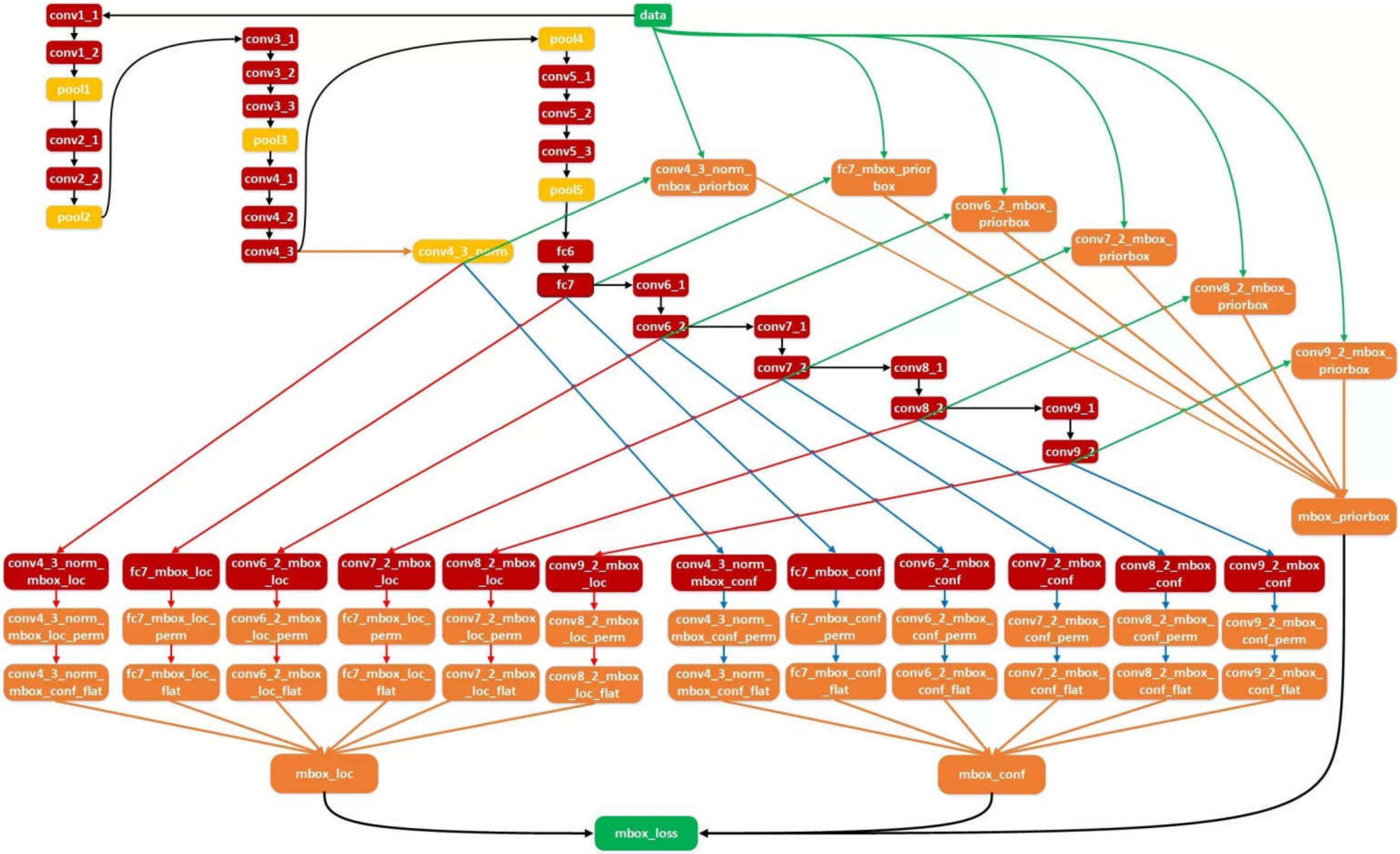
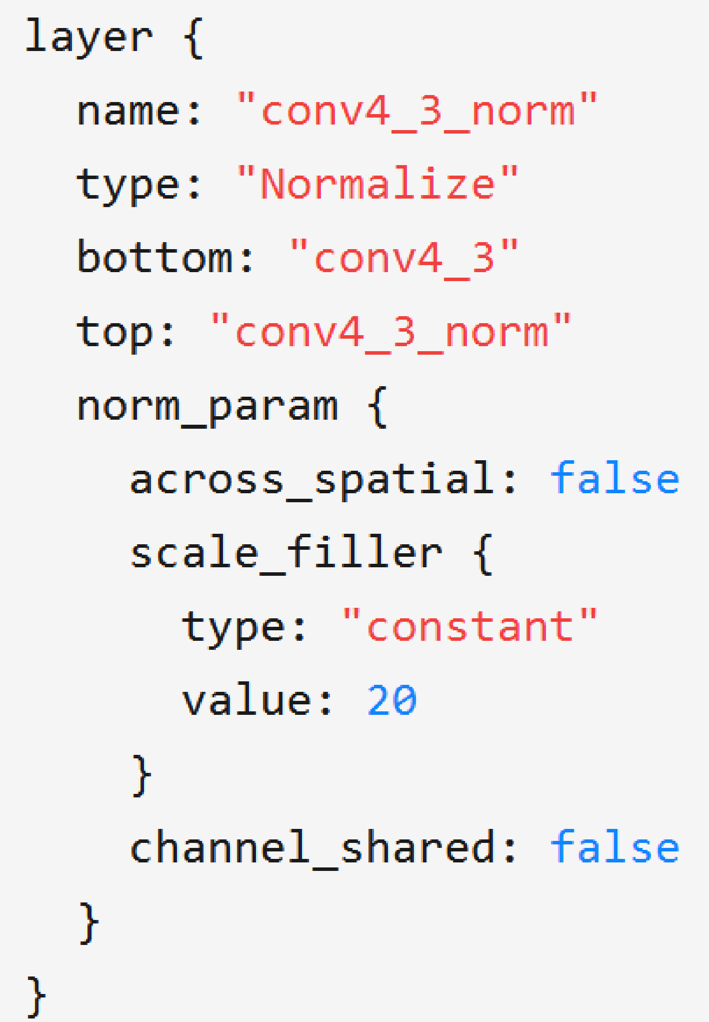
- SSD conv4_3_norm
- ParseNet
- SENet



