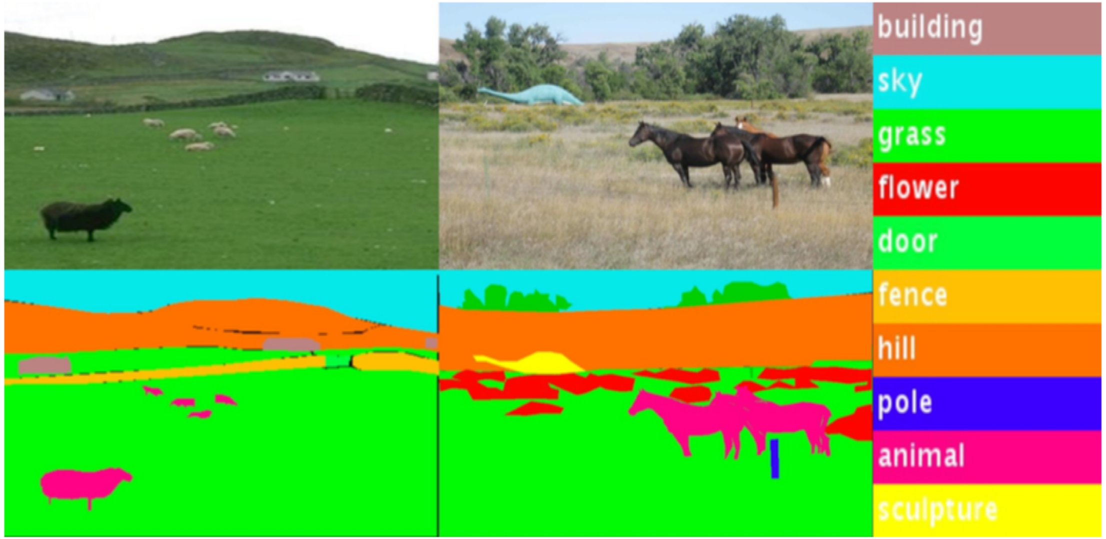
Context Encoding for Semantic Segmentation

CVPR 2018, Salt Lake City

Hang Zhang[1,2], Kristin Dana[1], Jianping Shi[3], Zhongyue Zhang[2], Xiaogang Wang[4], Ambrish Tyagi[2], and Amit Agrawal[2]

[1]Rutgers University, [2]Amzon Inc, [3]Sensetime, [4]CUHK

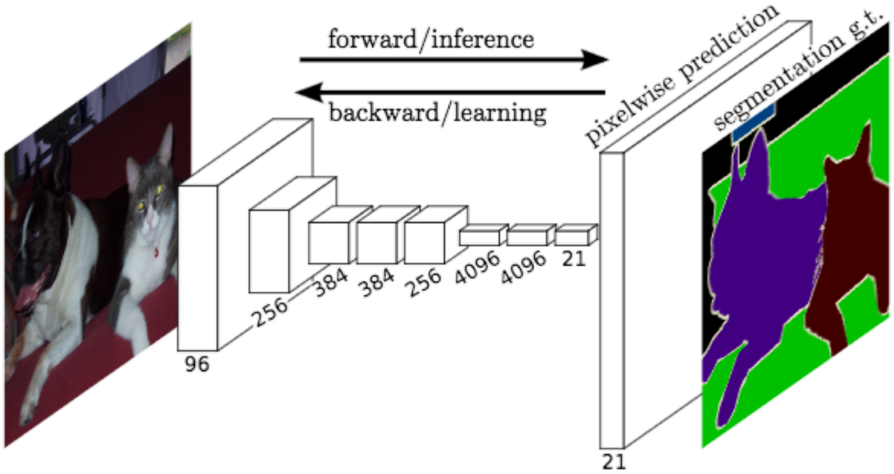
Semantic Segmentation



- Per-pixel predictions of object categories
- A comprehensive scene description (object category, location and shape)

Examples from ADE20K Dataset.

Fully Convolutional Network [1] (FCN)



- Meta algorithm for Semantic Segmentation
- Pre-trained CNN + Decoder
- Translation equivariant

Figure credit: Long et al.

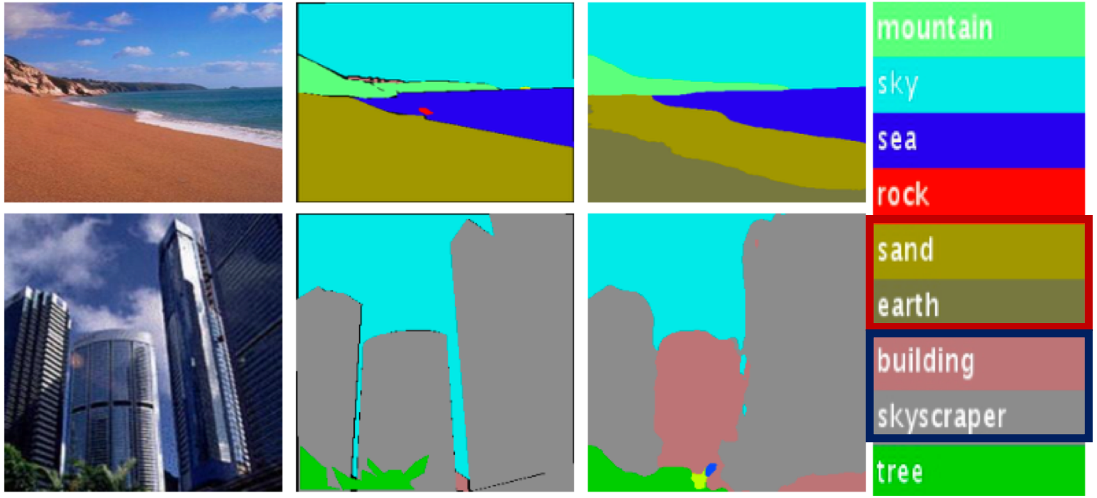
[1]Jonathan Long, Evan Shelhamer, & Trevor Darrell. “Fully Convolutional Networks for Semantic Segmentation”.CVPR2015

