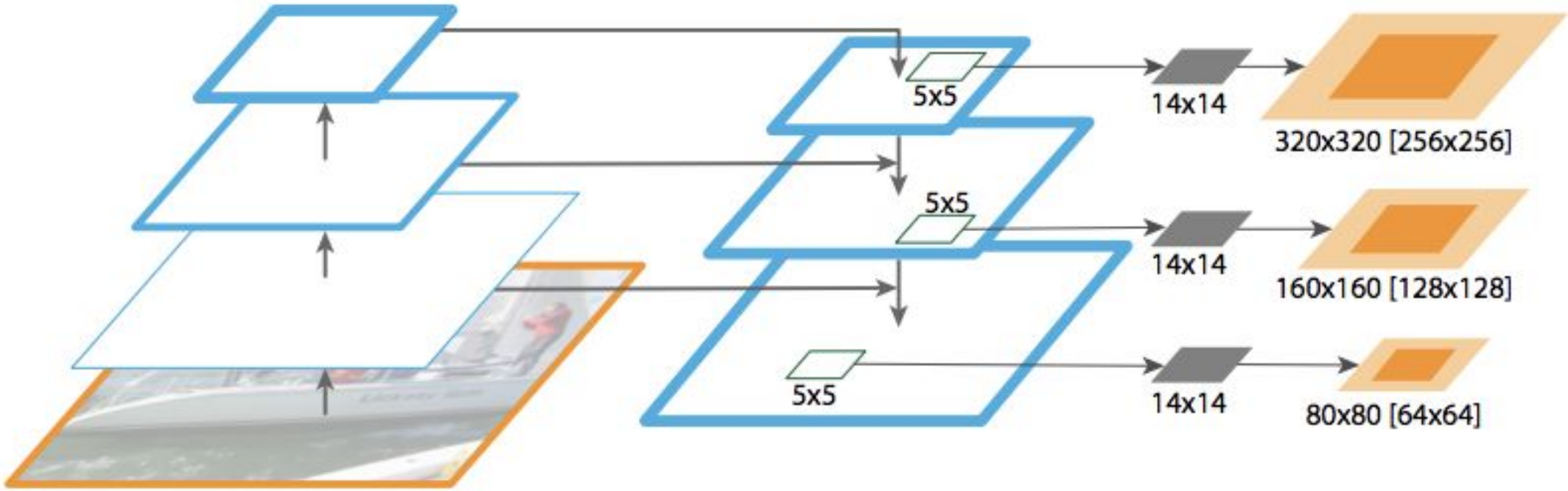
Mask RCNN



Face++ 旷视

Mask-RCNN, ICCV 2017

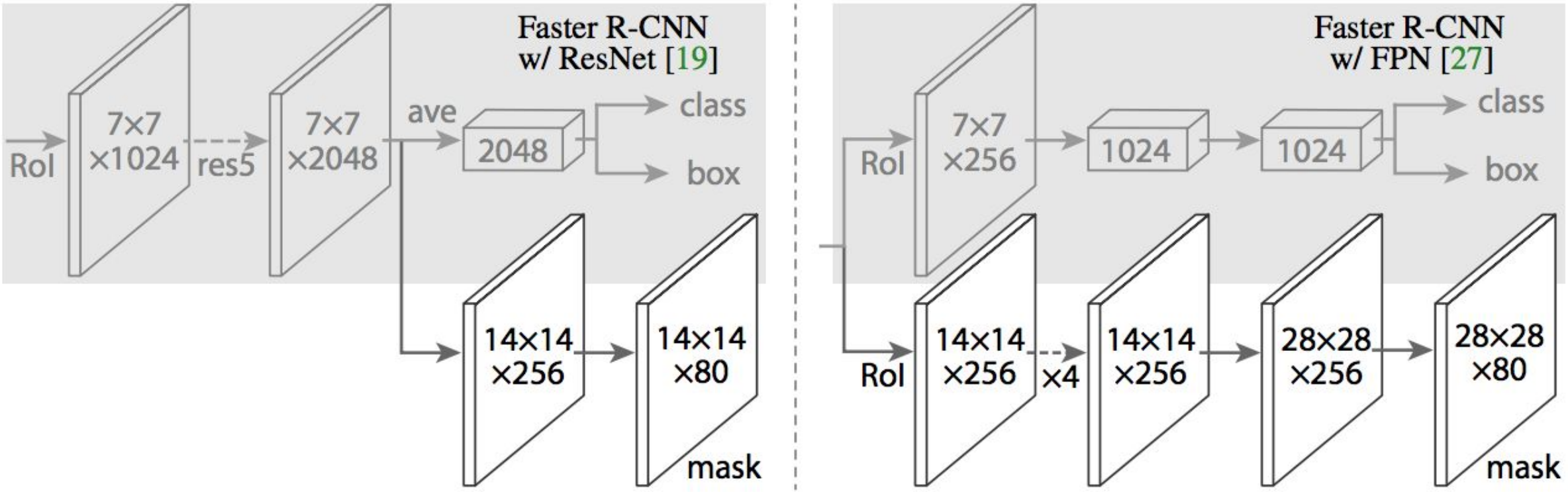
Mask RCNN



Face++ 旷视

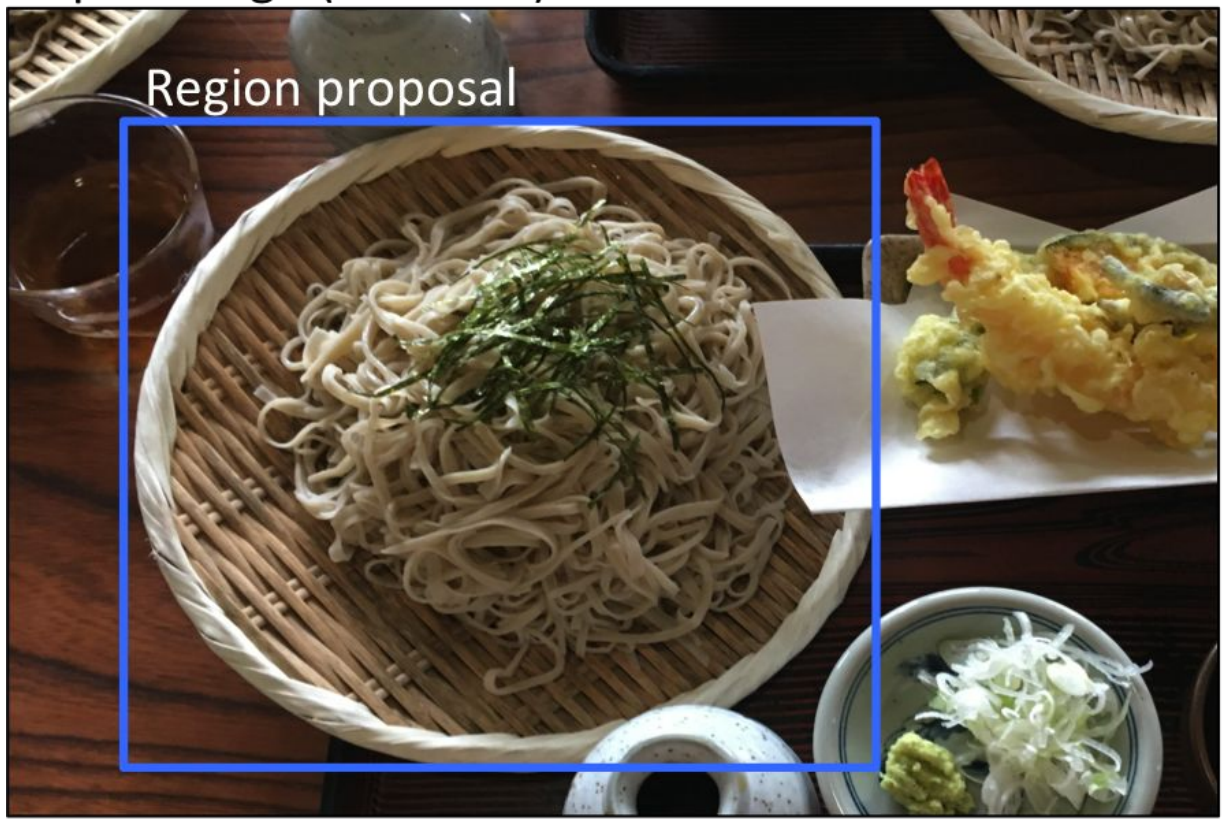
Mask-RCNN, ICCV 2017

Mask RCNN


Faster R-CNN
w/ ResNet [19]
RoI
7×7
×1024
res5
7×7
×2048
ave
2048
class
box