conv1_1
conv1_2
pool1
conv2_1
conv2_2
pool2
conv3_1
conv3_2
conv3_3
pool3
conv4_1
conv4_2
conv4_3
conv4_3_norm
pool4
conv5_1
conv5_2
conv5_3
pool5
fc6
fc7
data
conv6_1
conv6_2
conv7_1
conv7_2
conv8_1
conv8_2
conv9_1
conv9_2
conv4_3_norm_mbox_priorbox
fc7_mbox_prior box
conv6_2_mbox_ priorbox
conv7_2_mbox_ priorbox
conv8_2_mbox_ priorbox
conv9_2_mbox_ priorbox
mbox_priorbox
conv4_3_norm_ mbox_loc
fc7_mbox_loc
conv6_2_mbox _loc
conv7_2_mbox _loc
conv8_2_mbox _loc
conv9_2_mbox _loc
conv4_3_norm_ mbox_conf
fc7_mbox_conf
conv6_2_mbox _conf
conv7_2_mbox _conf
conv8_2_mbox _conf
conv9_2_mbox _conf
conv4_3_norm_ mbox_loc_perm
fc7_mbox_loc_ perm
conv6_2_mbox _loc_perm
fc7_mbox_loc_ perm
conv7_2_mbox _loc_perm
conv8_2_mbox _loc_perm
conv4_3_norm_ mbox_conf_perm
fc7_mbox_conf _perm
conv6_2_mbox_ conf_perm
conv7_2_mbox_ conf_perm
conv8_2_mbox_ conf_perm
conv9_2_mbox_ conf_perm
conv4_3_norm_ mbox_conf_flat
fc7_mbox_loc_ flat
conv6_2_mbox _loc_flat
fc7_mbox_loc_ flat
conv7_2_mbox _loc_flat
conv8_2_mbox _loc_flat
conv4_3_norm_ mbox_conf_flat
fc7_mbox_conf _flat
conv6_2_mbox_ conf_flat
conv7_2_mbox_ conf_flat
conv8_2_mbox_ conf_flat
conv9_2_mbox_ conf_flat
mbox_loc
mbox_conf
mbox_loss



```
layer {
  name: "conv4_3_norm"
  type: "Normalize"
  bottom: "conv4_3"
  top: "conv4_3_norm"
  norm_param {
    across_spatial: false
    scale_filler {
      type: "constant"
      value: 20
    }
    channel_shared: false
  }
}
```

Norm的原因（作者回答）
That was discovered in my other paper (ParseNet) that conv4_3 has different scale from other layers. That is why I add L2 normalization for conv4_3 only.

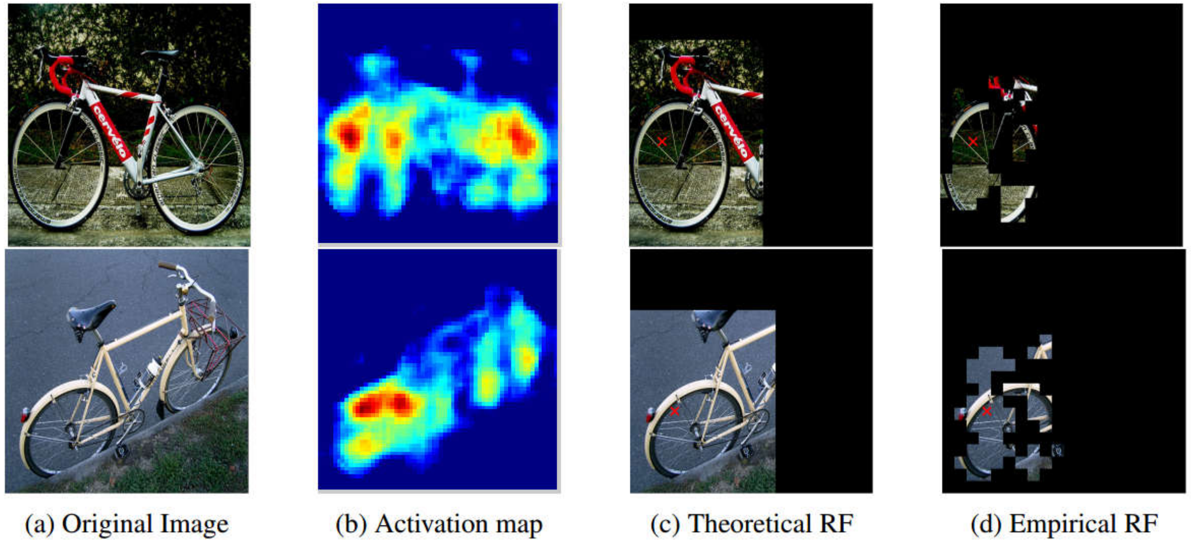
Value 的作用：scale

across_spatial含义：决定了normalization的范围
为true：整个feature map（c*h*w）。
为**false**，每个feaure vector（c*1*1）分别进行。