Difficulties in Predicting Categories and Shapes

- Work refining shapes/boundaries:
  - Dilated/Atrous Convolution [2,3]
  - CRF Post-processing [4]
  - Adding Lateral/Skip Connections [5]
  - Enlarging Spatial Resolution [6]
- Difficult to identifying categories

[2]Chen et al. "Rethinking Atrous Convolution for Semantic Image Segmentation". arXiv 2015
[3]Yu, Fisher, and Vladlen Koltun. "Multi-scale context aggregation by dilated convolutions." ICLR 2016
[4]Zheng, Shuai, et al. "Conditional random fields as recurrent neural networks. ICCV 2015
[5]Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation."
[6]Pohlen, Tobias, et al. "Full-resolution residual networks for semantic segmentation in street scenes." CVPR 2017

Challenges in Understanding Context



(a) Image (b) Ground Truth (c) FCN (baseline) (d) Legend

FCN results on ADE20K Dataset. (ResNet 50, stride 8)

Increasing Receptive Field?

Using pyramid representations

- PSPNet [7]
  Spatial Pyramid Pooling
- DeepLab-v3 [8]
  large rate Dilated/Atrous convolutions

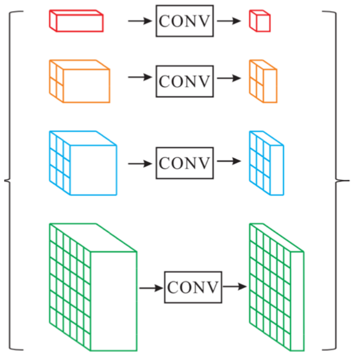


Figure credit: Zhao et al.

“Is *capturing contextual information the same as increasing the receptive-field size?*“

[7]Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia. “Pyramid Scene Parsing Network”. CVPR 2017.
[8]Chen et al. “Rethinking Atrous Convolution for Semantic Image Segmentation”. arXiv 2017.

Labeling an Image

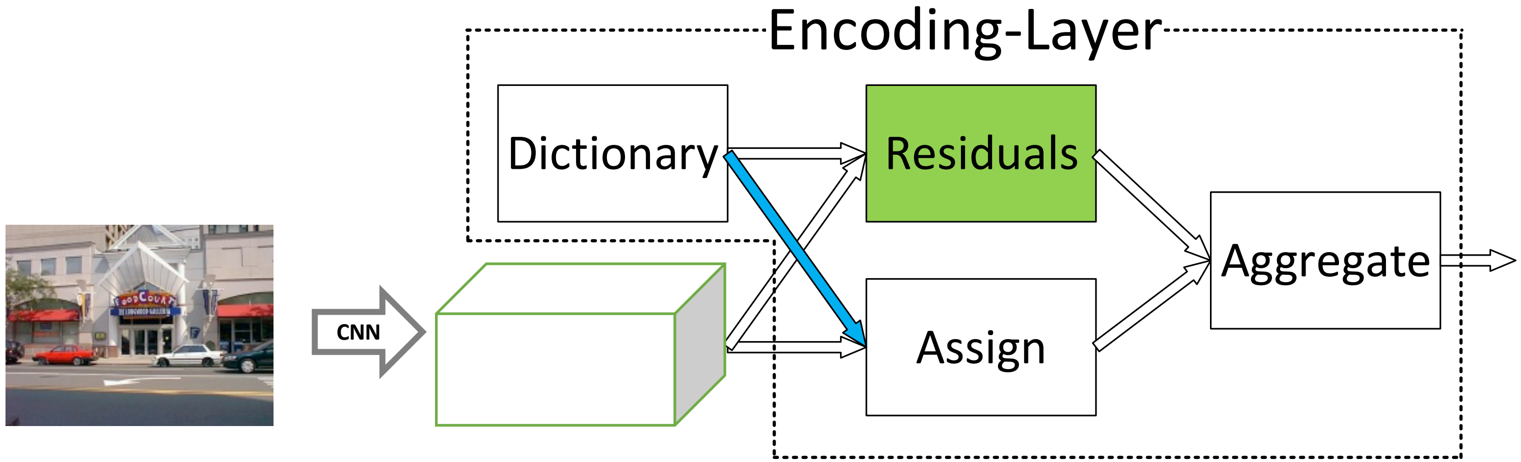
Consider labeling a new image for ADE20K dataset with **150 categories.**

Design a "Labeling Tool" for CNN

Examples from ADE20K Dataset.