14×14
×256
14×14
×80
mask
Faster R-CNN
w/ FPN [27]
RoI
7×7
×256
1024
1024
class
box
RoI
14×14
×256
×4
14×14
×256
28×28
×256
28×28
×80
mask

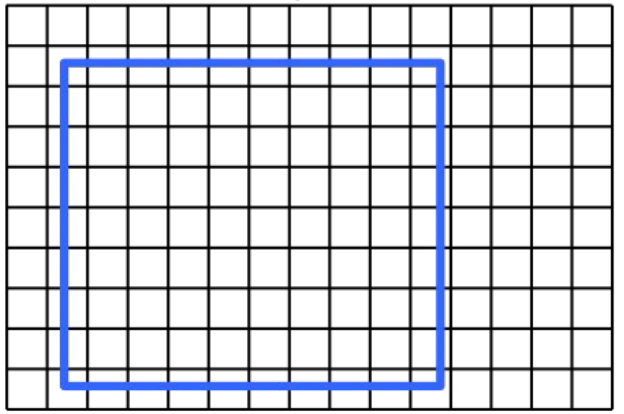
Mask RCNN

Input image (320x480)



CNN (stride = 32)

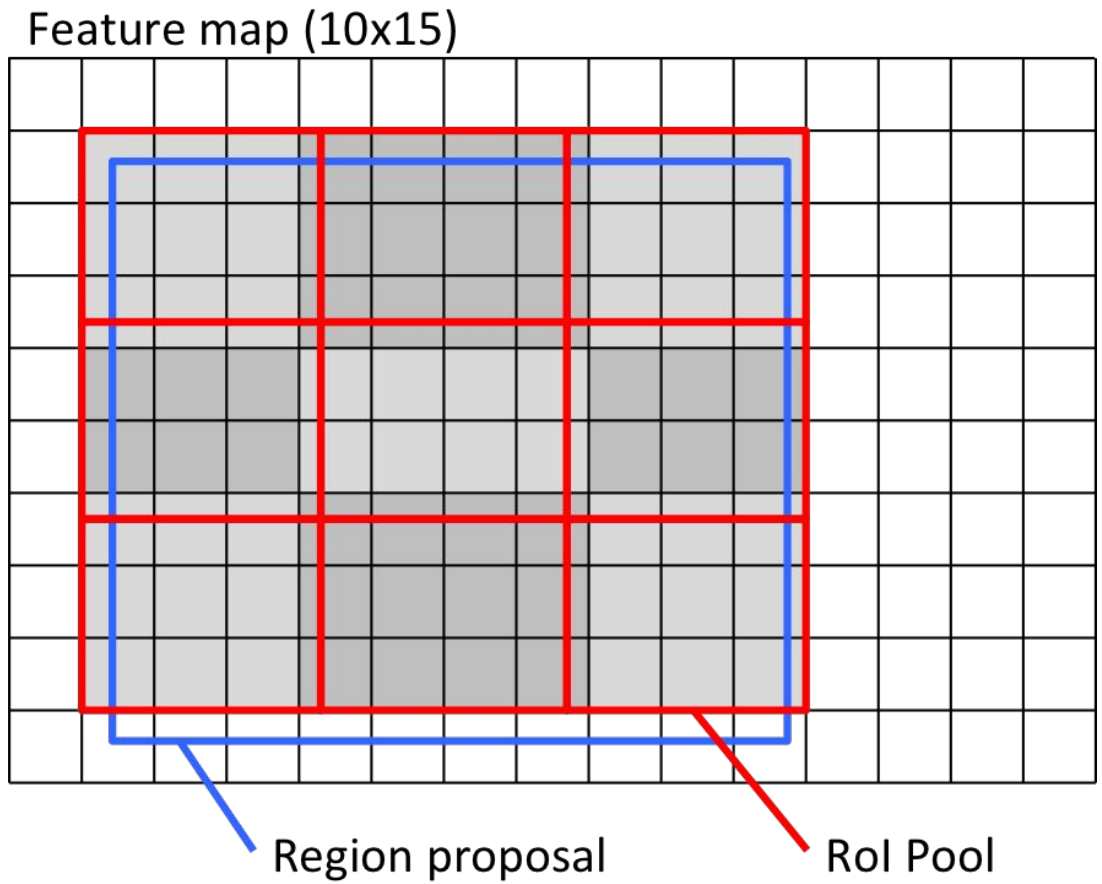
Feature map (10x15)