PARSENET: LOOKING WIDER TO SEE BETTER ICLR 2016

semantic segmetation

RF的实际情况低于理论计算值，
原因没说，激活函数导致？



(a) Original Image (b) Activation map (c) Theoretical RF (d) Empirical RF

Figure 2: **Receptive field (RF) size for last layer.** (a) original image; (b) activation map on bicycle from a channel of the last layer of a network; (c) theoretical receptive field of the maximum activation (marked by red cross) is defined by the network structure; (d) empirical receptive field affecting the activation. Clearly empirical receptive field is not large enough to capture the global context.

Adding global context for semantic segmentation is useful

Difficult in combining features:

the scale of features from different layers may be **quite different**, making it difficult to directly combine them for prediction.

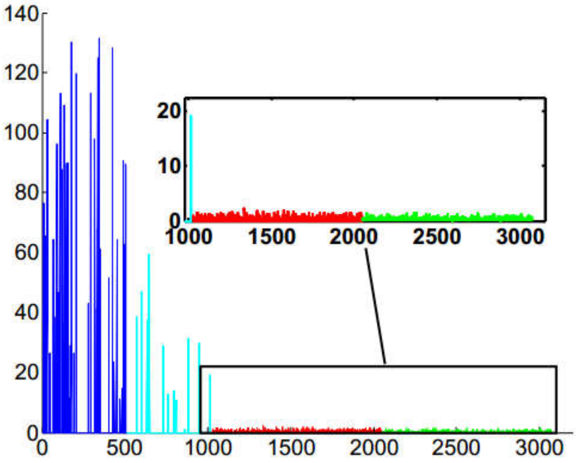


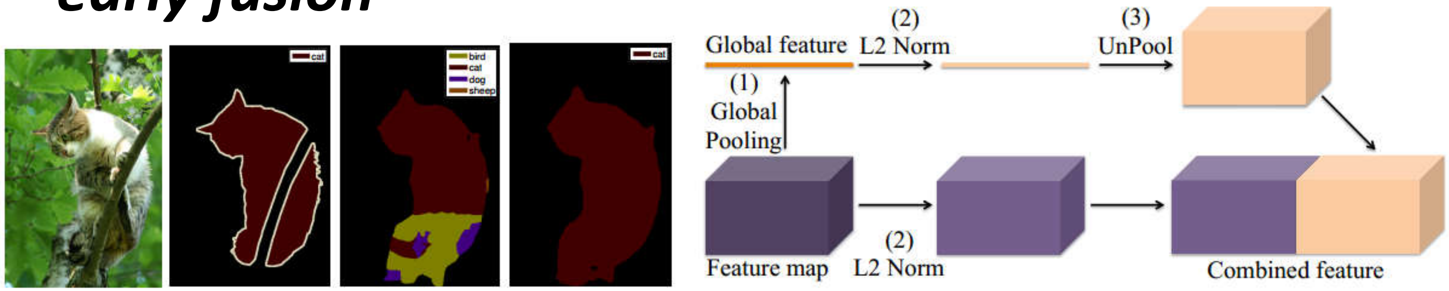
Figure 3: Features from 4 different layers have activations that are of drastically different scales. Each color corresponds to a different layers' feature. While *blue* and *cyan* are on a comparable scale, *red* and *green* features are of a scale 2 orders of magnitude less.

Solution:

We find that ***L_2* normalizing** features for each layer and combining them using **a scaling factor learned** through backpropagation works well to address this potential difficulty

two general standard paradigms

- ***early fusion***



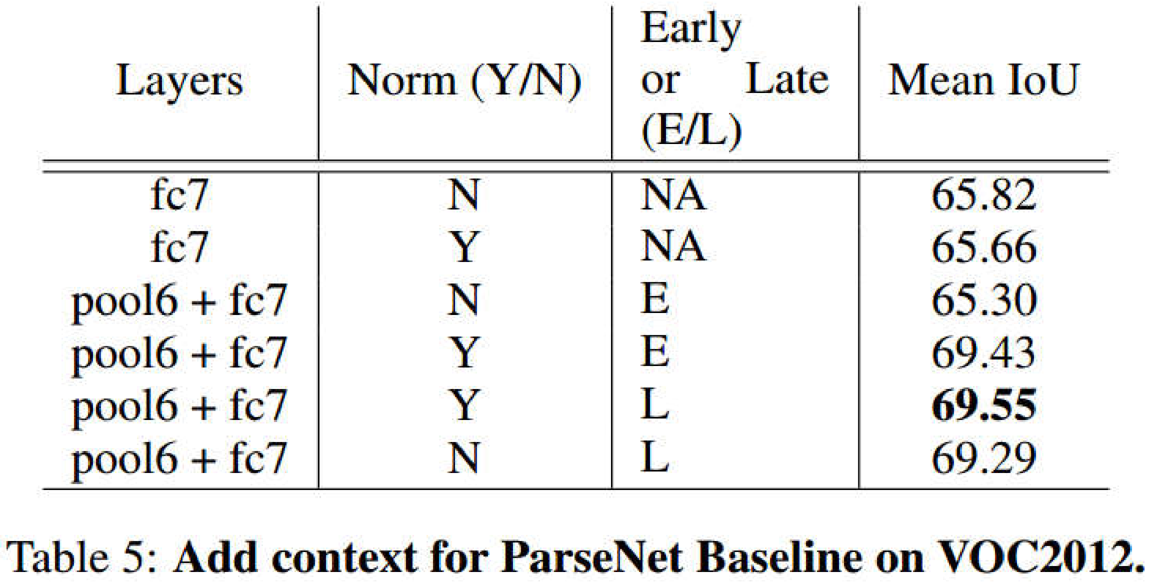
(a) Image (b) Truth (c) FCN (d) ParseNet (e) ParseNet contexture module overview.

