Learnable Histogram: Statistical Context Features for Deep Neural Networks

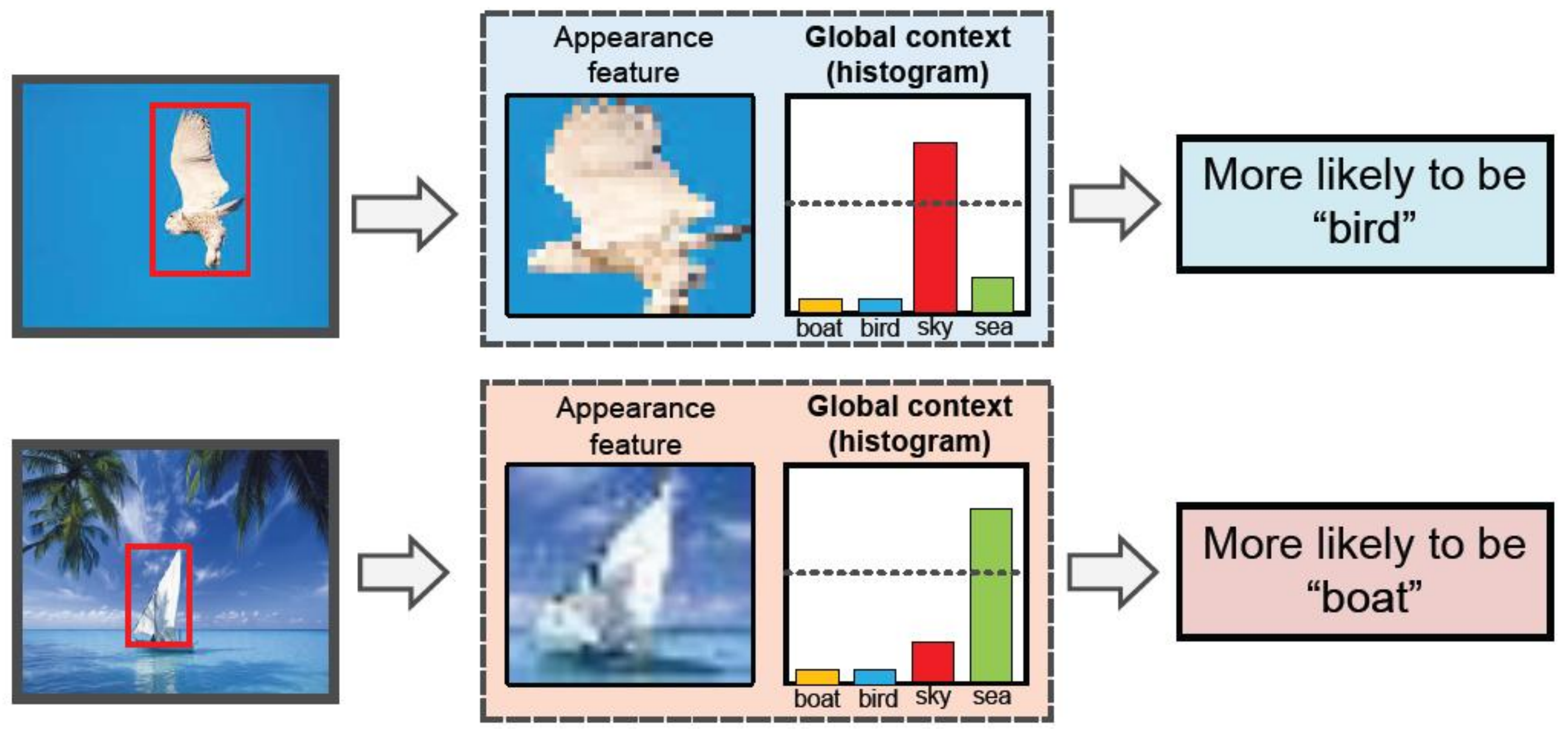
The Chinese University of Hong Kong
香港中文大學

Zhe Wang, Hongsheng Li, Wanli Ouyang, Xiaogang Wang
Department of Electronic Engineering, The Chinese University of Hong Kong
{zwang, hsli, wlouyang, xgwang}@ee.cuhk.edu.hk