Face++ 旷视

https://qiita.com/yu4u/items/5cbe9db166a5d72f9eb8

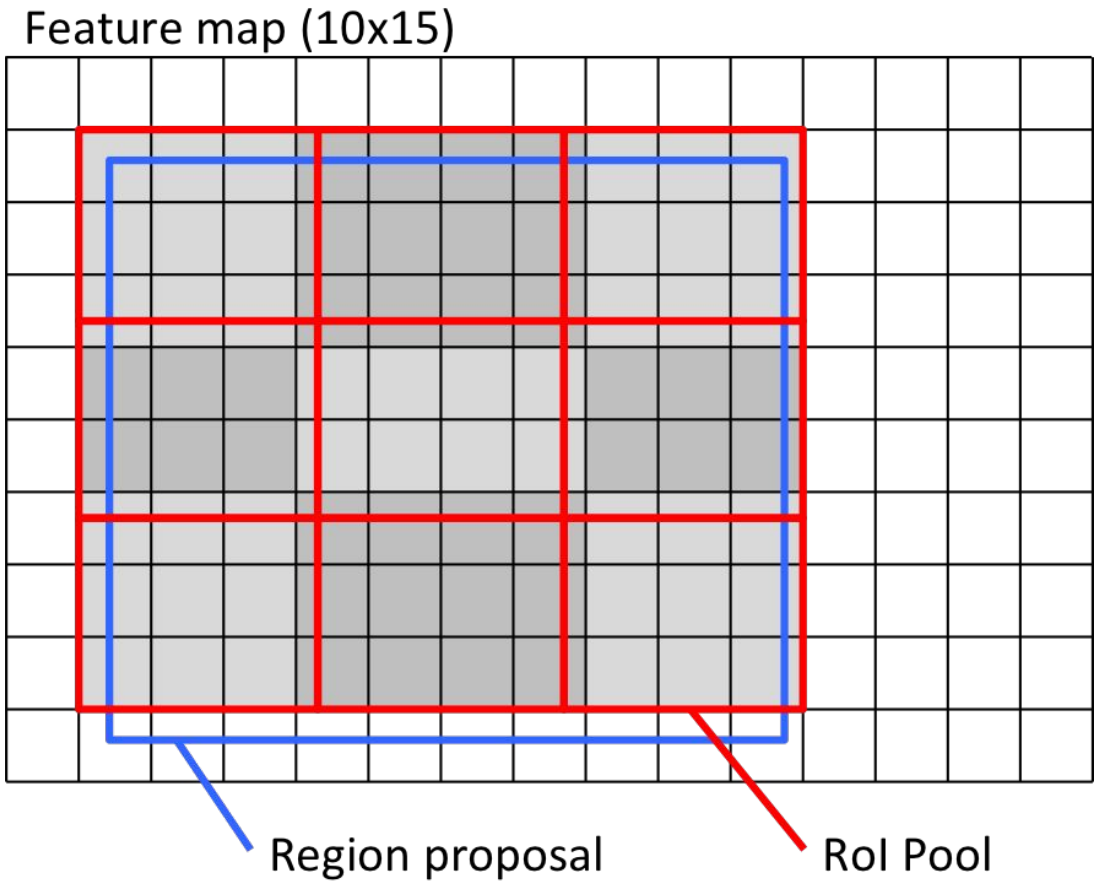
Mask RCNN

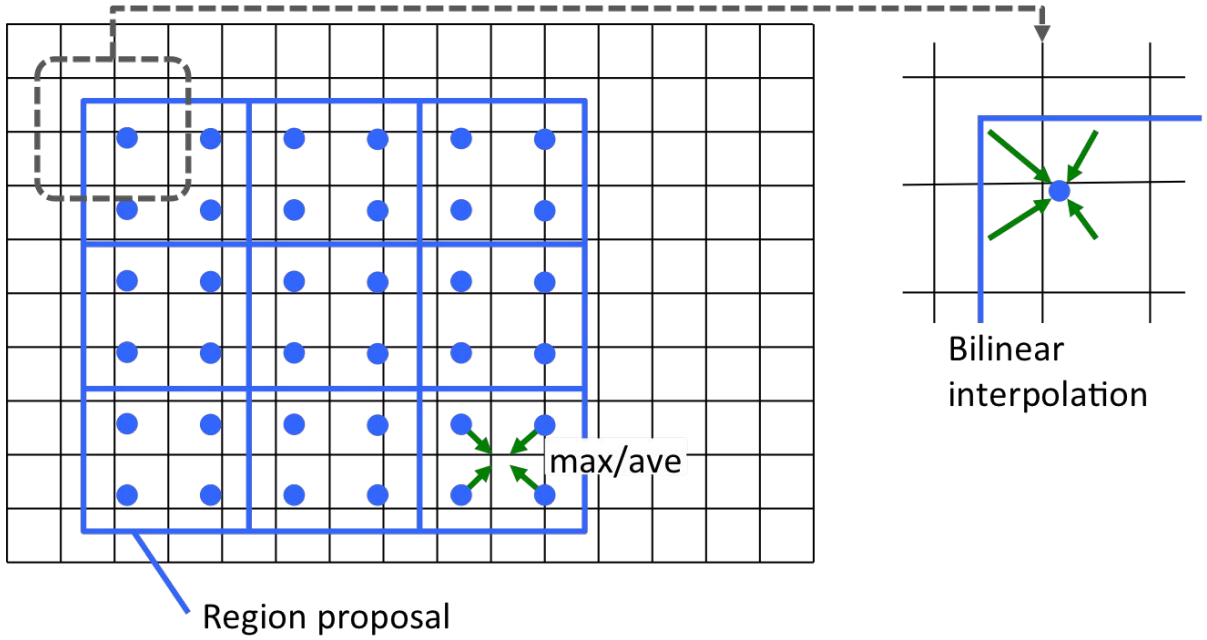


Face++ 旷视

https://qiita.com/yu4u/items/5cbe9db166a5d72f9eb8

Mask RCNN

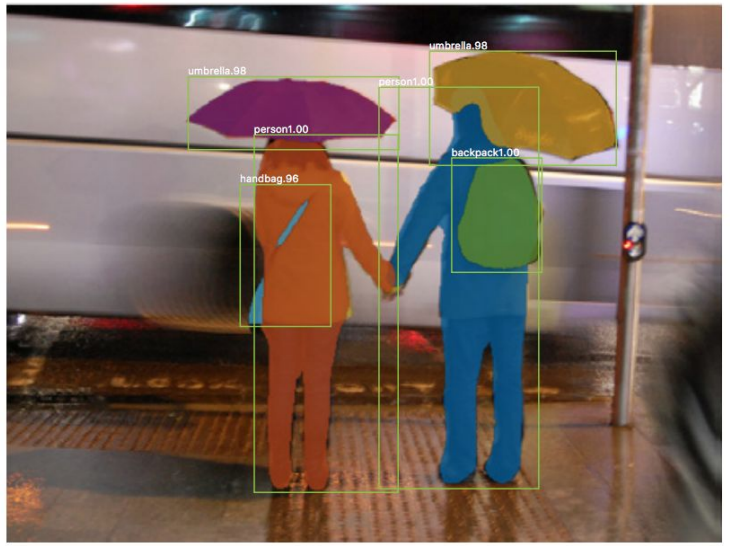