Figure 1: ParseNet uses extra global context to clarify local confusion and smooth segmentation.

- ***late fusion***

each feature is used to learn its own classifier, followed by merging the two predictions into a single classification core

- experiments show that both method works more or less the same if we normalize the feature properly for early fusion case



Layers	Norm (Y/N)	Early or Late (E/L)	Mean IoU
fc7	N	NA	65.82
fc7	Y	NA	65.66
pool6 + fc7	N	E	65.30
pool6 + fc7	Y	E	69.43
pool6 + fc7	Y	L	**69.55**
pool6 + fc7	N	L	69.29

Table 5: **Add context for ParseNet Baseline on VOC2012.**

Difficult in normalizing features

- Simply normalizing each input of a layer changes the scale of the layer and will slow down the learning if we do not scale it accordingly.
- For example, we tried to normalize a feature s.t. *L*2-norm is 1, yet we can hardly train the network **because the features become very small(?)**. However, if we normalize it to e.g. 10 or 20, the network begins to learn well

Solution

- Introduce a *scaling* parameter *γi*, for each channel, which scales the normalized value by *yi* = *γix^i*.

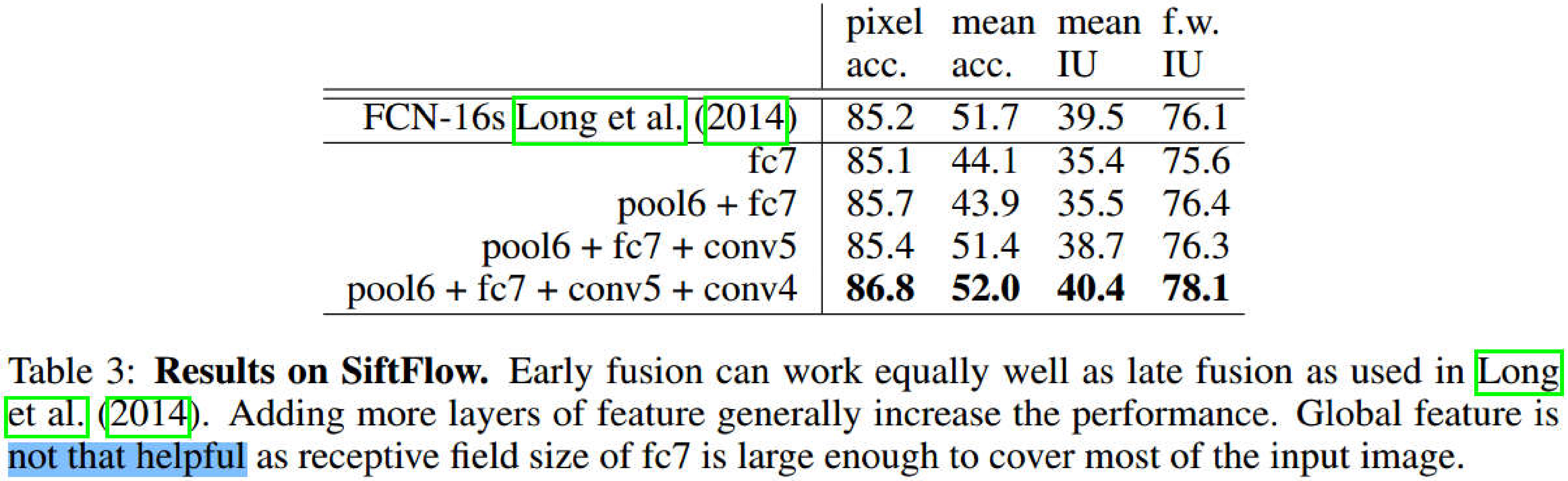
  The number of extra parameters is equal to total number of channels.(每个channel一个γ，分别normalize)

$$\frac{\partial \ell}{\partial \hat{\mathbf{x}}} = \frac{\partial \ell}{\partial \mathbf{y}} \cdot \gamma \qquad \frac{\partial \ell}{\partial \mathbf{x}} = \frac{\partial \ell}{\partial \hat{\mathbf{x}}}\left(\frac{\mathbf{I}}{||\mathbf{x}||_2} - \frac{\mathbf{x}\mathbf{x}^T}{||\mathbf{x}||_2^3}\right) \qquad \frac{\partial \ell}{\partial \gamma_i} = \sum_{y_i} \frac{\partial \ell}{\partial y_i}\hat{x}_i$$

