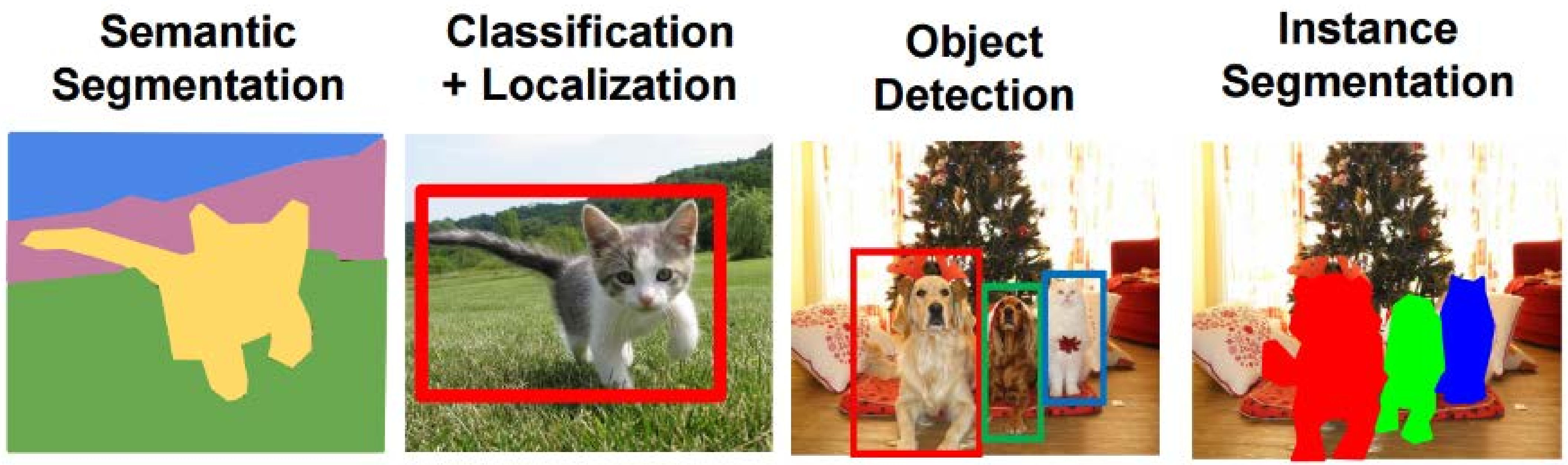
Learning Discriminative Feature Network for Semantic Segmentation

$Changqian\ Yu, et\ al.$, 2018, CVPR

2019/01/28, KangYeol Kim

DAVIAN
Data and Visual Analytics Lab

Tedious Recap.



From Weakly To Fully

Image-level label | Discriminative part of image

Weakly Supervised Localization →

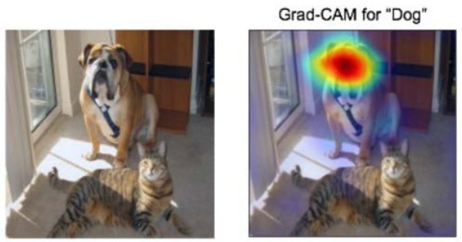
Image-level label | Pixel-level Classification

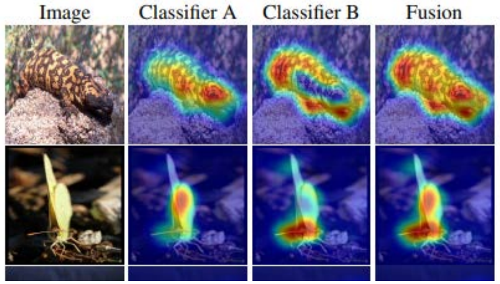
Weakly Supervised Semantic Segmentation →

Pixel-level label | Pixel-level Classification

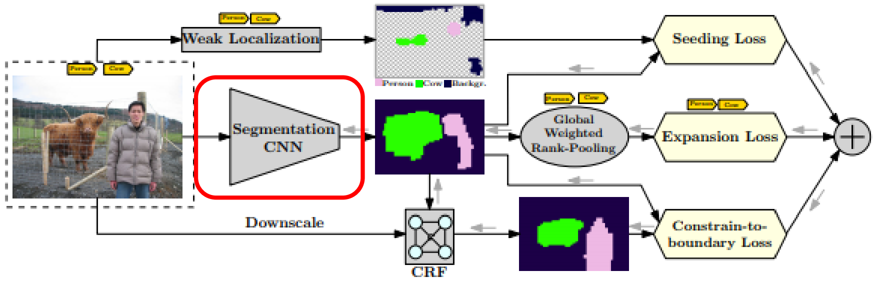
Fully Supervised Semantic Segmentation

- CAM
- Grad-CAM(2016)
- Two Phrase Learning for WSL(2017)
- ACoL for WSL(2018)