- Scene Context
- Narrowing the list of probable categories

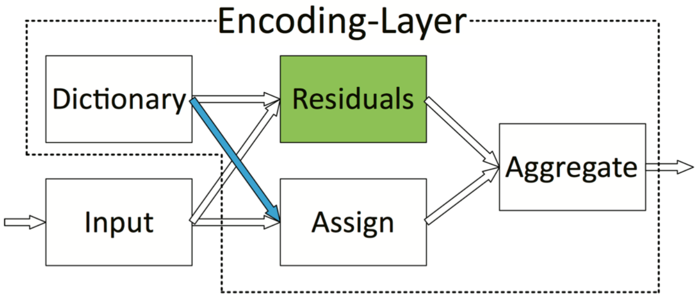
Capturing Contextual Info in Computer Vision



[9]Hang Zhang, Jia Xue, Kristin Dana. "Deep TEN: Texture Encoding Network". CVPR2017

Code available on GitHub

Context Encoding



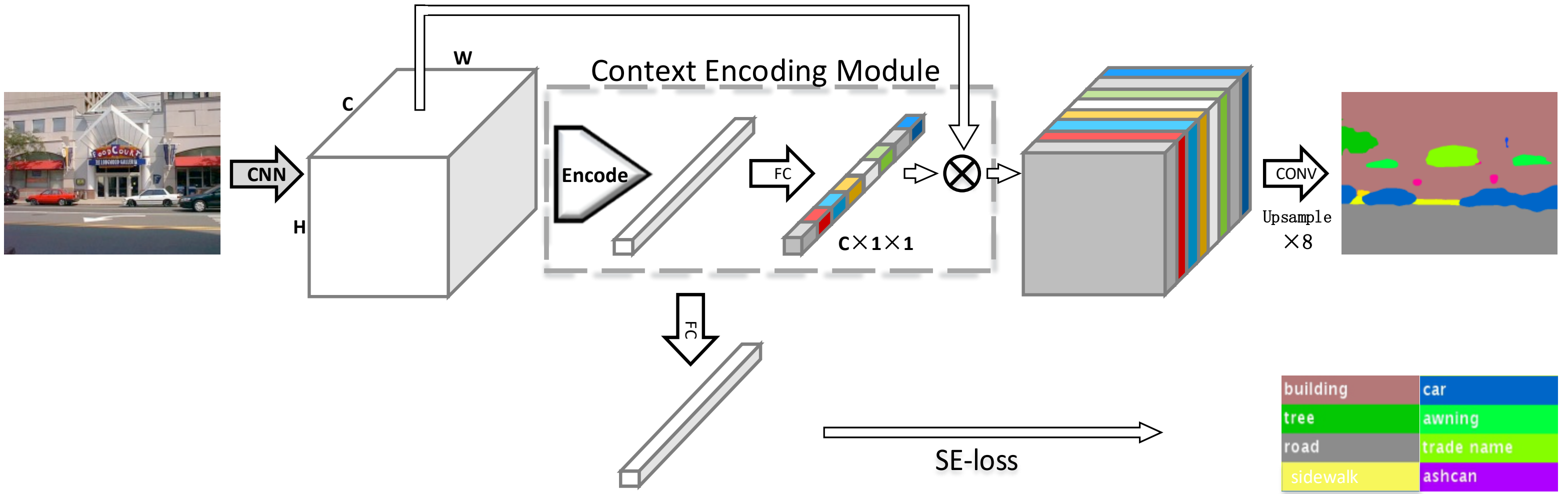
- Encoding Layer [9]
  - Considers $X \in \mathbb{R}^{C\times H\times W}$ as a set of C-dimensional features $X = \{x_1, \dots x_N\}$, where $N = H\times W$
  - Leans a codebook $D = \{d_1, \dots d_K\}$, smoothing factors $S = \{s_1, \dots s_K\}$
  - Outputs the residual encoder $e_k = \sum_{i=1}^{N} e_{ik}$:

$$e_{ik} = \frac{\exp(-s_k \|r_{ik}\|^2)}{\sum_{j=1}^{K} \exp(-s_j \|r_{ij}\|^2)} r_{ik}$$

Where the residuals are given by $r_{ik} = x_i - d_k$.

[9]Hang Zhang, Jia Xue, Kristin Dana. "Deep TEN: Texture Encoding Network". CVPR2017

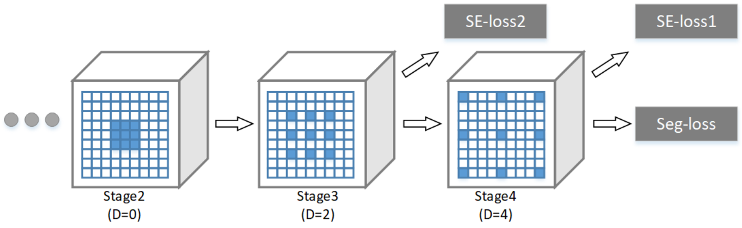
Context Encoding Network (EncNet)



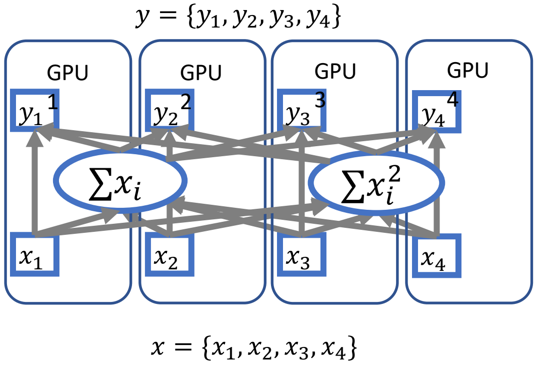
Notation: ***FC*** fully connected layer, ***Conv*** convolutional layer, ***Encode*** Encoding Layer[9], ⊗ channel-wise multiplication

[9] Hang Zhang, Jia Xue, Kristin Dana. “Deep TEN: Texture Encoding Network”. CVPR2017

Network Training of EncNet



- ResNet with Dilation Strategy (stride 8)
- Synchronize Cross-GPU Batch Normalization[10] (SyncBN)