◆ Motivation

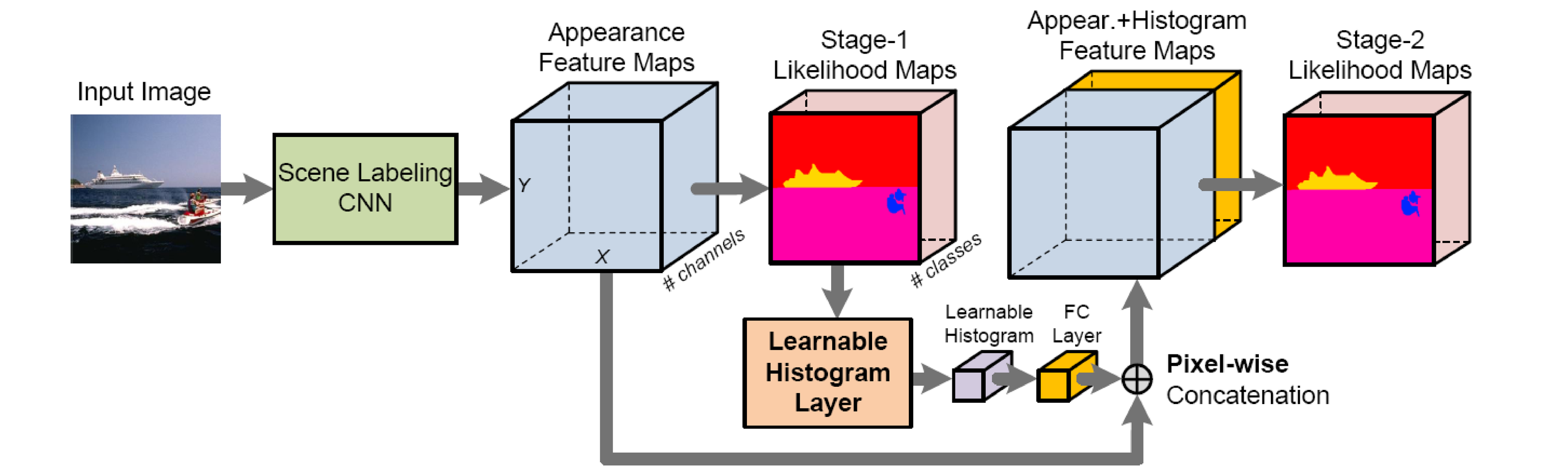
➢ Statistical context features are useful in classification problems, but currently can not be jointly optimized with the deep model.

➢ We propose a learnable histogram layer which can be trained within a deep model in an end-to-end manner.

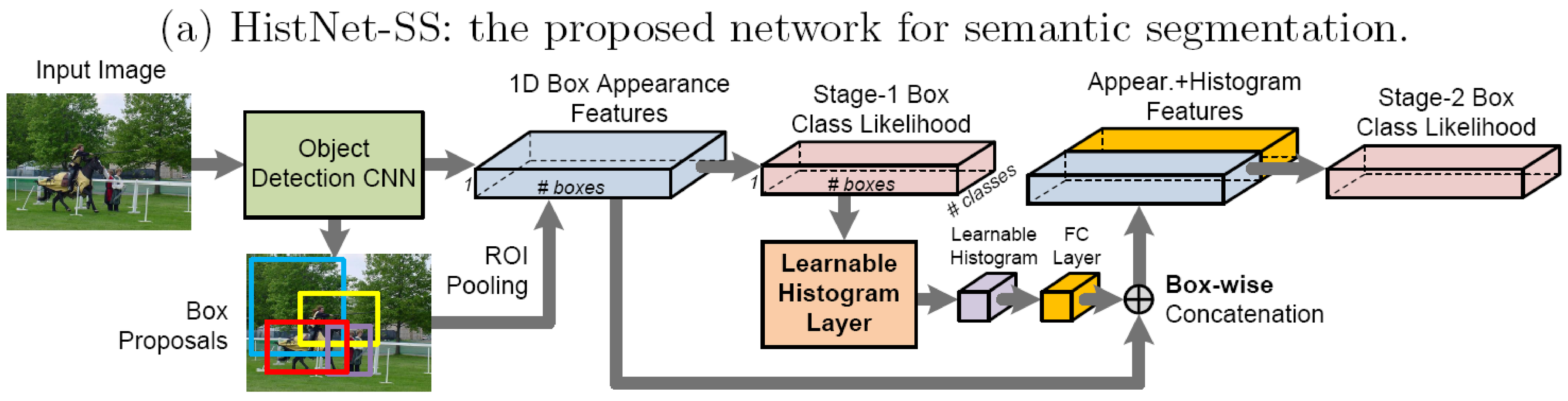


◆ Combining histogram features and deep models

➢ We designed two networks combined with the proposed histogram features for semantic segmentation and object detection, respectively.