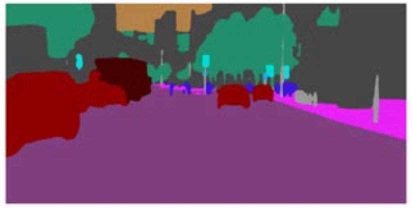




- SEC: Seed, Expand and Constrain(2016)
- SEC, Online PSL(Prohibitive Segmantation Learning)
- w/o GT, train **semantic segmentation network** using cues generated by weakly supervised manner



- FCN
- SegNet
- DeepLab v1
- PSPNet
- DeepLab v2
- **DFNet (Today's Paper)**
- Deep Lab v3
- DenseASPP



DAVIAN
Data and Visual Analytics Lab

Current Ranking

	mean	aero plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motor bike	person	potted plant	sheep	sofa	train	tv/ monitor	submission date
DeepLabv3+_JFT [?]	**89.0**	**97.5**	77.9	96.2	80.4	**90.8**	**98.3**	**95.5**	**97.6**	58.8	**96.1**	79.2	95.0	97.3	94.1	**93.8**	78.5	**95.5**	74.4	93.8	81.6	09-Feb-2018
SRC-B-MachineLearningLab [?]	88.5	97.2	78.6	**97.1**	80.6	89.7	97.4	93.7	96.7	59.1	95.4	81.1	93.2	**97.5**	94.2	92.9	73.5	93.3	74.2	91.0	85.0	19-Apr-2018
DeepLabv3+_AASPP [?]	88.5	97.4	**80.3**	**97.1**	80.1	89.3	97.4	94.1	96.9	**61.9**	95.1	77.2	94.2	**97.5**	**94.4**	93.0	72.4	93.8	72.6	93.3	83.3	22-May-2018
MSCI [?]	88.0	96.8	76.8	97.0	80.6	89.3	97.4	93.8	97.1	56.7	94.3	78.3	93.5	97.1	94.0	92.8	72.3	92.6	73.6	90.8	**85.4**	08-Jul-2018
ExFuse [?]	87.9	96.8	**80.3**	97.0	**82.5**	87.8	96.3	92.6	96.4	53.3	94.3	78.4	94.1	94.9	91.6	92.3	**81.7**	94.8	70.3	90.1	83.8	22-May-2018
DeepLabv3+ [?]	87.8	97.0	77.1	**97.1**	79.3	89.3	97.4	93.2	96.6	56.9	95.0	79.2	93.1	97.0	94.0	92.8	71.3	92.9	72.4	91.0	84.9	09-Feb-2018
DeepLabv3-JFT [?]	86.9	96.9	73.2	95.5	78.4	86.5	96.8	90.3	97.1	51.4	95.0	73.4	94.0	96.8	94.0	92.3	81.5	95.4	67.2	90.8	81.8	05-Aug-2017
DIS [?]	86.8	94.0	73.3	93.5	79.1	84.8	95.4	89.5	93.4	53.6	94.8	79.0	93.6	95.2	91.5	89.6	78.1	93.0	**79.4**	**94.3**	81.3	13-Sep-2017
** Gluon DeepLabV3 152 ** [?]	86.7	96.5	74.3	96.1	80.2	85.2	97.0	93.8	96.4	49.7	93.6	77.6	**95.1**	95.3	93.9	89.6	75.8	94.4	70.8	89.7	78.7	03-Oct-2018
CASIA_IVA_SDN [?]	86.6	96.9	78.6	96.0	79.6	84.1	97.1	91.9	96.6	48.5	94.3	78.9	93.6	95.5	92.1	91.1	75.0	93.8	64.8	89.0	84.6	29-Jul-2017
IDW-CNN [?]	86.3	94.8	67.3	93.4	74.8	84.6	95.3	89.6	93.6	54.1	94.9	79.0	93.3	95.5	91.7	89.2	77.5	93.7	79.2	94.0	80.8	30-Jun-2017
DFN [?]	86.2	96.4	78.6	95.5	79.1	86.4	97.1	91.4	95.0	47.7	92.9	77.2	91.0	96.7	92.2	91.7	76.5	93.1	64.4	88.3	81.2	15-Jan-2018