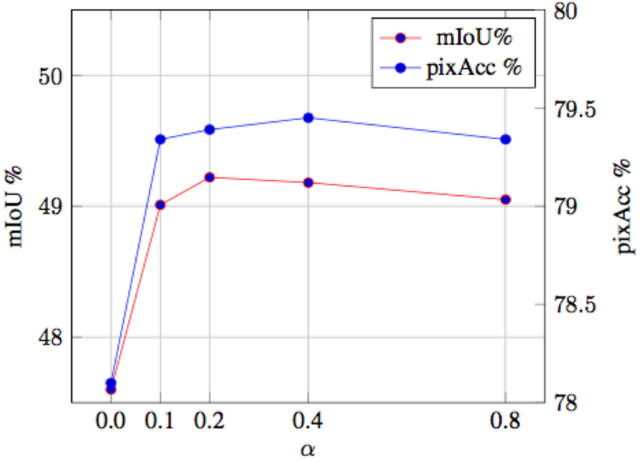
"Sync Once" Cross GPU BN implementation

[10] Ioffe and Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." ICML. 2015.

Ablation Study of EncNet on PASCAL Context

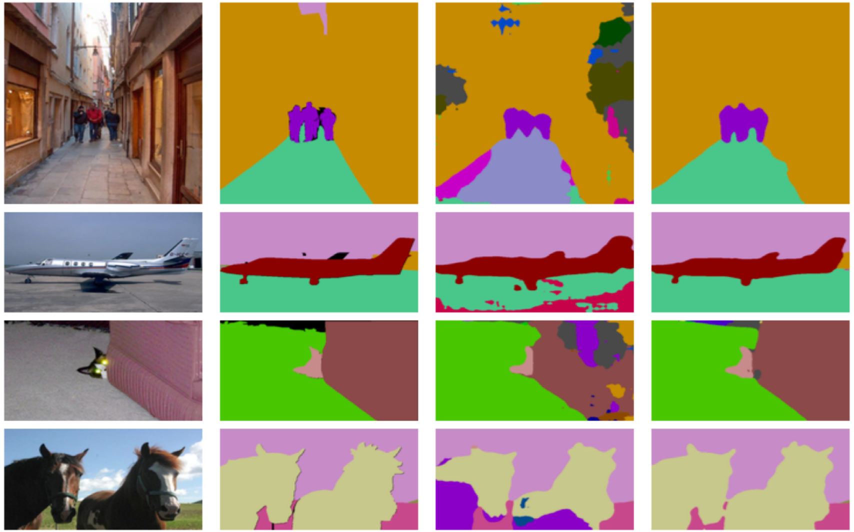
Method	BaseNet	Encoding	SE-loss	MS	pixAcc%	mIoU%
FCN	Res50				73.4	41.0
EncNet	Res50	✓			78.1	47.6
EncNet	Res50	✓	✓		79.4	49.2
EncNet	Res101	✓	✓		80.4	51.7
EncNet	Res101	✓	✓	✓	81.2	52.6

Semantic segmentation results on PASCAL-Context dataset. (mIoU on 59 classes w/o background)



mIoU and pixAcc as a function of SE-loss weight α.

EncNet Results on PASCAL Context



(a) Image (b) Ground Truth (c) FCN (d) EncNet (ours)

Method	BaseNet	mIoU%
FCN-8s [36]		37.8
CRF-RNN [58]		39.3
ParseNet [34]		40.4
BoxSup [9]		40.5
HO_CRF [2]		41.3
Piecewise [32]		43.3
VeryDeep [49]		44.5
DeepLab-v2 [5]	Res101-COCO	45.7
RefineNet [31]	Res152	47.3
EncNet (ours)	Res101	**51.7**

Segmentation results on PASCAL-Context dataset. (mIoU on 60 classes w/ background)

EncNet Results on PASCAL VOC 2012

Method	aero	bike	bird	boat	bottle	mIoU
FCN [37]	76.8	34.2	68.9	49.4	60.3	62.2
DeepLabv2 [4]	84.4	54.5	81.5	63.6	65.9	71.6
CRF-RNN [60]	87.5	39.0	79.7	64.2	68.3	72.0
DeconvNet [41]	89.9	39.3	79.7	63.9	68.2	72.5
GCRF [49]	85.2	43.9	83.3	65.2	68.3	73.2
DPN [36]	87.7	59.4	78.4	64.9	70.3	74.1
Piecewise [32]	90.6	37.6	80.0	67.8	74.4	75.3
ResNet38 [52]	**94.4**	**72.9**	94.9	68.8	78.4	82.5
PSPNet [59]	91.8	71.9	94.7	71.2	75.8	82.6
EncNet (ours)[3]	94.1	69.2	**96.3**	**76.7**	**86.2**	**82.9**

Results on PASCAL VOC 2012, showing per-class IoU on first 5 categories.

Method	aero	bike	bird	boat	bottle	mIoU
CRF-RNN [60]	90.4	55.3	88.7	68.4	69.8	74.7
Dilation8 [54]	91.7	39.6	87.8	63.1	71.8	75.3
DPN [36]	89.0	61.6	87.7	66.8	74.7	77.5
Piecewise [32]	94.1	40.7	84.1	67.8	75.9	78.0
DeepLabv2 [5]	92.6	60.4	91.6	63.4	76.3	79.7
RefineNet [31]	95.0	73.2	93.5	78.1	84.8	84.2
ResNet38 [52]	96.2	75.2	**95.4**	74.4	81.7	84.9
PSPNet [59]	95.8	72.7	95.0	78.9	84.4	85.4
DeepLabv3 [6]	**96.4**	76.6	92.7	77.8	**87.6**	85.7
EncNet (ours)[4]	95.3	**76.9**	94.2	**80.2**	85.2	**85.9**

Results on PASCAL VOC 2012 with COCO pre-training, showing per-class IoU on first 5 categories.