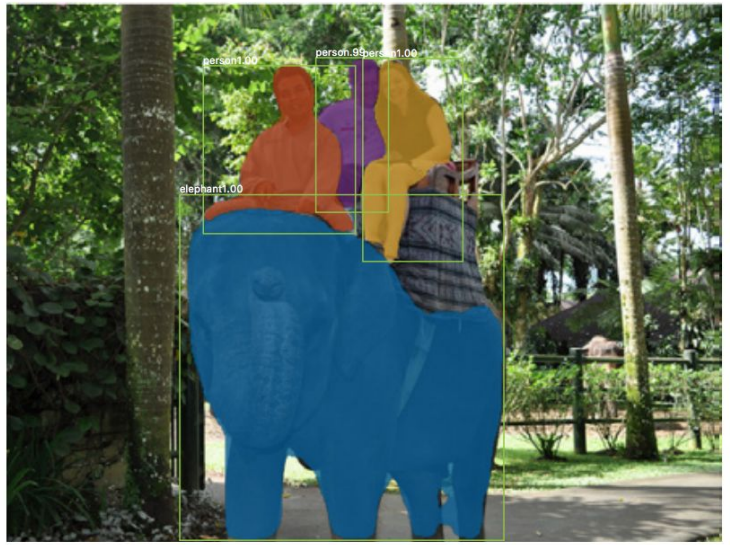


Face++ 旷视

https://qiita.com/yu4u/items/5cbe9db166a5d72f9eb8

Mask RCNN





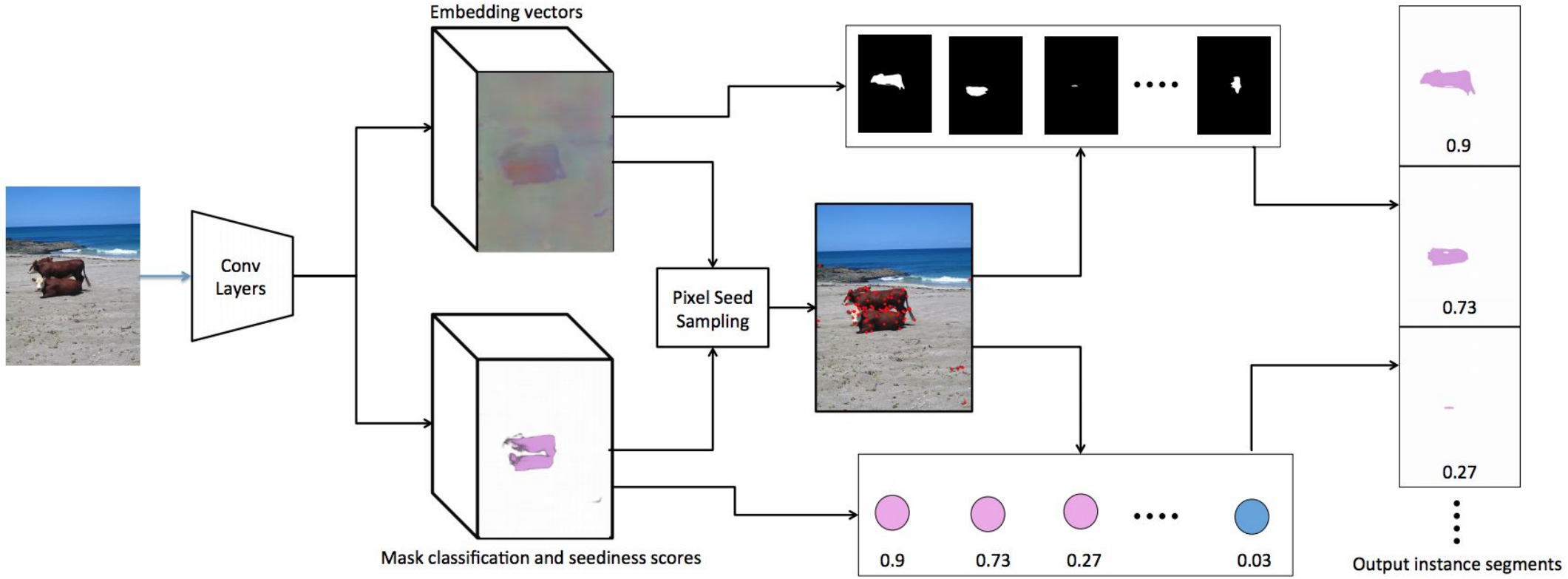
Face++ 旷视

Mask-RCNN, ICCV 2017

Mask RCNN

- State-Of-Art framework on InstanceSeg
- Best paper of ICCV 2017!
- Simple framework!