12^{th} placement @ VOC2012 *leader board*

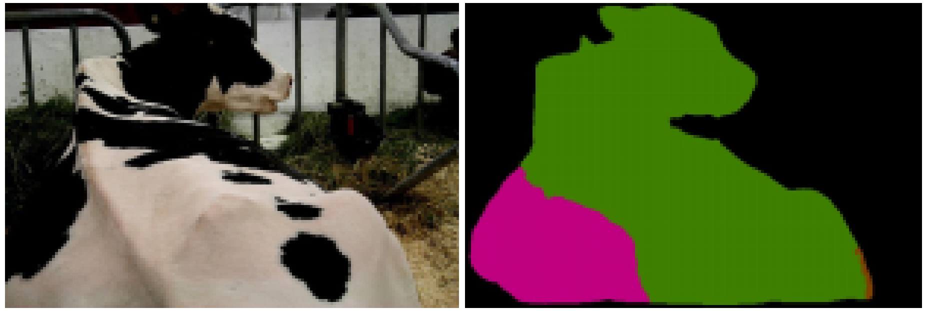
Main Idea:Two Weaknesses in SS

1. **Intra-class inconsistency**
The patches which share the
same semantic label
but
different appearances

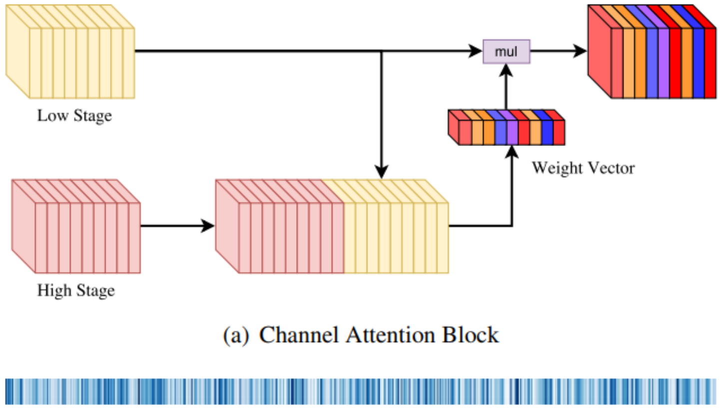
2. **Inter-class indistinction**
The two adjacent patches which have
different semantic labels
but with
similar appearances

Intra-class Inconsistency

- **What?**
  - The predictions can be incontinuous without delicate consideration of neighboring pixels
- **Why?**
  - Mainly due to **LACK OF CONTEXT**
- **How?**
  - Combining different scale context
    - PSPNet, Deeplab v3
  - [=>] Utilizing the inherent multi-scale context of different stages
    - **[LIMIT]** Just summing up the features by channel(RefineNet) => Ignores the diverse consistency in different stages
    - **[+] GAP @ last layer => Add global information**
    - **[+] Channel Attention Block to utilize different consistency information**



(a) Input (b) FCN Based Model

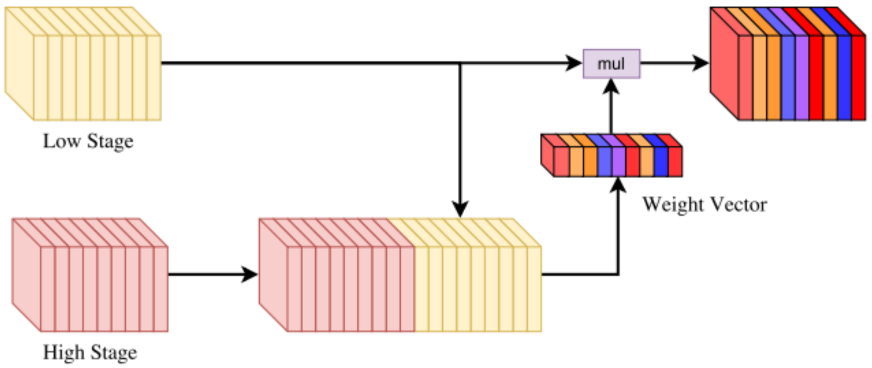


(a) Channel Attention Block

(b) Attention Vector

Intra-class Inconsistency

- The features in different stages have different degrees of discrimination, which results in different consistency prediction.
- In order to obtain the intra-class consistent prediction, we should extract the discriminative features and inhibit the indiscriminative features.
- Motivated by SENet, this paper adapted channel-wise weight parameters. With this, then network can obtain discriminative features stage-wise