[11] http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?challengeid=11&compid=6

EncNet Results on ADE20K

Method	BaseNet	pixAcc%	mIoU%
FCN [36]		71.32	29.39
SegNet [3]		71.00	21.64
DilatedNet [52]		73.55	32.31
CascadeNet [59]		74.52	34.90
RefineNet [31]	Res152	-	40.7
PSPNet [57]	Res101	81.39	43.29
PSPNet [57]	Res269	**81.69**	**44.94**
FCN (baseline)	Res50	74.57	34.38
EncNet (ours)	Res50	79.73	41.11
EncNet (ours)	**Res101**	**81.69**	44.65

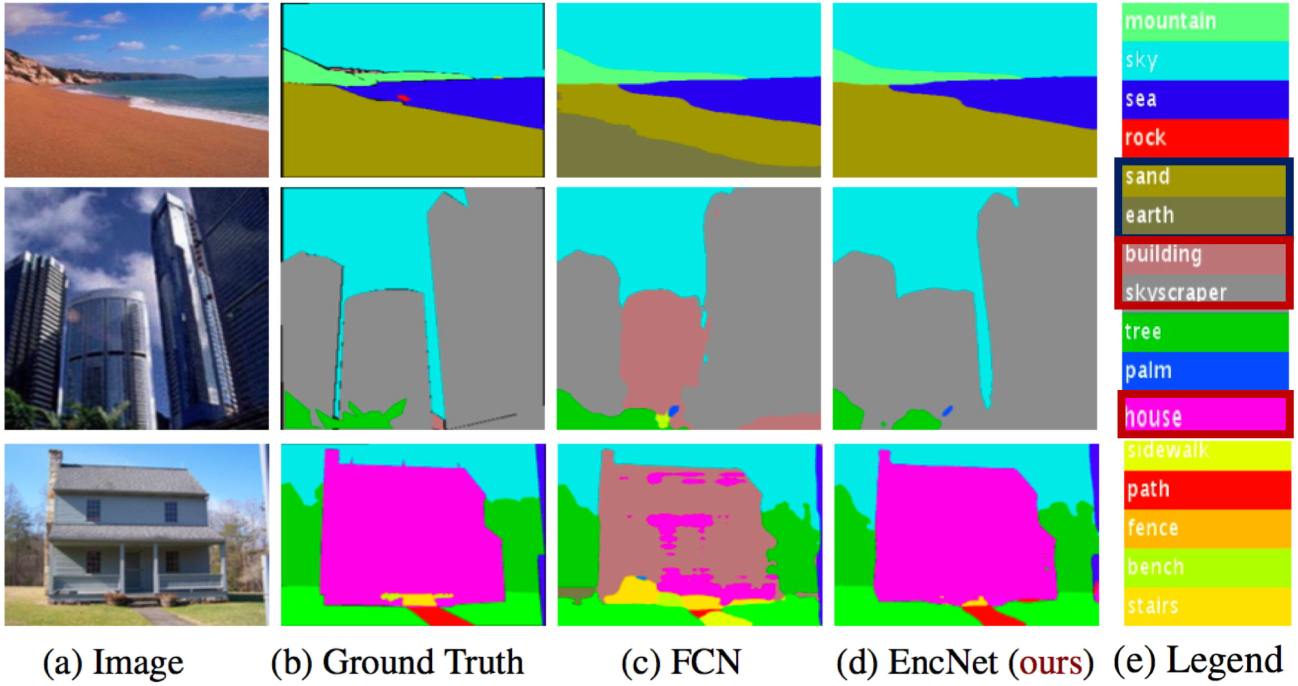
Results on ADE20K validation set.

rank	Team	Final Score
-	(EncNet-101, single model ours)	**0.5567**[6]
1	CASIA_IVA_JD	0.5547
2	WinterIsComing	0.5544
-	(PSPNet-269, single model) [57]	0.5538

Results on ADE20K test set, ranks in COCO-Place challenge 2017. Our single model surpass the winning entry of the COCO-Place challenge and PSPNet-269 (1st place in 2016).