Bottom-up Pipeline

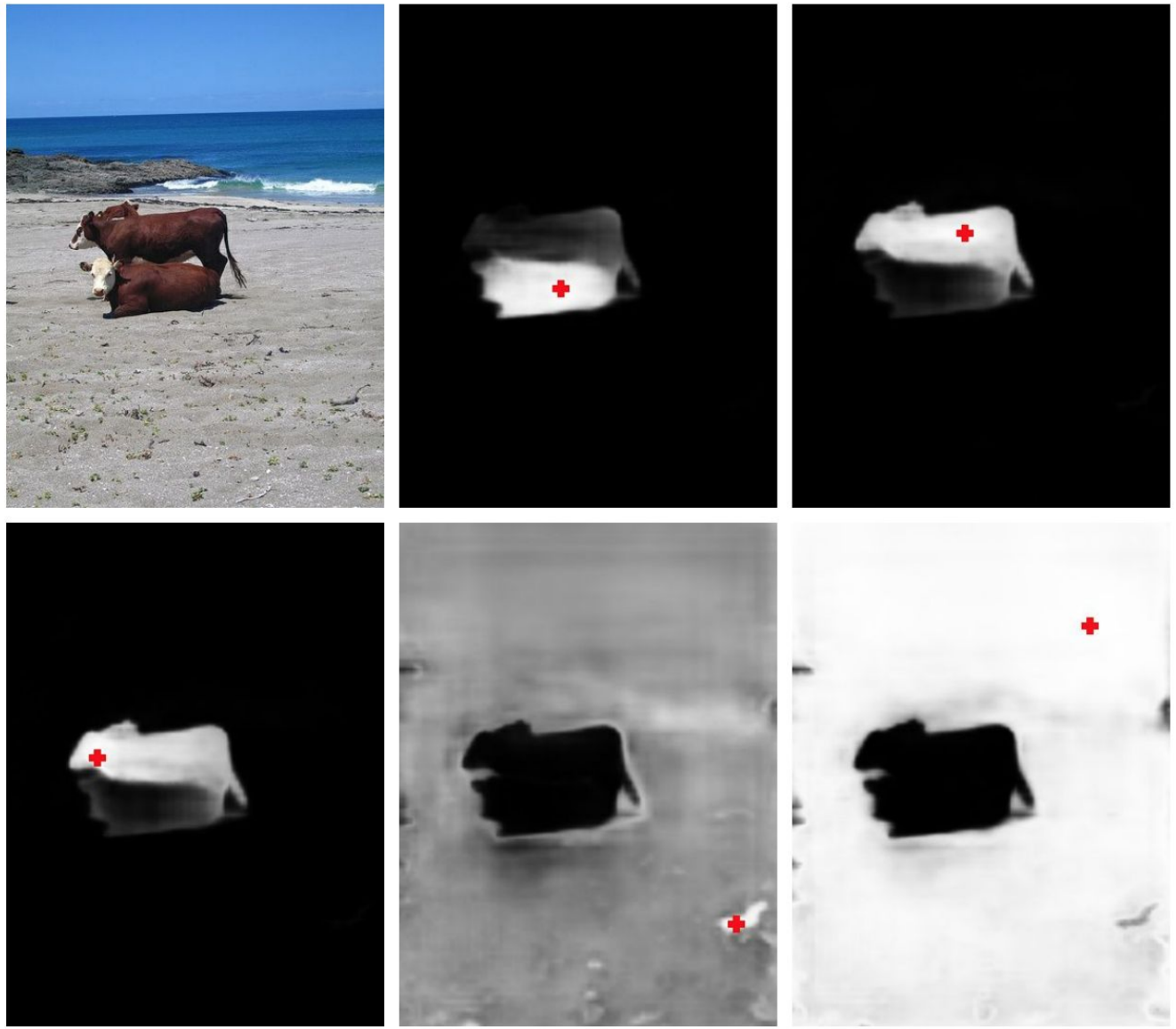


Bottom-up Pipeline

$$p_t = \arg \max_{p \notin p_{1:t-1}} [\log(S_p) + \alpha \log(D(p, p_{1:t-1})]$$

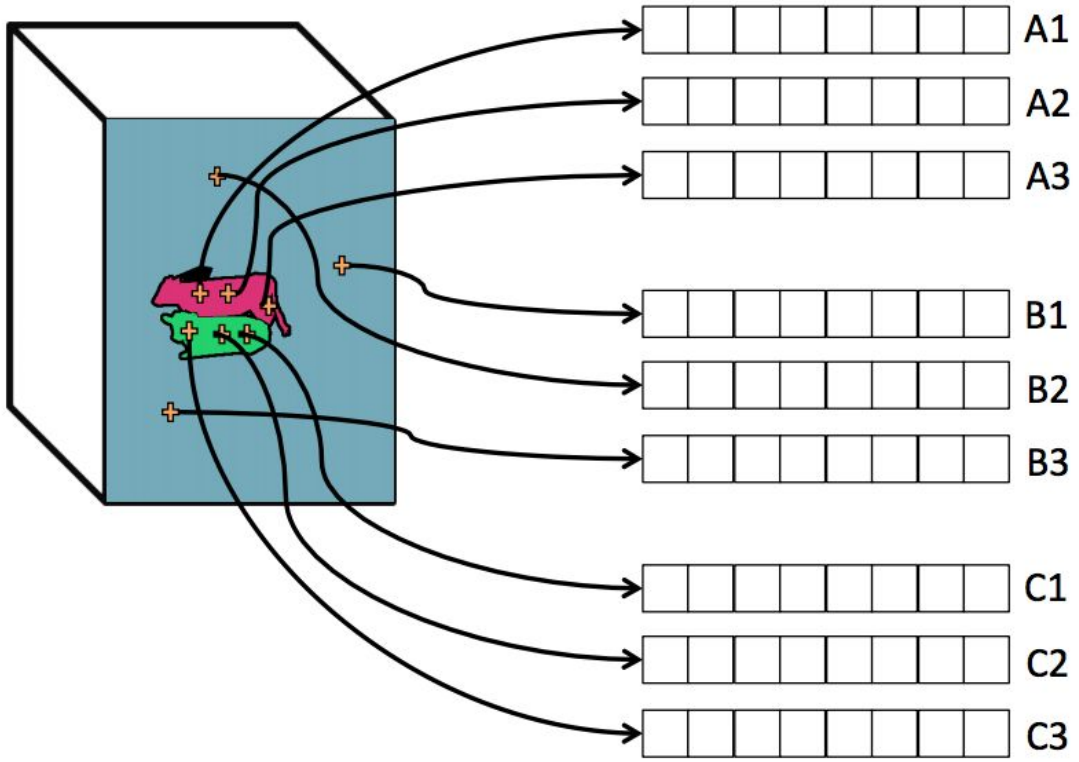
where

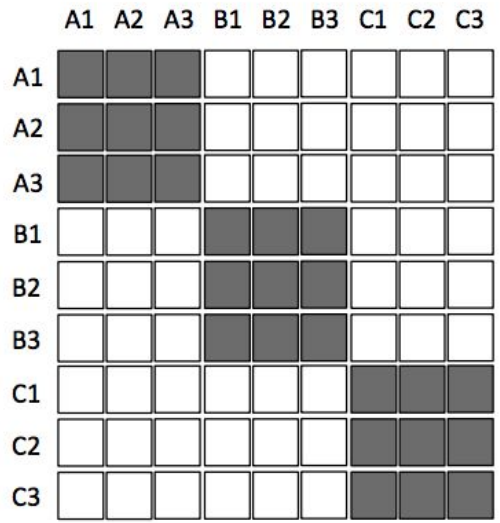
$$D(p, p_{1:t-1}) = \min_{q \in p_{1:t-1}} ||e_p - e_q||_2^2$$