(a) Channel Attention Block

(b) Attention Vector

$$\bar{y} = \alpha y = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ \vdots \\ y_K \end{bmatrix} = \begin{bmatrix} \alpha_1 w_1 \\ \vdots \\ \alpha_K w_K \end{bmatrix} \times \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix} \quad (3)$$

where $\bar{y}$ is the new prediction of network and $\alpha = Sigmoid(x; w)$

Inter-class indistinction

- **What?**
  - The predictions can have misconception regarding the object which has a similar appearance.
- **Why?**
  - Mainly due to **VAGUE BOUNDARY**
- **How?**
  - Semantic boundary to guide the learning of the features => [+] Variational features
  - Details:
    - GT – Apply 'canny edge detection' on GT SS labels => Reshape it into (# of classes, H, W) where The channel in the part where the true label and edge are located is full of 1's
    - The output of Board Network is also (# of classes, H, W)

(d) Input　　(e) FCN Based Model

$$FL(p_k) = -(1 - p_k)^{\gamma} \log p_k$$

- ✓ Fosal loss to train hard for abstruse cases
- ✓ $p_k \uparrow \;\Rightarrow\; Weight \downarrow$
- ✓ $p_k \downarrow \;\Rightarrow\; Weight \uparrow$

Full Structure

$$\ell_s = SoftmaxLoss(y; w)$$
$$\ell_b = FocalLoss(y; w)$$
$$L = \ell_s + \lambda\, \ell_b$$