[12] Leaderboard at http://sceneparsing.csail.mit.edu/

Visual Examples of EncNet in ADE20K



Conclusion

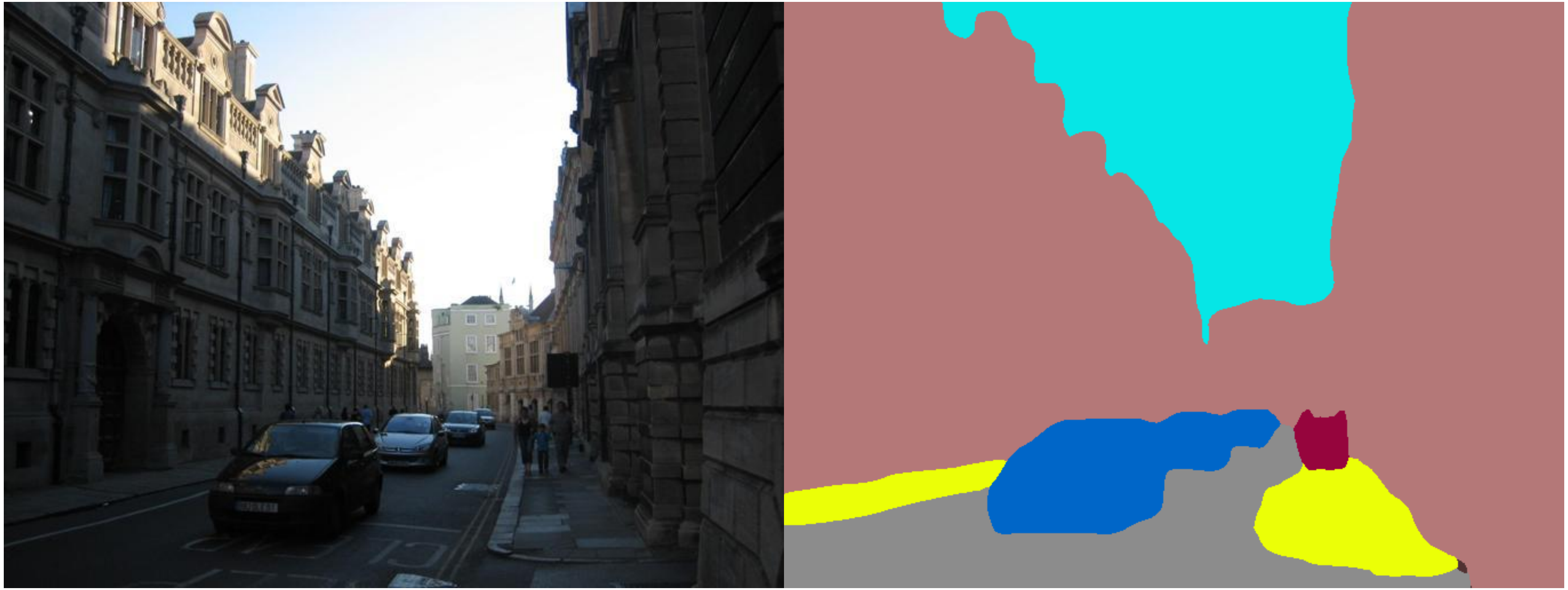
- Context Encoding Module with EncNet
  - straightforward, light-weight
  - compatible with FCN based approaches
- Superior performance on gold-standard benchmarks.
- The complete systems are publicly available (including SyncBN)
  - Source training/evaluation code and pretrained models

  https://github.com/zhanghang1989/PyTorch-Encoding
- Poster #A5

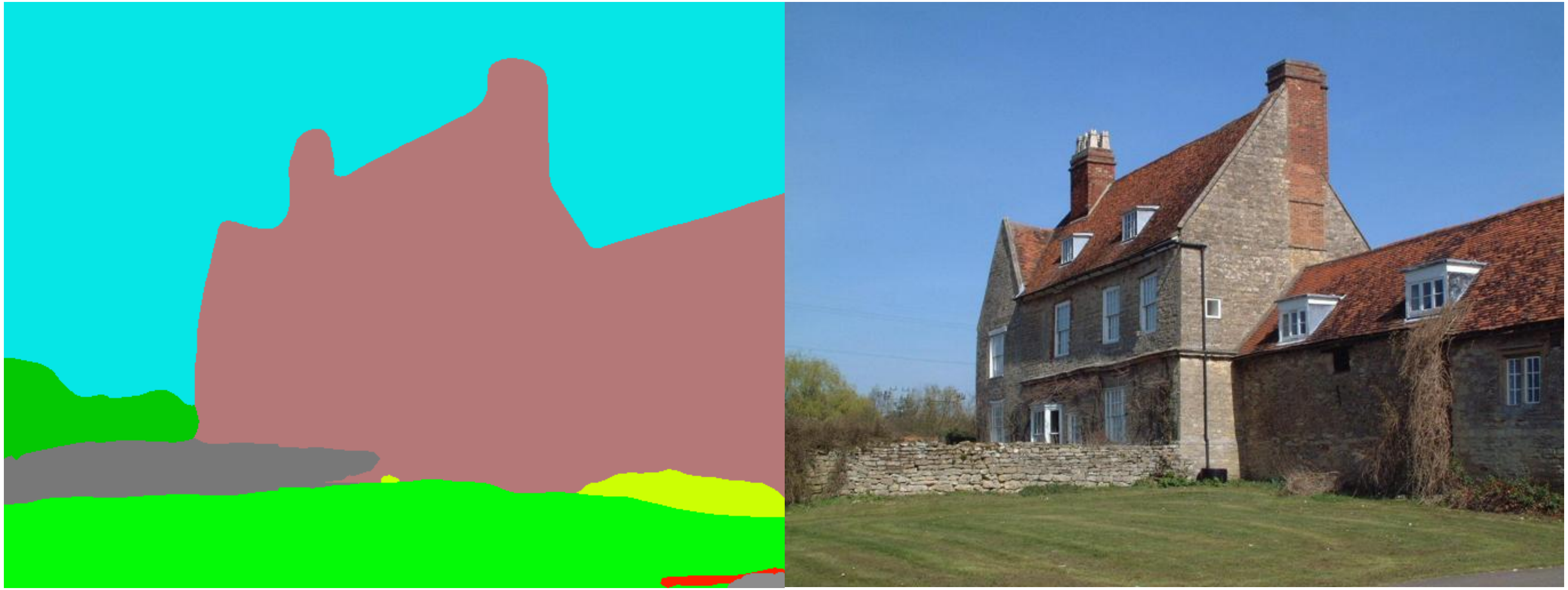
The authors would like to thank Sean Liu from Amazon Lab 126, Sheng Zha and Mu Li from Amazon AI for helpful discussions and comments. We thank Amazon Web Service (AWS) for providing free EC2 access.