(a) HistNet-SS: the proposed network for semantic segmentation.



(b) HistNet-OD: the proposed network for object detection.

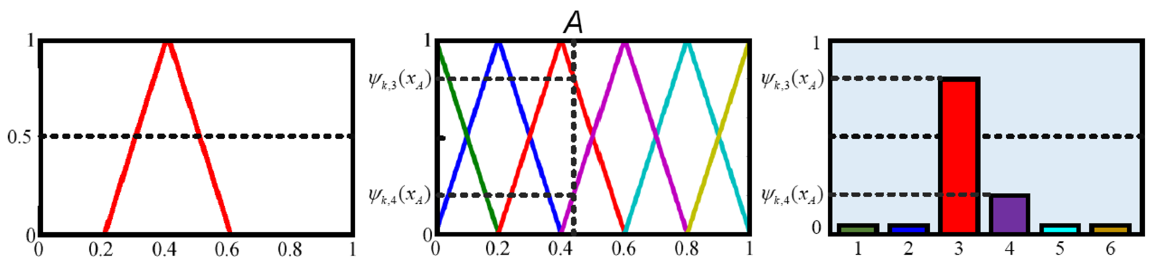
◆ Learnable histogram layer

➢ The bth bin of class k in the learnable histogram is modeled by a piecewise linear basis function:

$$\psi_{k,b}(x_k) = \max\{0, 1 - |x_k - \mu_{k,b}| \times w_{k,b}\}$$

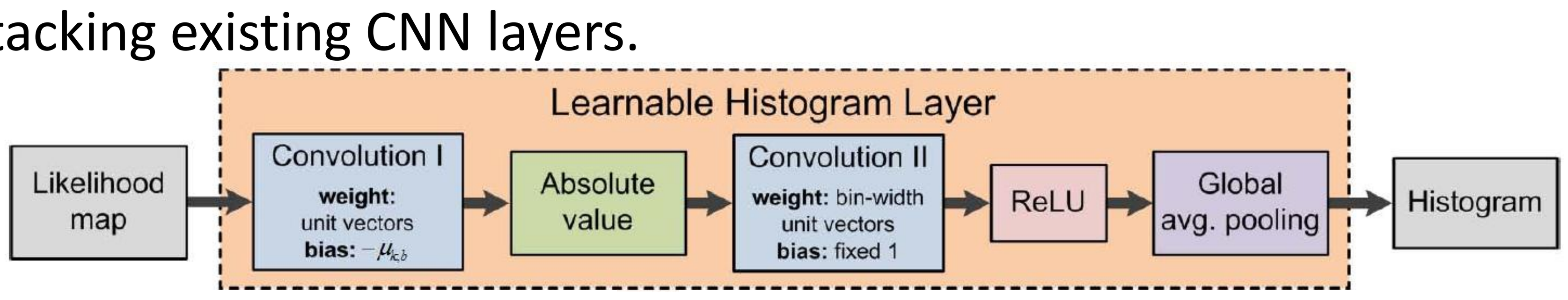
➢ The bin centers and widths can be updated by stochastic gradient descent:

$$\frac{\partial E}{\partial w_{k,b}} = \begin{cases} |x_k - \mu_{k,b}|, & \psi_{k,b}(x_k) > 0, \\ 0, & \text{otherwise.} \end{cases} \qquad \frac{\partial E}{\partial u_{k,b}} = \begin{cases} w_{k,b}, & \psi_{k,b}(x_k) > 0 \text{ and } x_k - \mu_{k,b} > 0, \\ -w_{k,b}, & \psi_{k,b}(x_k) > 0 \text{ and } x_k - \mu_{k,b} < 0, \\ 0, & \text{otherwise.} \end{cases}$$



(a) (b) (c)

➢ The proposed learnable histogram layer can be modeled by stacking existing CNN layers.



$$H^{\mathrm{I}}_{k,b}(c) = \begin{cases} 1, & c = k, \\ 0, & \text{otherwise,} \end{cases} \qquad H^{\mathrm{II}}_{k,b}(c) = \begin{cases} w_{k,b}, & c = (k-1)B + k, \\ 0, & \text{otherwise.} \end{cases}$$

where $H^{\mathrm{I}}_{k,b}(c)$ and $H^{\mathrm{II}}_{k,b}(c)$ are kernels for Convolution I and II.