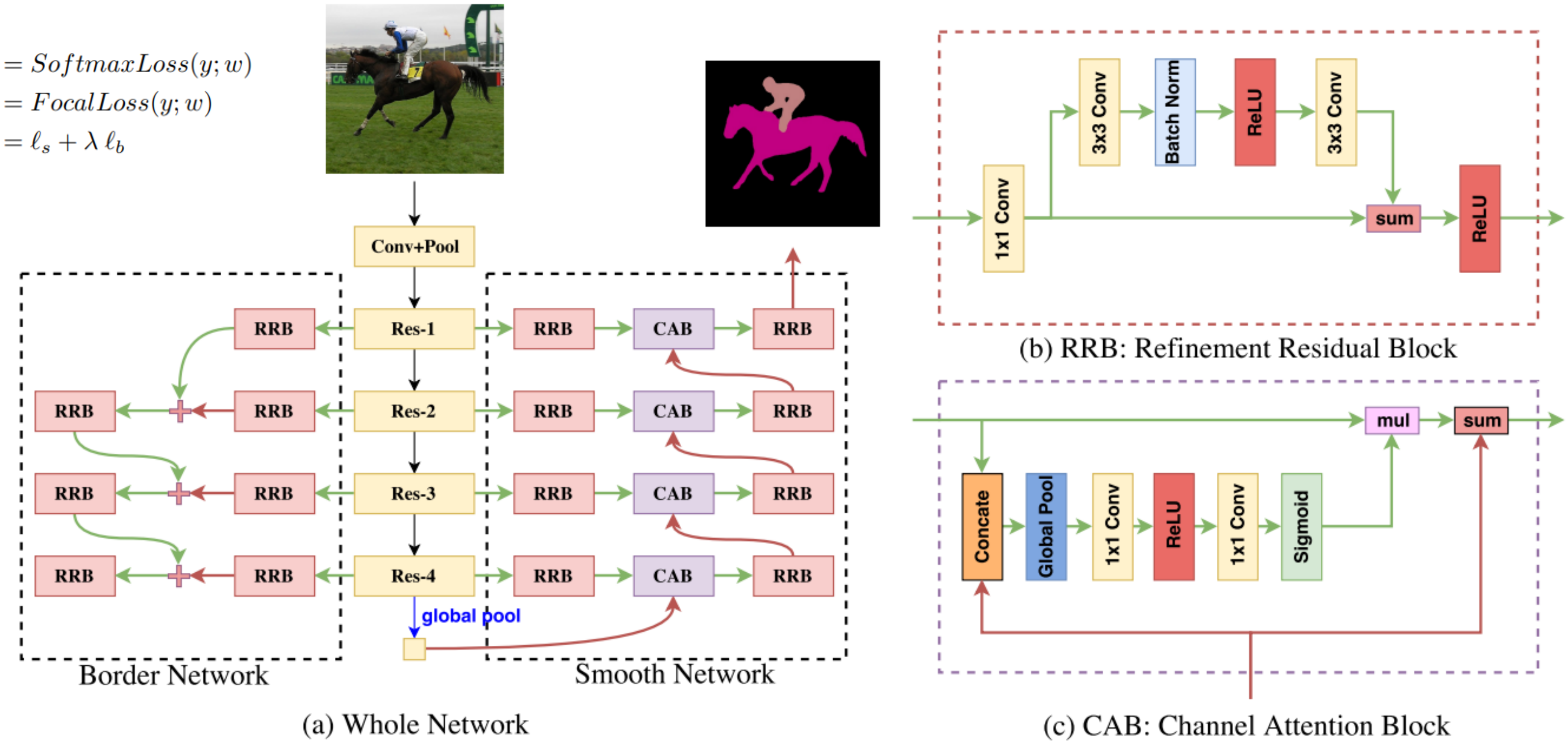


Figure 2. An overview of the Discriminative Feature Network. (a) Network Architecture. (b) Components of the Refinement Residual Block (RRB). (c) Components of the Channel Attention Block (CAB). The red and blue lines represent the upsample and downsample operators, respectively. The green line can not change the size of feature maps, just a path of information passing.

Smooth Network Effect

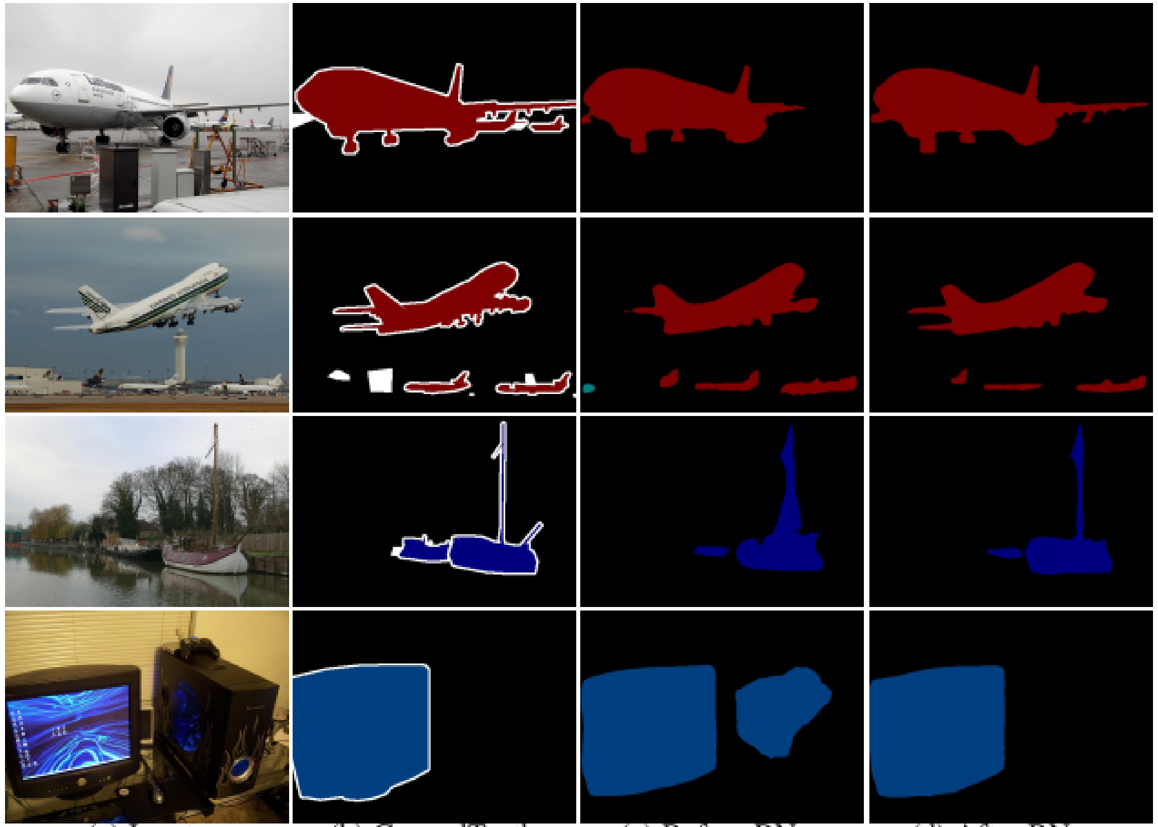
(a) Input (b) GroundTruth (c) Before SN (d) After SN