More EncNet Examples on ADE20K Dataset

More EncNet Examples on ADE20K Dataset



More EncNet Examples on ADE20K Dataset



More EncNet Examples on ADE20K Dataset

Conclusion

- Context Encoding Module with EncNet
  - straightforward, light-weight
  - compatible with FCN based approaches
- Superior performance on gold-standard benchmarks.
- The complete systems are publicly available (including SyncBN)
  - Source training/evaluation code and pretrained models

  https://github.com/zhanghang1989/PyTorch-Encoding
- Poster #A5

The authors would like to thank Sean Liu from Amazon Lab 126, Sheng Zha and Mu Li from Amazon AI for helpful discussions and comments. We thank Amazon Web Service (AWS) for providing free EC2 access.

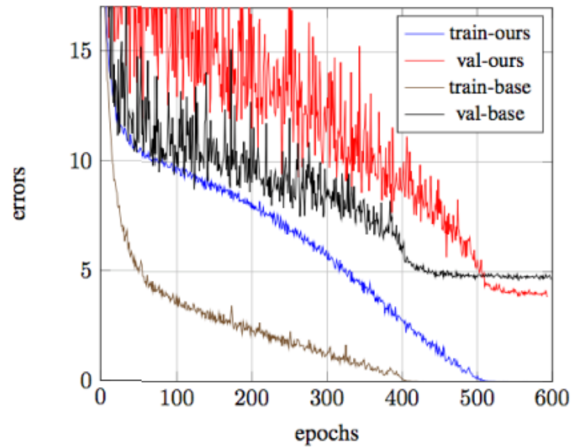
Prior Work in Featuremap Attention

- Spatial Attention: Spatial Transformer Network
- Channel-wise manipulation:
  - AdaIN or MSG-Net in style transfer
  - SE-Net
- Relations and Differences with SE-Net:
  - Semantic Encoding, an explicit representations for global context
  - EncNet directly highlight the class-dependent feature.

EncNet Experiments on CIFAR-10

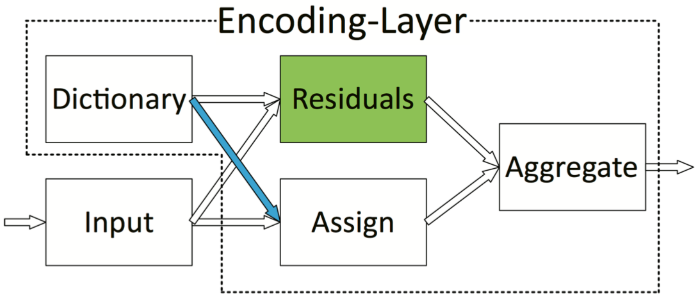
Method	Depth	Params	Error
ResNet (pre-act) [19]	1001	10.2M	4.62
Wide ResNet 28×10 [56]	28	36.5M	3.89
ResNeXt-29 16×64d [53]	29	68.1M	3.58
DenseNet-BC (k=40) [21]	190	25.6M	3.46
ResNet 64d (baseline)	14	2.7M	4.93
Se-ResNet 64d (baseline)	14	2.8M	4.65
EncNet 16k64d (ours)	14	**3.5M**	3.96
EncNet 32k128d (ours)	**14**	16.8M	**3.45**

Comparison of model depth, number of parameters, test errors (%) on CIFAR-10.



Train and validation curves of EncNet-32k64d and the baseline Se-ResNet-64d on CIFAR-10 dataset.

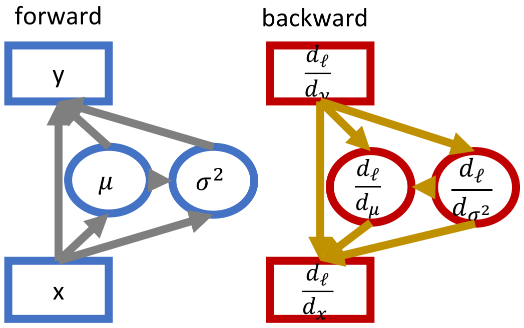
Context Encoding



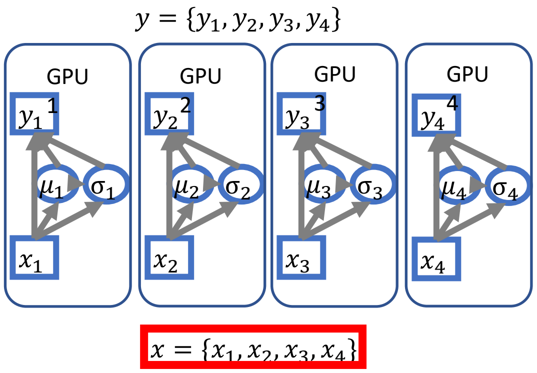
- Encoding Layer [9]
  - Outputs the residual encoder as encoded semantics $\mathrm{e} = \sum_{k=1}^{K} \phi(e_k)$
- Featuremap Attention
  - *FC* on encoded semantics, outputs scaling factors $\gamma = \delta(We)$, where W is the layer weight and δ is sigmoid function.
  - Channel-wise multiplication $Y = X \otimes \gamma$

[9]Hang Zhang, Jia Xue, Kristin Dana. "Deep TEN: Texture Encoding Network". CVPR2017

Standard BN and Data Parallelism



Batch Normalization [5] in training mode.

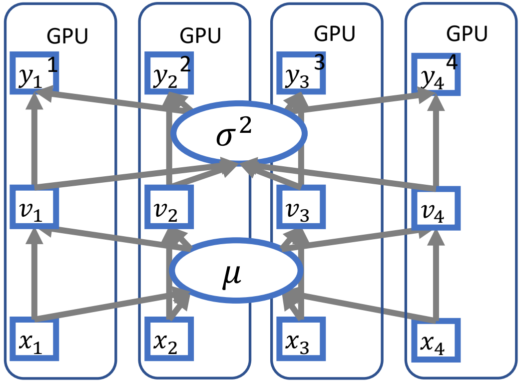


Standard BN with data parallel implementation.