Face++ 旷视

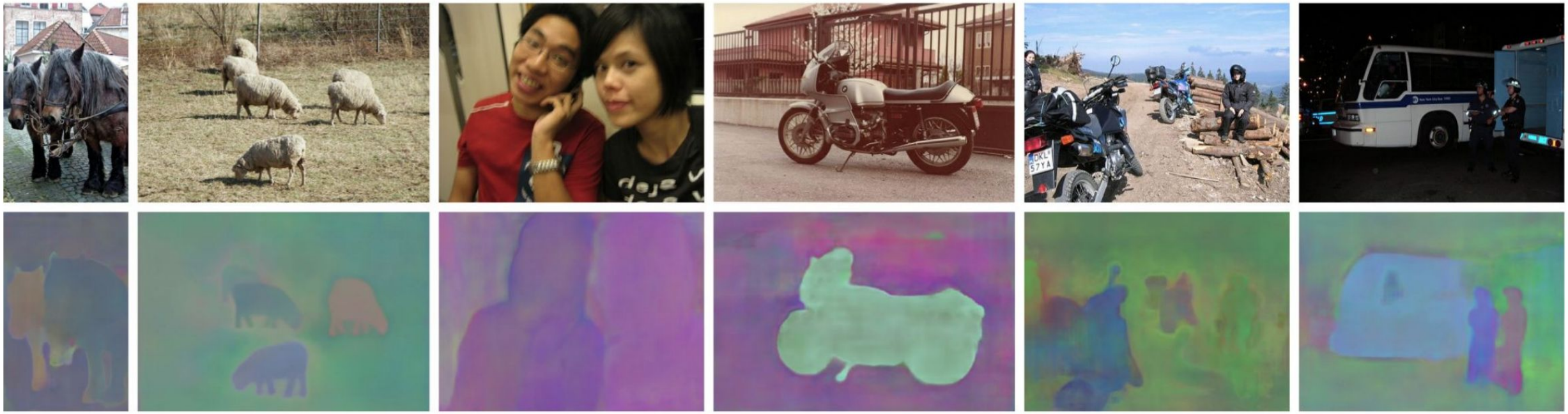
Bottom-up Pipeline





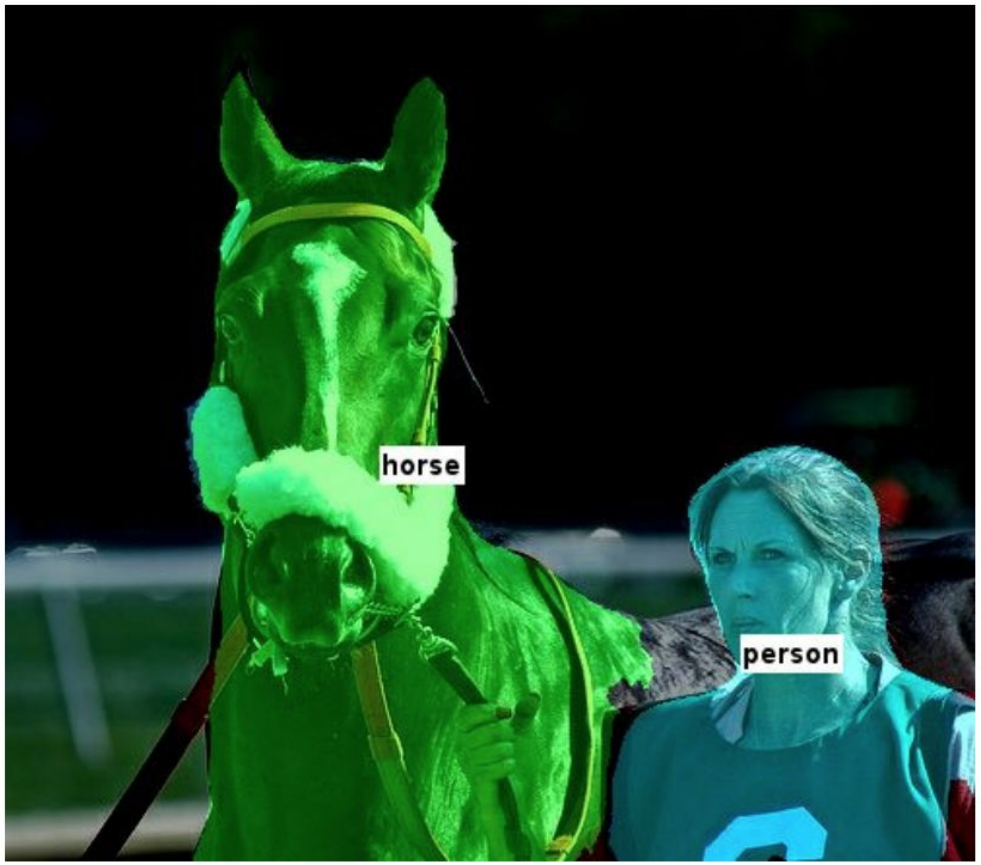
Sigmoid cross entropy loss on the similarity between pairs of embedding vectors