Method	Mean IOU(%)
Res-101	72.86
Res-101+RRB	76.65
Res-101+RRB+GP	78.20
Res-101+RRB+GP+CAB	79.31
Res-101+RRB+DS	77.08
Res-101+RRB+GP+DS	78.51
Res-101+RRB+GP+CAB+DS	79.54

GP – Global Pooling
DS – Deep supervision (Add auxiliary loss)

DAVIAN
Data and Visual Analytics Lab

Border Network Effect

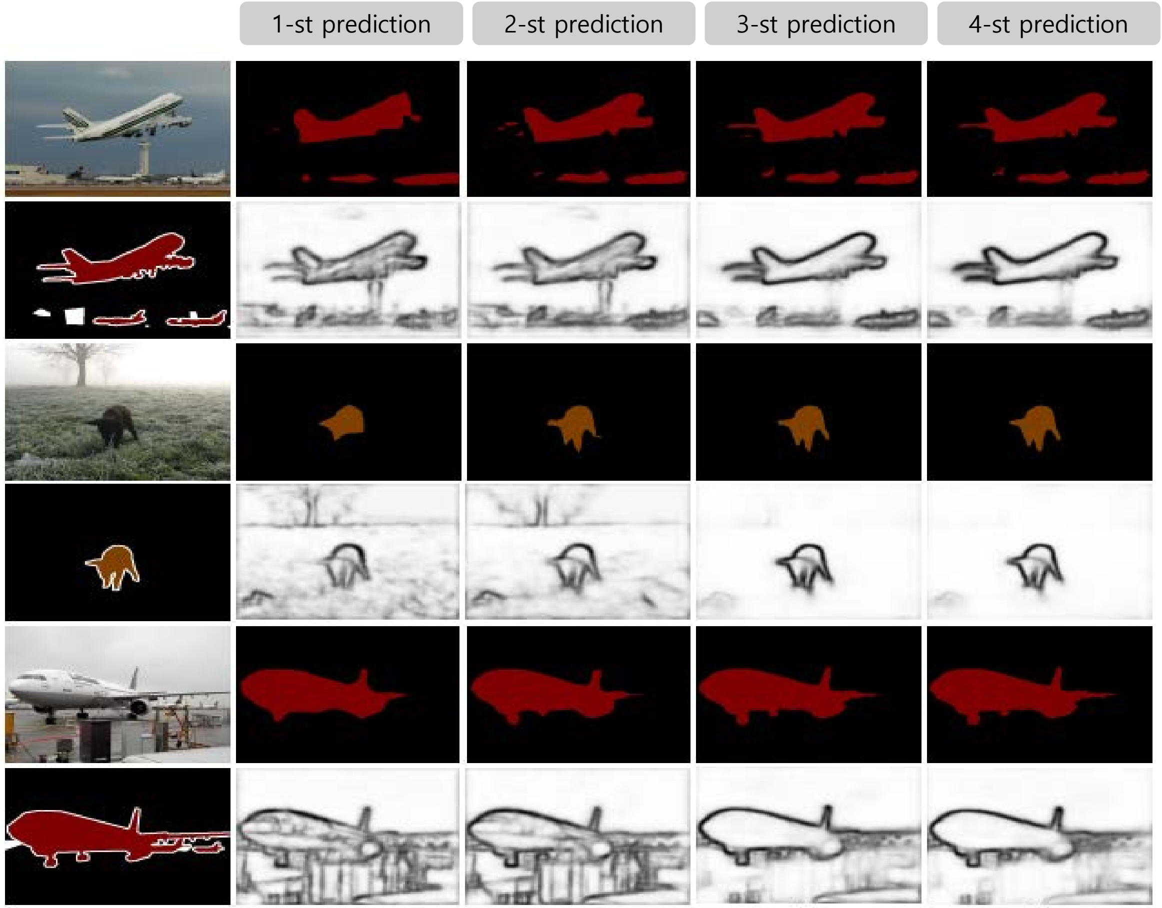


(a) Input (b) GroundTruth (c) Before BN (d) After BN

Method	Mean IOU(%)
Res-101+SN	79.54
Res-101+SN+BN	79.67
Res-101+SN+MS_Flip	79.90
Res-101+SN+BN+MS_Flip	80.01