◆ Experiments

➢ Datasets

➢ Semantic segmentation
- SIFTFlow
- Stanford Background
- PASCAL VOC 2012 segmentation

➢ Object detection
- PASCAL VOC 2007 detection

➢ Base models

➢ Our learnable histogram layer can be flexibly integrated into various network structures. The Base models are selected as follows:

Datasets	Base model
SIFTFlow	VGG-FCN [1]
Stanford Background	VGG-FCN [1]
PASCAL VOC 2012 segmentation	Deeplab [2]
PASCAL VOC 2007 detection	Faster-RCNN [3]

➢ Results

➢ SIFTFlow and Stanford Background

Methods	Per-pixel	Per-class
Tighe et al. [32]	0.769	0.294
Liu et al. [25]	0.748	n/a
Farabet et al. [12]	0.785	0.296
Pinheiho et al. [18]	0.777	0.298
Sharma et al. [29]	0.796	0.336
Yang et al. [1]	0.798	0.487
Eigen et al. [31]	0.868	0.464
FCN [19]	0.851	**0.517**
FCN (our implement)	0.860	0.457
FCN+FC-CRF	0.865	0.468
HistNet-SS stage-1	0.876	0.505
HistNet-SS	**0.879**	0.5
HistNet-SS+FC-CRF	**0.879**	0.512

(a) SIFTFlow dataset

Method	Per-pixel	Per-class
Gould et al. [2]	0.764	n/a
Tighe et al. [32]	0.775	n/a
Socher et al. [28]	0.781	n/a
Lempitzky et al. [33]	0.819	0.724
Farabet et al. [12]	0.814	0.76
Pinheiho et al. [18]	0.802	0.699
Sharma et al. [29]	0.823	0.791
FCN (our implement)	0.851	0.811
FCN+FC-CRF	0.862	0.82
FCN+MOPCNN [17]	0.863	0.811
HistNet-SS stage-1	0.871	**0.838**
HistNet-SS	0.871	0.837
HistNet-SS+FC-CRF	**0.881**	0.837

(b) Stanford background dataset

➢ PASCAL VOC 2012 segmentation

Ours achieved a mean IOU of 67.5% while the Base model is 64.2%.

➢ PASCAL VOC 2007 detection

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
RCNN [11]	73.4	77.0	63.4	45.4	44.6	75.1	78.1	79.8	40.5	73.7
fast RCNN [36]	74.5	78.3	69.2	53.2	36.6	77.3	78.2	82.0	40.7	72.7
faster RCNN [23]	69.1	78.3	68.9	55.7	49.8	77.6	79.7	85.0	51.0	76.1
HistNet-OD stage-1	68	80.3	74.1	55.7	53.3	83.6	80.2	85.1	53.7	74.2
HistNet-OD	67.6	80.3	74.1	55.6	53.2	83.4	80.2	85.1	53.6	74

	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	**mAP**
RCNN [11]	62.2	79.4	78.1	73.1	64.2	35.6	66.8	67.2	70.4	71.1	66.0
fast RCNN [36]	67.9	79.6	79.2	73.0	69.0	30.1	65.4	70.2	75.8	65.8	66.9
faster RCNN [23]	64.2	82.0	80.5	76.2	75.8	38.5	71.4	65.4	77.8	66.1	69.5
HistNet-OD stage-1	69.3	82.5	84.9	76.5	77.7	44.2	71.7	66.6	75.5	71.8	71.4
HistNet-OD	69.3	82.5	84.8	76.3	77.6	44.1	71.9	66.8	75.4	71.9	71.4

➢ Ablation study

➢ Learnable histogram v.s. fix-bin histogram v.s. unlocked histogram.