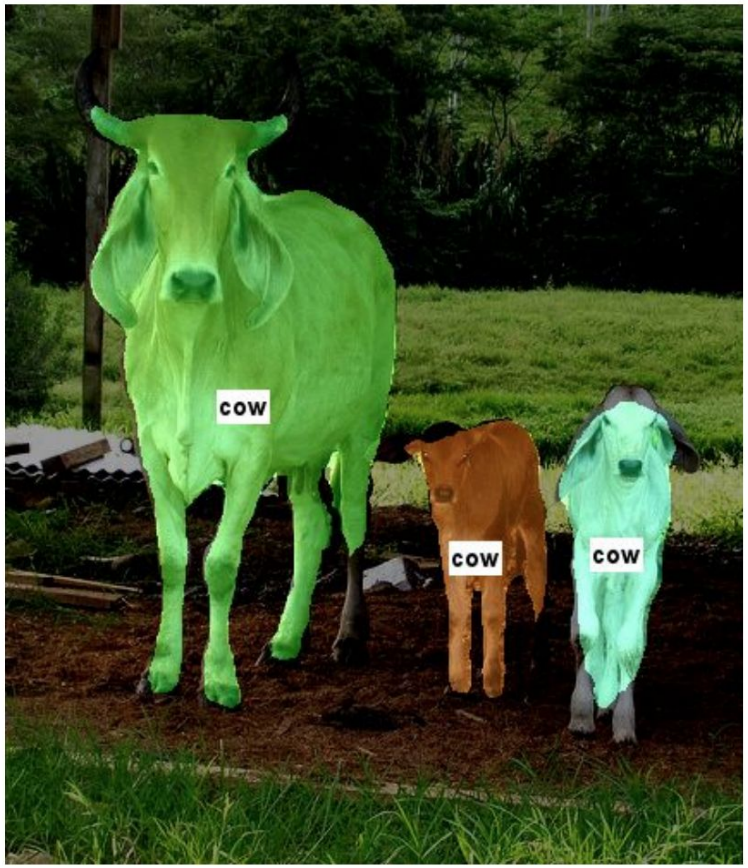
Bottom-up Pipeline



Face++ 旷视

Bottom-up Pipeline





Face++ 旷视

Semantic Instance Segmentation via Deep Metric Learning, arxiv

Bottom-up Pipeline

- Alternative framework to InstanceSeg
- Tricky to implement
- Incorporating the metric learning

Face++ 旷视

Top-Down

- Instance ---> Segmentation (for each instance)
- MainStream
- Start-Of-Art
- Easy to implement
- Difficulty: shrink the gap between det and seg

Bottom-Up

- Segmentation (for image) ---> instance
- Alternative
- Sub-Optimal
- Tricky to implement
- Difficulty: generate better instance

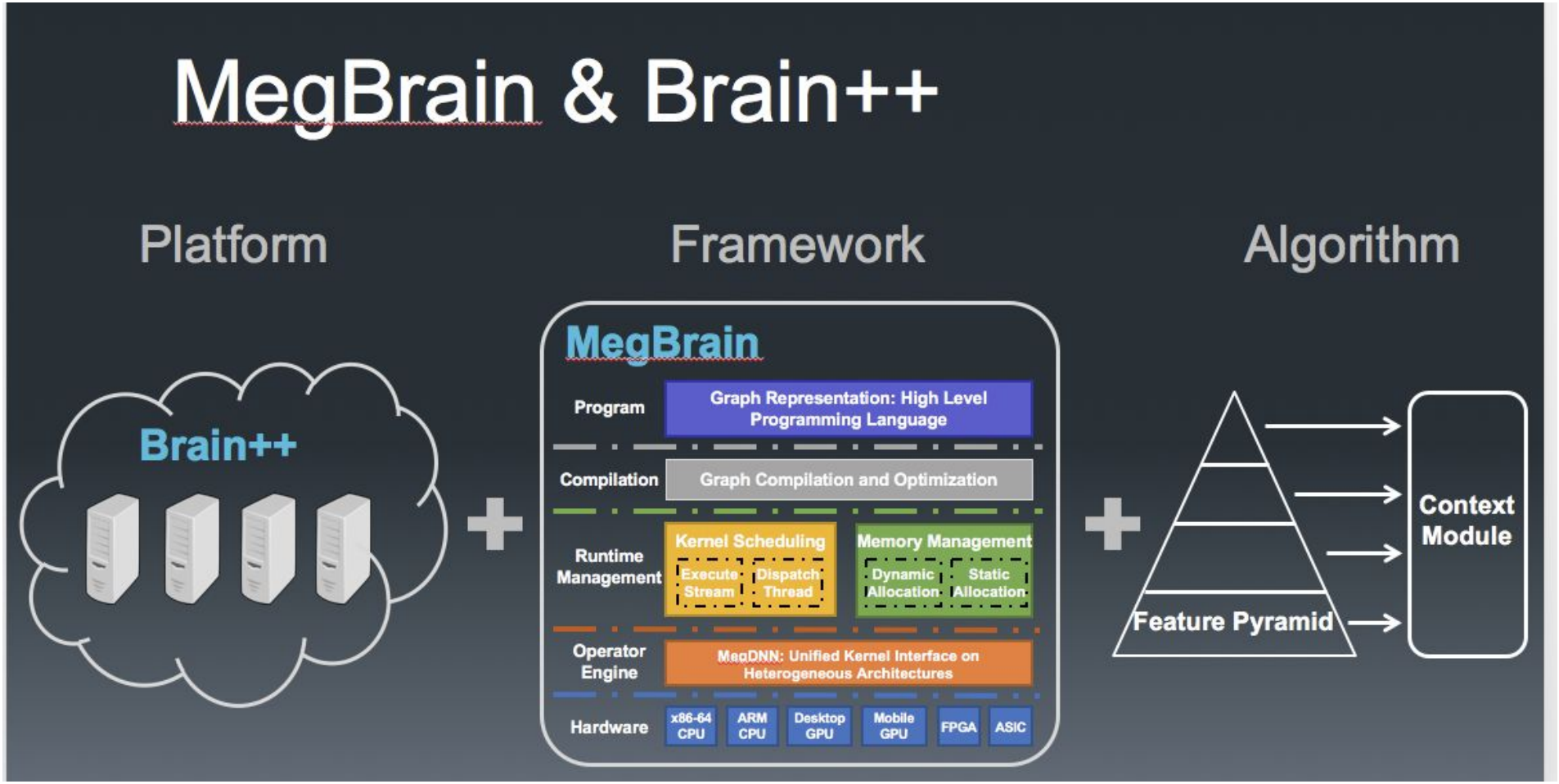
Re-Cap

- Segmentation with CNN: FCN, Deeplab, GCN ...
- Segmentation with CRF: DenseCRF, CRFAsRNN, ...
- Different Convolutions: Dilated Conv, Global Conv, Deformable, ...
- Top-Down pipeline for Instance Segmentation: FCIS, Mask-RCNN
- Bottom-Up pipeline