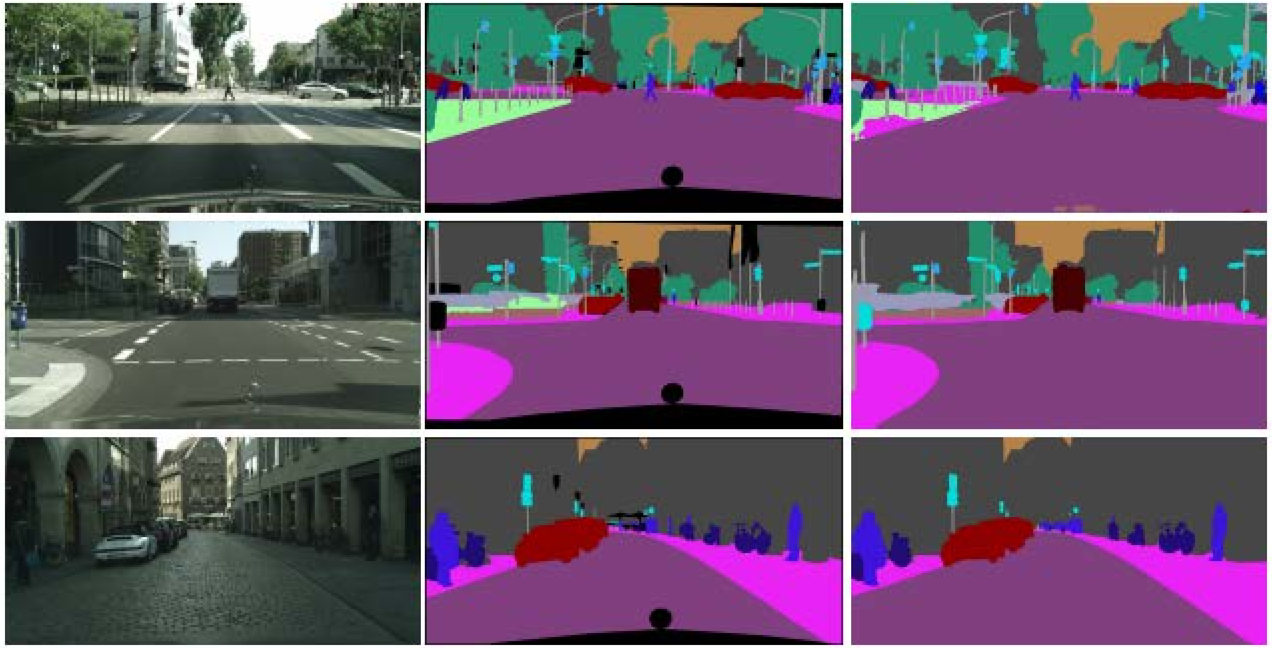
MS – Multi-scale training
It can be possible since last upsampling layer resize last feature map into original size at any given sized one.

Multi-stage Effect + VOC2012



Method	Mean IOU(%)
FCN [27]	62.2
Zoom-out [29]	69.6
ParseNet [24]	69.8
Deeplab v2-CRF [5]	71.6
DPN [26]	74.1
Piecewise [20]	75.3
LRR-CRF [11]	75.9
PSPNet [40]	82.6
Ours	**82.7**
DLC^{+} [18]	82.7
DUC^{+} [34]	83.1
GCN^{+} [30]	83.6
RefineNet^{+} [19]	84.2
ResNet-38^{+} [35]	84.9
PSPNet^{+} [40]	85.4
Deeplab v3^{+} [6]	85.7
Ours^{+}	**86.2**

Cityscapes data result



(a) Input (b) GroundTruth (c) DFN

Method	Mean IOU(%) w/o coarse	Mean IOU(%) w/ coarse
CRF-RNN [41]	62.5	-
FCN [27]	65.3	-
DPN [26]	66.8	59.1
LRR [11]	69.7	71.8
Deeplab v2-CRF [5]	70.4	-
Piecewise [20]	71.6	-
RefineNet [19]	73.6	-
SegModel [10]	78.5	79.2
DUC [34]	77.6	80.1
PSPNet [40]	78.4	80.2
Ours	**79.3**	**80.3**

Thanks a lot !!
Any Questions?