➢ Statistical context v.s. non-statistical context.

Methods	SIFTFlow		Stanford background		# extra parameters
	per-pixel	per-class	per-pixel	per-class	(SIFTFlow/Stanford)
FCN baseline	0.860	0.450	0.851	0.811	0
FCN-fix-hist	0.872	0.481	0.860	0.829	~ 190,000 / 36,000
FCN-free-all	0.870	0.489	0.862	0.824	~ 190,000 / 36,000
FCN-fc7-global	0.870	0.462	-	-	~ 960,000 / 23,000
FCN-score-global	0.873	0.480	0.863	0.825	~ 150,000 / 35,000
R-HistNet-SS	**0.880**	0.486	**0.872**	**0.845**	~ 380,000 / 72,000
HistNet-SS (ours)	0.879	**0.5**	0.871	0.837	~ 190,000 / 36,000

➢ Example results on semantic segmentation on SIFTFlow dataset

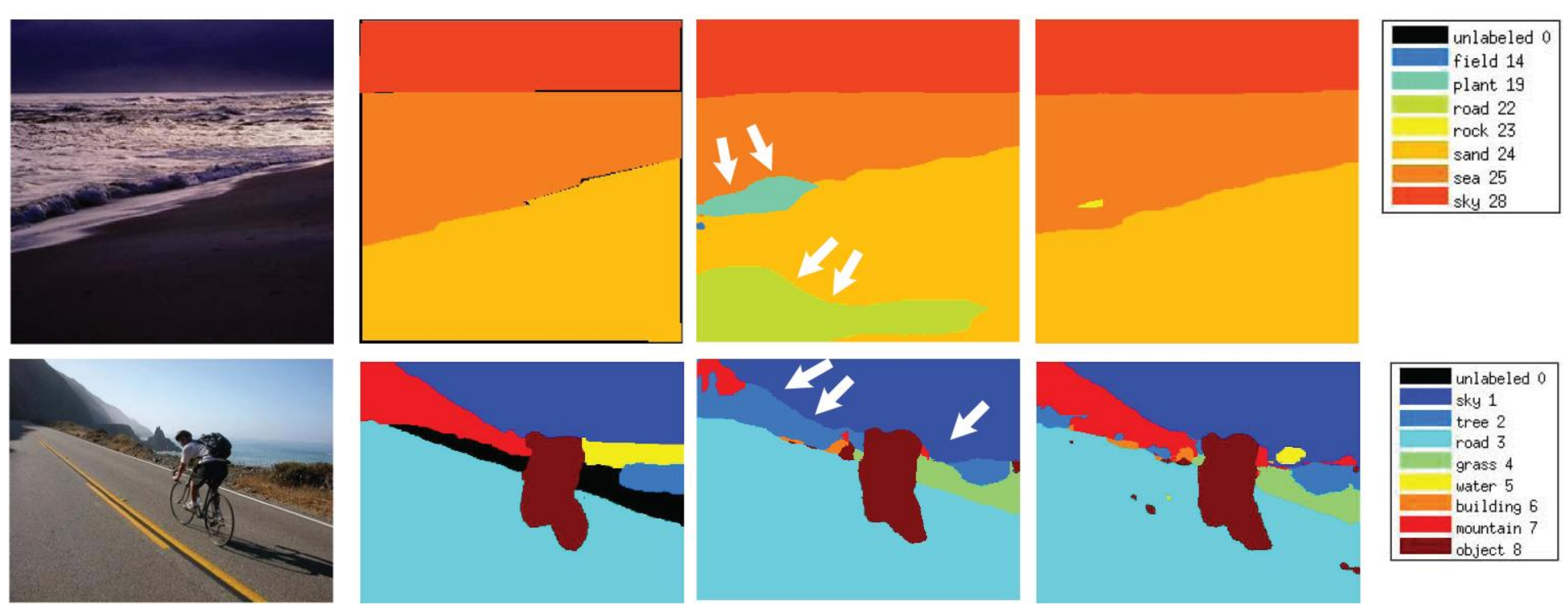


Image Label FCN HistNet-SS (ours)

1. Long, J., E.Shelhamer, Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proc. CVPR. (2014)
2. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In:Proc. ICLR. (2015)
3. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proc. NIPS. (2015)