Experiment

FCN-32s use **pool6** as the global context feature

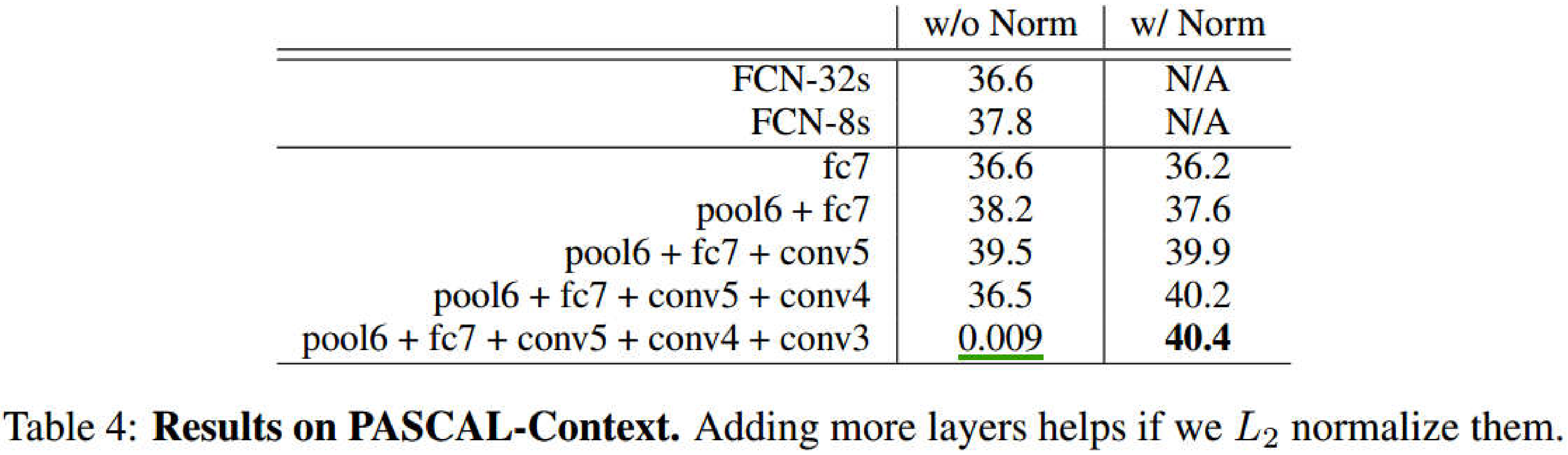
This is perhaps due to the small image size (256 × 256)



	pixel acc.	mean acc.	mean IU	f.w. IU
FCN-16s Long et al. (2014)	85.2	51.7	39.5	76.1
fc7	85.1	44.1	35.4	75.6
pool6 + fc7	85.7	43.9	35.5	76.4
pool6 + fc7 + conv5	85.4	51.4	38.7	76.3
pool6 + fc7 + conv5 + conv4	**86.8**	**52.0**	**40.4**	**78.1**

Table 3: **Results on SiftFlow.** Early fusion can work equally well as late fusion as used in Long et al. (2014). Adding more layers of feature generally increase the performance. Global feature is not that helpful as receptive field size of fc7 is large enough to cover most of the input image.

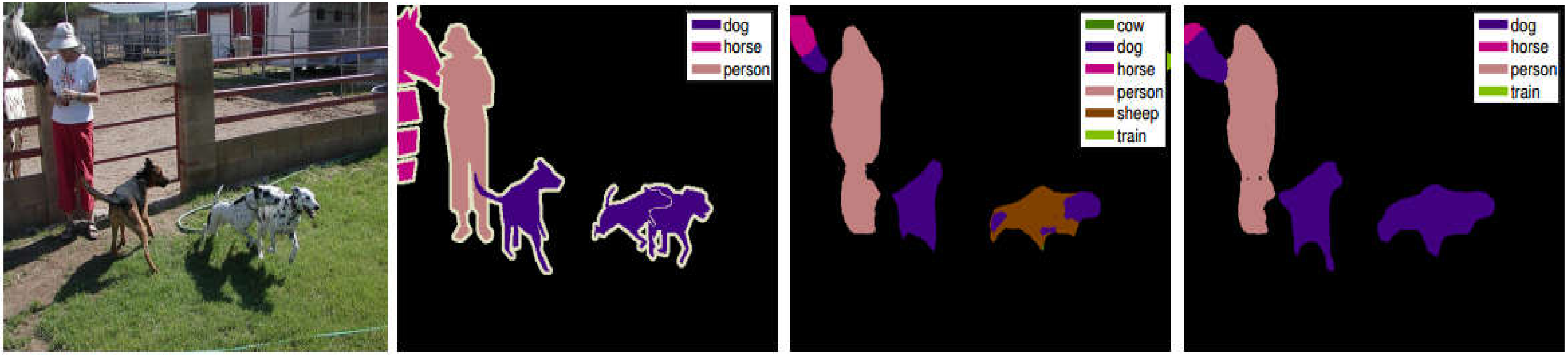
- if we naively keep adding conv4, it starts decreasing the performance a bit; and if we add conv3, the network collapses.



	w/o Norm	w/ Norm
FCN-32s	36.6	N/A
FCN-8s	37.8	N/A
fc7	36.6	36.2
pool6 + fc7	38.2	37.6
pool6 + fc7 + conv5	39.5	39.9
pool6 + fc7 + conv5 + conv4	36.5	40.2
pool6 + fc7 + conv5 + conv4 + conv3	0.009	**40.4**

Table 4: **Results on PASCAL-Context.** Adding more layers helps if we L_2 normalize them.

