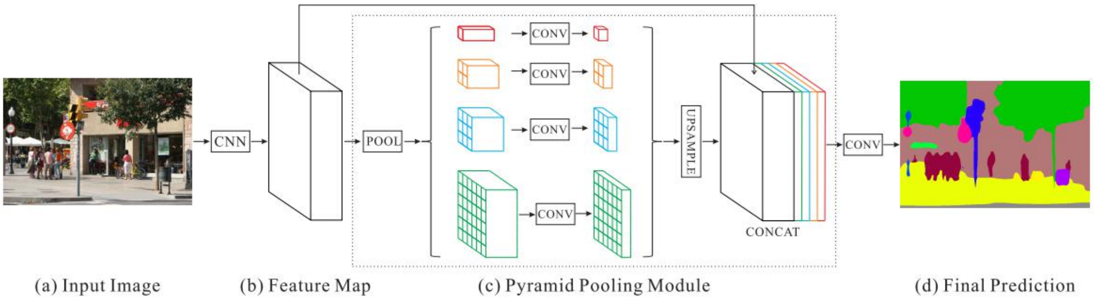
# Introduction



PSPN propose a new network structure focus on different size sub-region. Thus it can using more background information to predict. This is different from the pose estimation or segmentation problem I have learned.

They also use a auxiliary loss in the ResNet to developed an effective optimization strategy.

# Important Observations (the key points to consider)

## Mismatched Relationship

We may predict wrong labels ‘car’ on a boat floating on the river. But the common knowledge is that a car is seldom over a river. Lack of the ability to collect contextual information increases the chance of misclassification.

### Confusion Categories

It’ s hard th separates confusion categories eg. mountain and hill; wall, house, building and skyscraper.

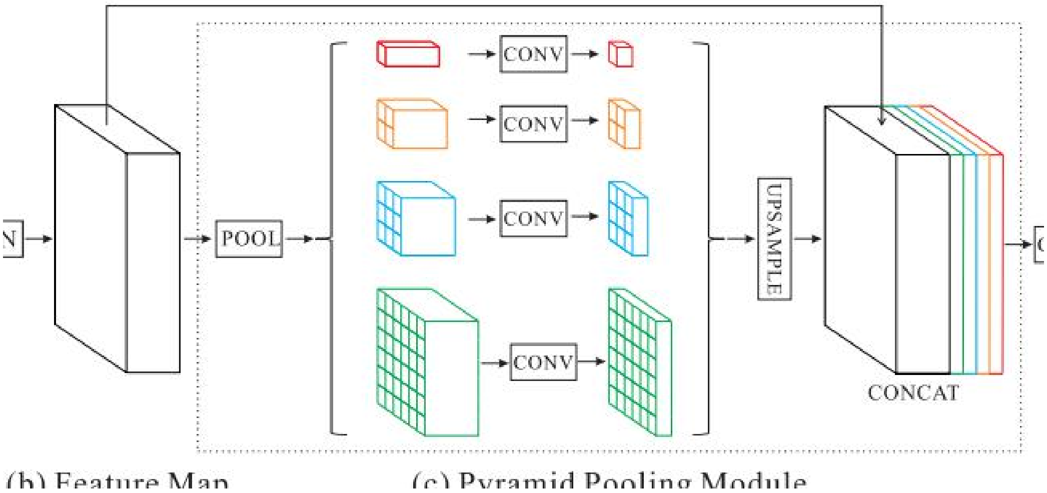
### Inconspicuous Classes

Several small-size objects and stuff, like streetlight and signboard, are hard to find while they may be of great importance.

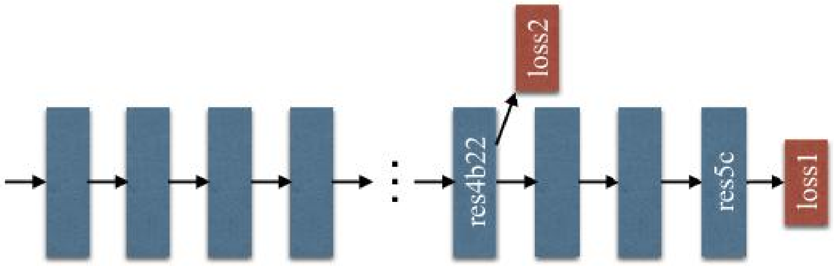
## Implementing Details

### Pyramid Pooling Module

We use pooling to obtain different size (1*1,2*2,3*3,7*7) feature map.Later we use 1*1 convolution layer to reduce the dimension.Then we directly upsample the low-dimension feature maps to get the same size feature as the original feature map via bilinear interpolation. Finally, different levels of features are concatenated as the final pyramid pooling global feature.