COCO & Places Challenge 2017

Face++ 旷视

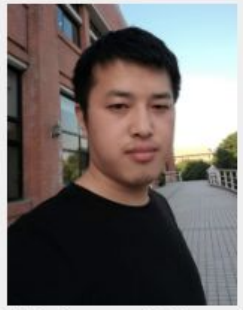
COCO & Places Challenge 2017



Face++ 旷视

COCO & Places Challenge 2017


Yilun Chen*


Zhicheng Wang*

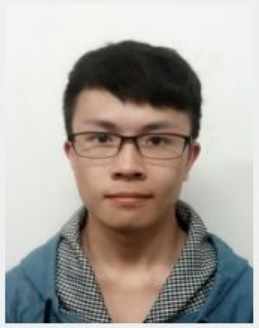

Xiangyu Peng

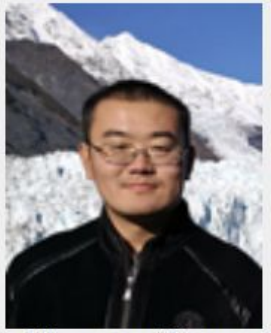

Zhiqiang Zhang


Gang Yu

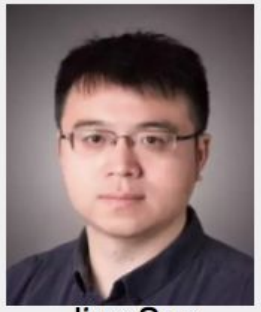

Chao Peng


Tete Xiao


Zeming Li

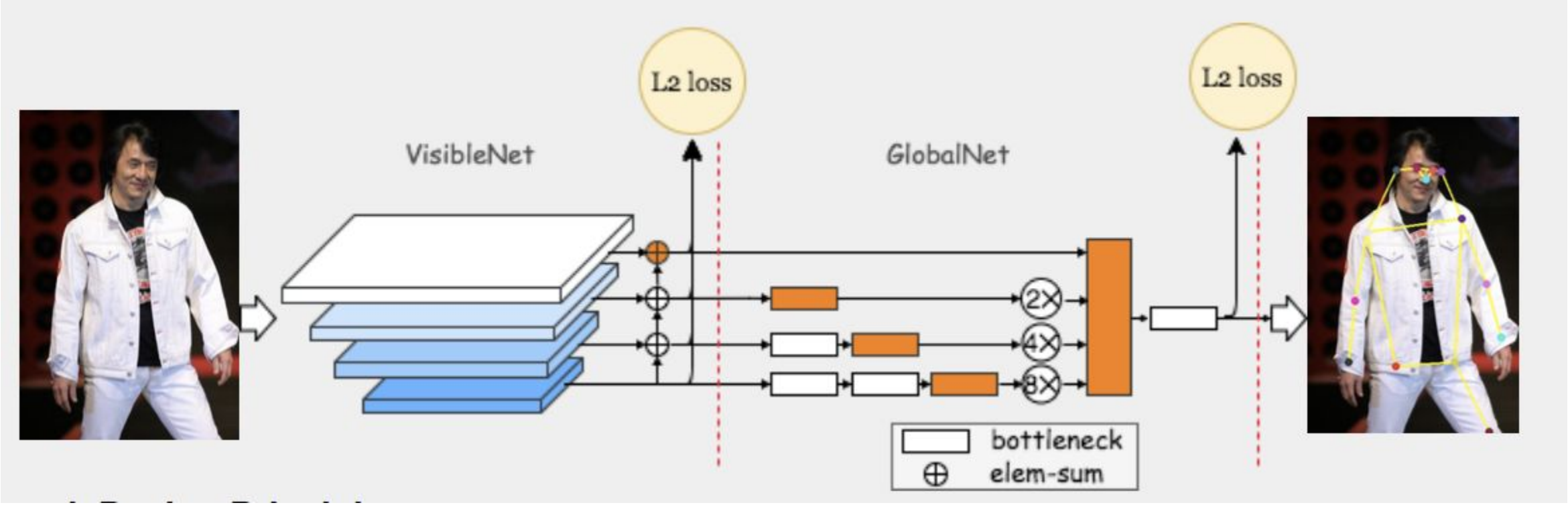

Xiangyu Zhang


Yuning Jiang


Jian Sun

Face++ 旷视

COCO & Places Challenge 2017



Face++ 旷视

COCO & Places Challenge 2017

Track	Rank	Ensemble	Single
COCO BBox Detection	1st	52.8	50.5
Places InstanceSeg	1st	30.7	28.7
COCO Keypoint	1st	72.6	70.9
COCO InstanceSeg	2nd	46.4	45.0

Face++ 旷视

COCO Challenge 2017 BBOX

Megvii (Face++)	0.526	0.730	0.585	0.343	0.556
UCenter	0.510	0.705	0.558	0.326	0.539
MSRA	0.507	0.717	0.566	0.343	0.529
FAIR Mask R-CNN	0.503	0.720	0.558	0.328	0.537

Face++ 旷视

COCO Challenge 2017 BBOX

Megvii (Face++)	0.526	0.730	0.585	0.343	0.556
UCenter	0.510	0.705	0.558	0.326	0.539
MSRA	0.507	0.717	0.566	0.343	0.529
FAIR Mask R-CNN	0.503	0.720	0.558	0.328	0.537

Face++ 旷视

Our Single Model is Here: 50.5.

Places Challenge 2017 InstanceSeg

Team Name	mean AP
Megvii (Face++)	0.2977
G-RMI	0.2415
BlueSky	0.1551
…	…
baseline	0.200

Megvii (Face++)
*Tete Xiao, *Ruixuan Luo, *Borui Jiang,
Shuai Shao, Yuning Jiang, Yadong Mu,
Jieqi Shi, Chi Zhang, Jian Sun
Megvii Research and Peking University

G-RMI
Alireza Fathi, Nori Kanazawa, Kai Yang,
Kevin Murphy
Google Research Machine Intelligence

Face++ 旷视

COCO Challenge 2017 Keypoint

Megvii (Face++)	0.721	0.905	0.789	0.679
oks	0.714	0.894	0.781	0.659
bangbangren	0.706	0.880	0.765	0.656
G-RMI	0.691	0.859	0.752	0.660
FAIR Mask R-CNN	0.689	0.892	0.752	0.637

Face++ 旷视

COCO Challenge 2017 Keypoint

SenseTime

Google Research

Megvii (Face++)	0.721	0.905	0.789	0.679
oks	0.714	0.894	0.781	0.659
bangbangren	0.706	0.880	0.765	0.656
G-RMI	0.691	0.859	0.752	0.660
FAIR Mask R-CNN	0.689	0.892	0.752	0.637

Face++ 旷视

COCO Challenge 2017 InstanceSeg

UCenter	0.463	0.694	0.507	0.258
Megvii (Face++)	0.460	0.707	0.503	0.263
FAIR Mask R-CNN	0.435	0.687	0.466	0.234
MSRA	0.426	0.680	0.460	0.240

Face++ 旷视

Thanks

Face++ 旷视