[5]Ioffe and Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." ICML. 2015.

Cross-GPU Batch Norm ("Sync twice")

- $\mu = \frac{\sum x_i}{N}$
- $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N} + \epsilon}$



"Sync twice" Implementation[6,7]

[6]Peng, Chao, et al. "MegDet: A Large Mini-Batch Object Detector." CVPR2018
[7]Liu, Shu, et al. "Path Aggregation Network for Instance Segmentation." CVPR2018

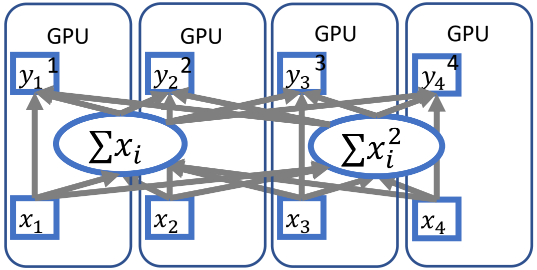
Cross-GPU Batch Norm ("Sync once")

- $\mu = \frac{\sum x_i}{N}$
- $\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{N} + \epsilon}$

$$= \sqrt{\frac{\sum x_i^2}{N} - \mu^2 + \epsilon}$$

$$= \boxed{\sqrt{\frac{\sum x_i^2}{N} - \frac{(\sum x_i)^2}{N^2} + \epsilon}}$$

$y = \{y_1, y_2, y_3, y_4\}$



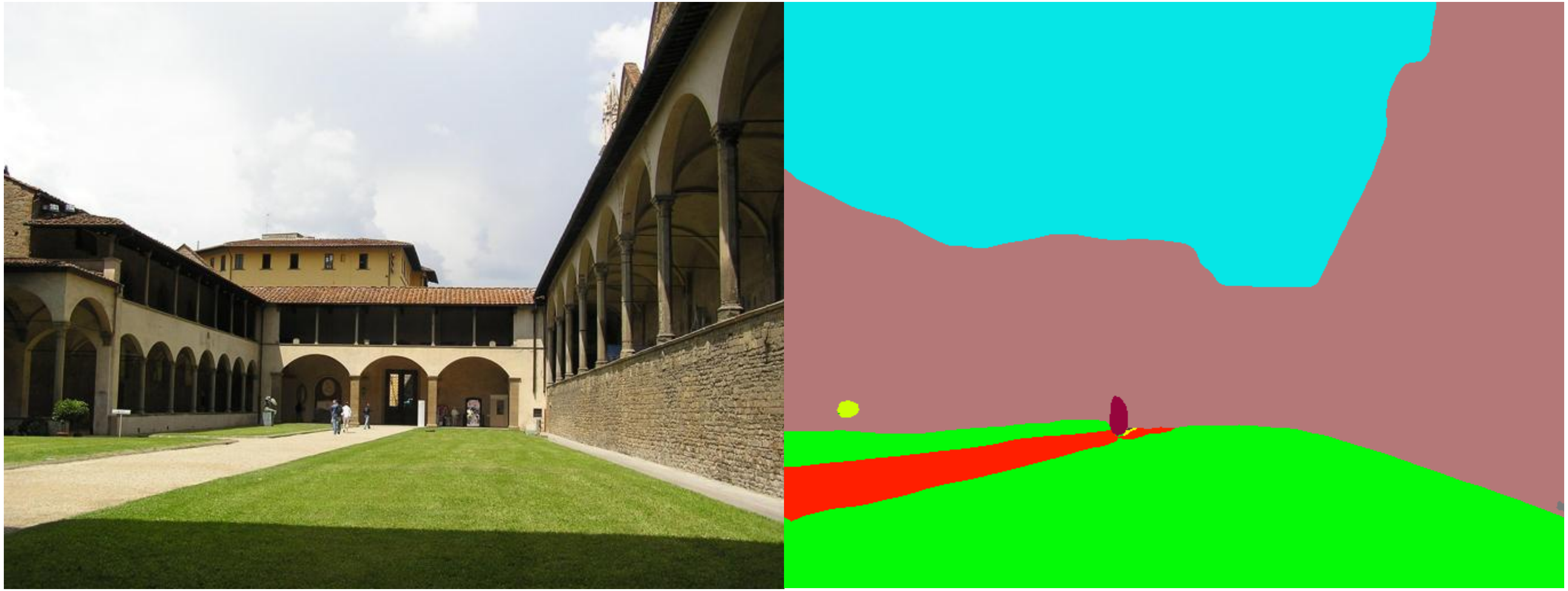
$x = \{x_1, x_2, x_3, x_4\}$

Our "Sync Once" implementation

[6]Peng, Chao, et al. "MegDet: A Large Mini-Batch Object Detector." CVPR2018
[7]Liu, Shu, et al. "Path Aggregation Network for Instance Segmentation." CVPR2018

More EncNet Examples on ADE20K Dataset

More EncNet Examples on ADE20K Dataset



Failure Examples of EncNet

Highlights

- Introduce a novel CNN architecture:
  - Context Encoding Network, EncNet (“Ink-Net”)
- State-of-the-art performance:
  - 85.9% mIoU on PASCAL VOC 2012, 51.7% mIoU on PASCAL Context
  - On ADE20K dataset, out single model surpass the winning entry of COCO Places challenge 2017 (achieving a final score of 0.5567)
- The complete system are publicly available.
  - Source code and pretrained models