- Sometimes context also brings confusion for prediction
- the global context feature definitely captured the spotty dog information that it used to help discriminate sheep from dog. However, it also added bias to classify the spotty horse as a dog.



(a) Original Image (b) Ground truth (c) ParseNet Baseline (d) ParseNet

Figure 5: Global context confuse local patch predictions.

summary

- 1.感受野实际值<理论值，需要global context
- 2.不同特征层由于feature数值差异大导致fusion困难：融合前按channel norm
- 3.l2 norm为1训练慢：每个channel分别scale，scale数值学习得到
- 4.end-to-end，extra开销小
- 5.深层feature rf足够时，融入gobal context没啥效果(eg:image size is small)
- 6. Sometimes context also brings confusion for prediction

最后一届 ImageNet 2017 竞赛 Image Classification 任务的冠军

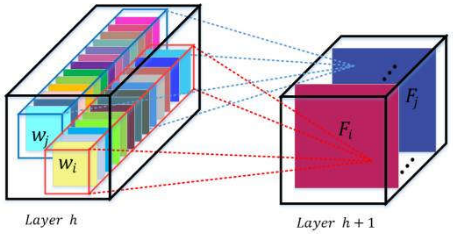
Squeeze-and-Excitation (SE) Networks

- If a network can be enhanced from the aspect of **channel relationship**?
- **Motivation:**
  - Explicitly model channel-interdependencies within modules
  - Feature recalibration
    - Selectively enhance useful features and suppress less useful ones

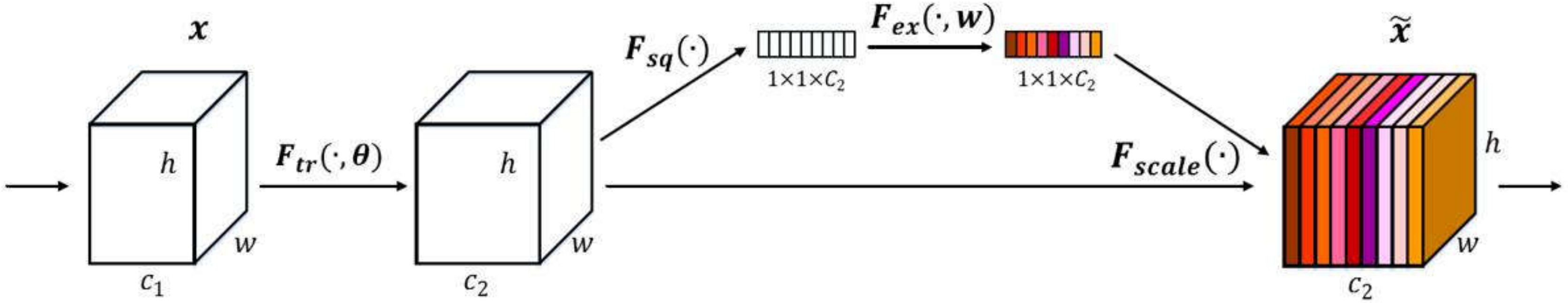
Convolution

A convolutional filer is expected to be an informative combination

- Fusing **channel-wise** and **spatial** information
- Within local receptive fields



Squeeze-and-Excitation Module



$\boldsymbol{x}$
h
w
c_1
$\boldsymbol{F}_{tr}(\cdot,\boldsymbol{\theta})$
h
w
c_2
$\boldsymbol{F}_{sq}(\cdot)$
$1\times1\times C_2$
$\boldsymbol{F}_{ex}(\cdot,\boldsymbol{w})$
$1\times1\times C_2$
$\boldsymbol{F}_{scale}(\cdot)$
$\widetilde{\boldsymbol{x}}$
h
w
c_2

Squeeze

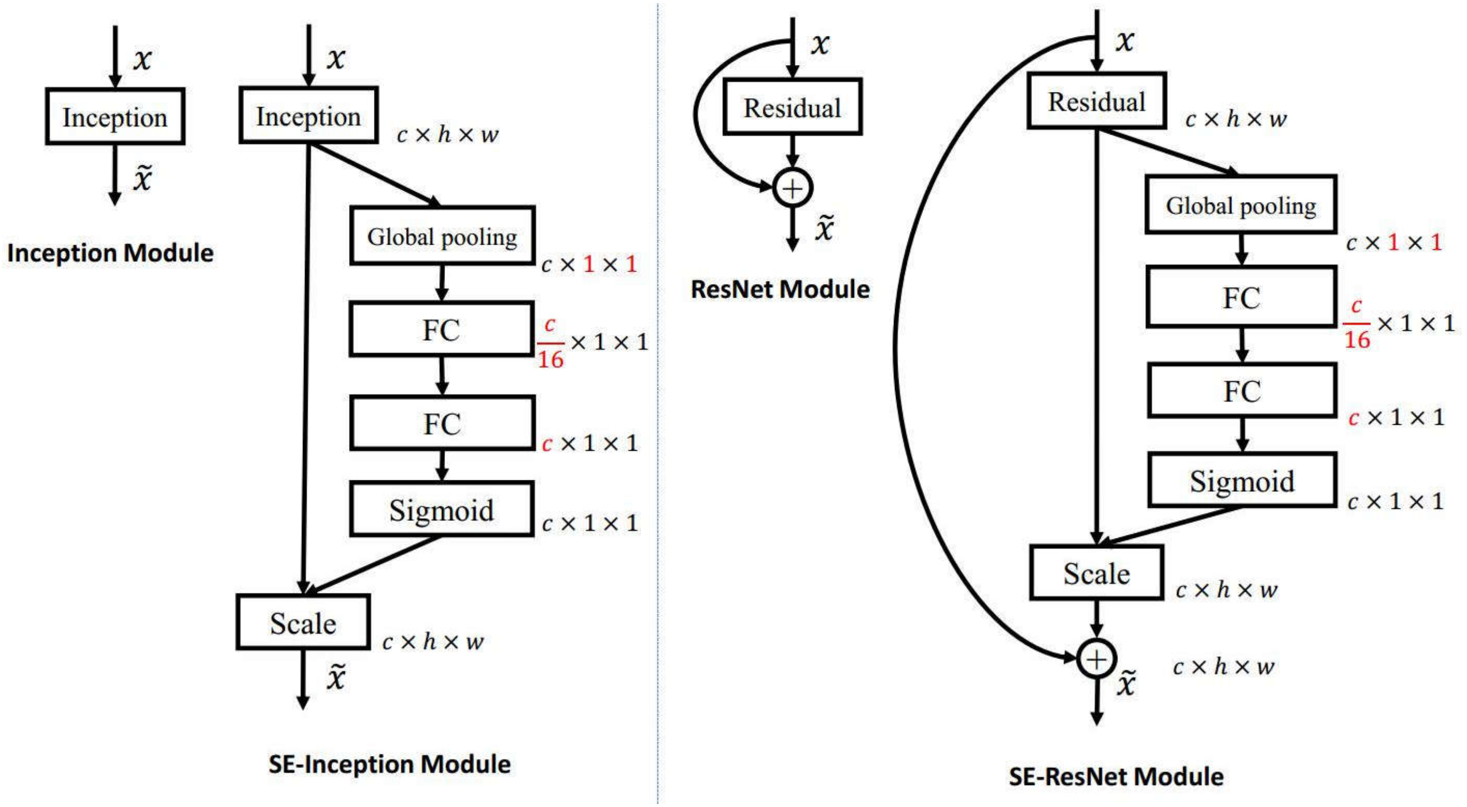
- Shrinking feature maps $\in \mathbb{R}^{w\times h\times c_2}$ through spatial dimensions ($w \times h$)
- Global distribution of channel-wise responses

Excitation

- Learning $W \in \mathbb{R}^{c_2\times c_2}$ to explicitly model channel-association
- Gating mechanism to produce channel-wise weights

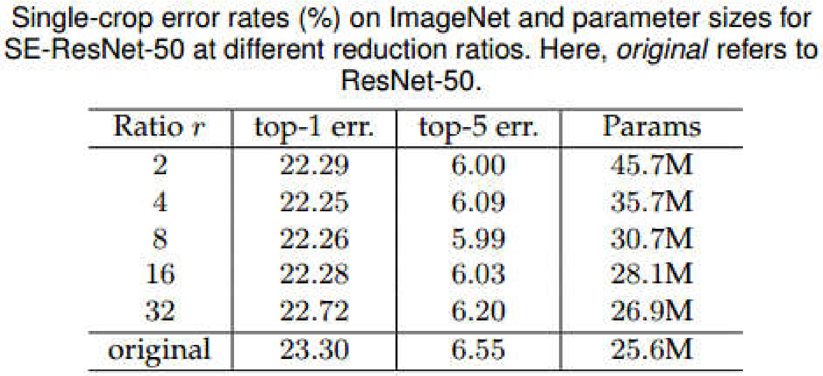
Scale

- Reweighting the feature maps $\in \mathbb{R}^{w\times h\times c_2}$



然后通过一个 Sigmoid 的门获得 0~1 之间归一化的权重，最后通过一个 Scale 的操作来将归一化后的权重加权到每个通道的特征上。

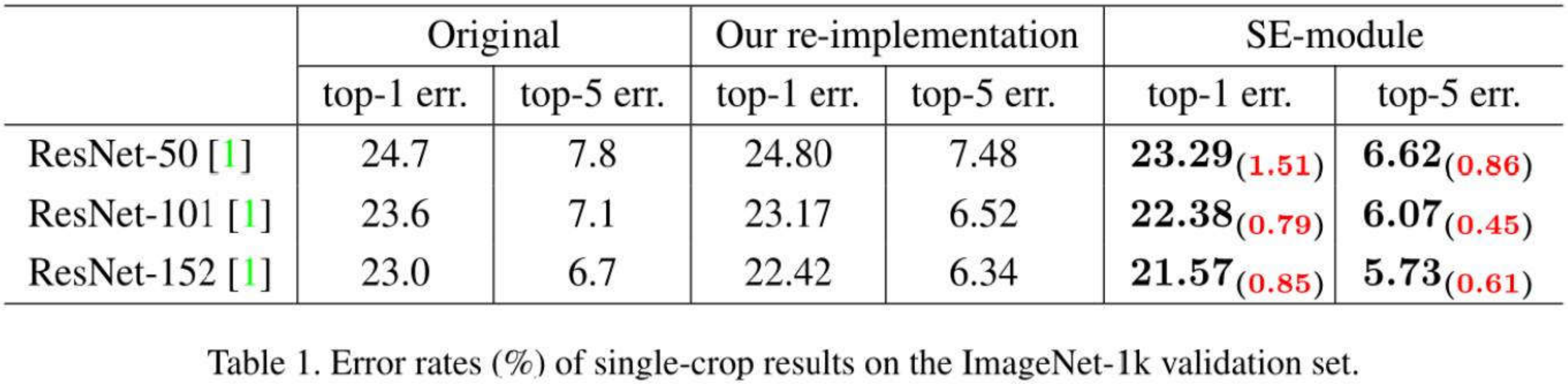
1）具有更多的非线性，可以更好地拟合通道间复杂的相关性；
2）极大地减少了参数量和计算量。



Single-crop error rates (%) on ImageNet and parameter sizes for SE-ResNet-50 at different reduction ratios. Here, *original* refers to ResNet-50.

Ratio r	top-1 err.	top-5 err.	Params
2	22.29	6.00	45.7M
4	22.25	6.09	35.7M
8	22.26	5.99	30.7M
16	22.28	6.03	28.1M
32	22.72	6.20	26.9M
original	23.30	6.55	25.6M

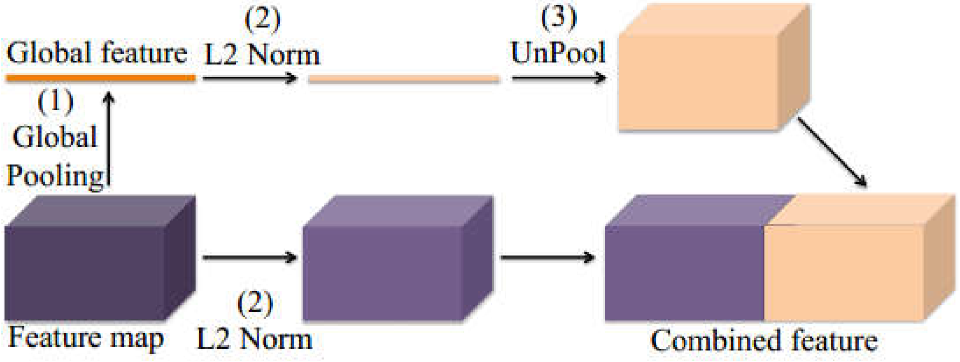
Benefits against Network Depth



	Original		Our re-implementation		SE-module	
	top-1 err.	top-5 err.	top-1 err.	top-5 err.	top-1 err.	top-5 err.
ResNet-50 [1]	24.7	7.8	24.80	7.48	$\mathbf{23.29}_{(\mathbf{1.51})}$	$\mathbf{6.62}_{(\mathbf{0.86})}$
ResNet-101 [1]	23.6	7.1	23.17	6.52	$\mathbf{22.38}_{(\mathbf{0.79})}$	$\mathbf{6.07}_{(\mathbf{0.45})}$
ResNet-152 [1]	23.0	6.7	22.42	6.34	$\mathbf{21.57}_{(\mathbf{0.85})}$	$\mathbf{5.73}_{(\mathbf{0.61})}$

Table 1. Error rates (%) of single-crop results on the ImageNet-1k validation set.

对比



(e) ParseNet contexture module overview.

操作很相似

Parsenet： concatenate feature maps

SENet: self-attetion

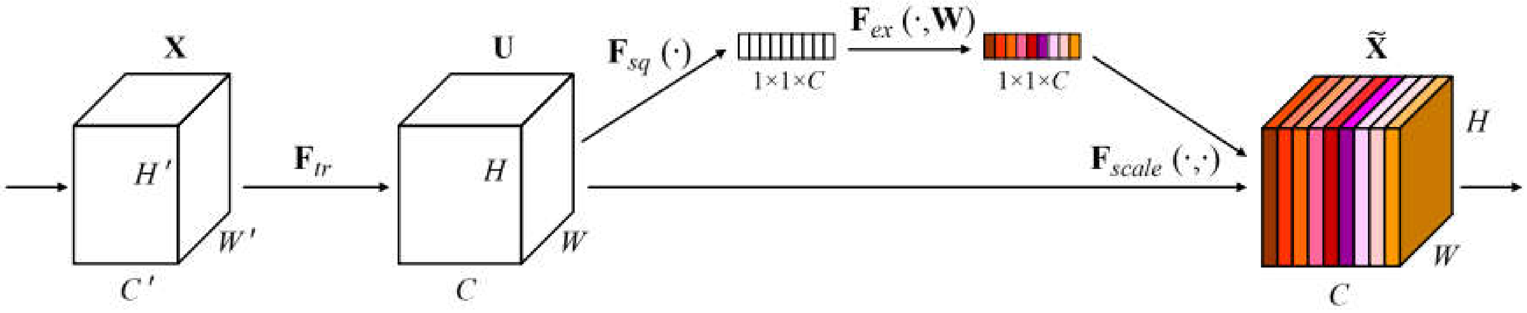


Fig. 1. A Squeeze-and-Excitation block.

谢谢

请批评指正