## Deep Supervision for ResNet-Based FCN



Apart from the main branch using softmax loss to train the final classifier, another classifier is applied after the fourth stage, i.e., the res4b22 residue block. Different from relay backpropagation that blocks the backward auxiliary loss to several shallow layers, we let the two loss functions pass through all previous layers. The auxiliary loss helps optimize the learning process, while the master branch loss takes the most responsibility. We add weight to balance the auxiliary loss.

## Others

We use the “poly” learning rate policy where current learning rate equals to the base one multiplying (1 - iter maxiter ) power. We set base learning rate to 0.01 and power to 0.9. Momentum and weight decay are set to 0.9 and 0.0001 respectively.

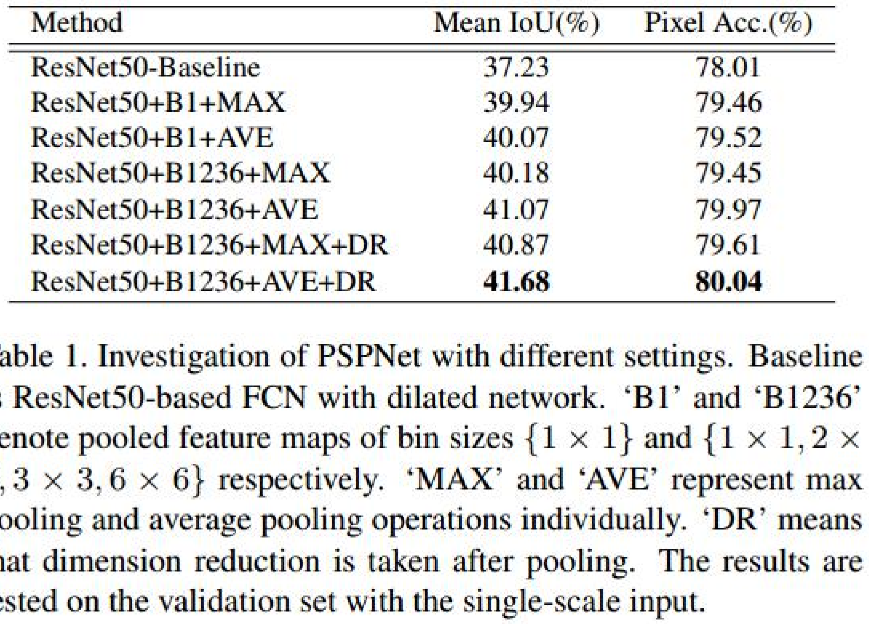
For data augmentation, we adopt random mirror and random resize between 0.5 and 2 for all datasets, and additionally add random rotation between - 10 and 10 degrees, and random Gaussian blur for ImageNet and PASCAL VOC.

An appropriately large “cropsize” can get good performance and “batchsize” in the batch normalization [14] layer is of great importance. We set the “batchsize” to 16 during training.

4.For the auxiliary loss, we set the weight to 0.4 in experiments.

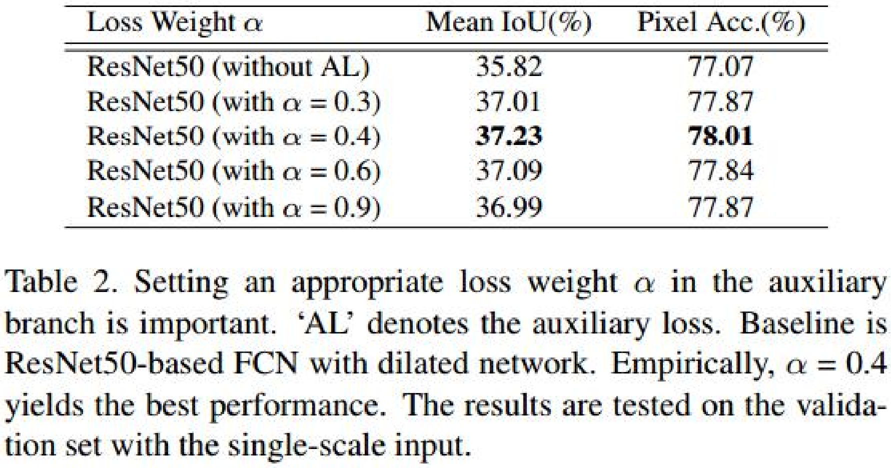
# Results

## Ablation Study



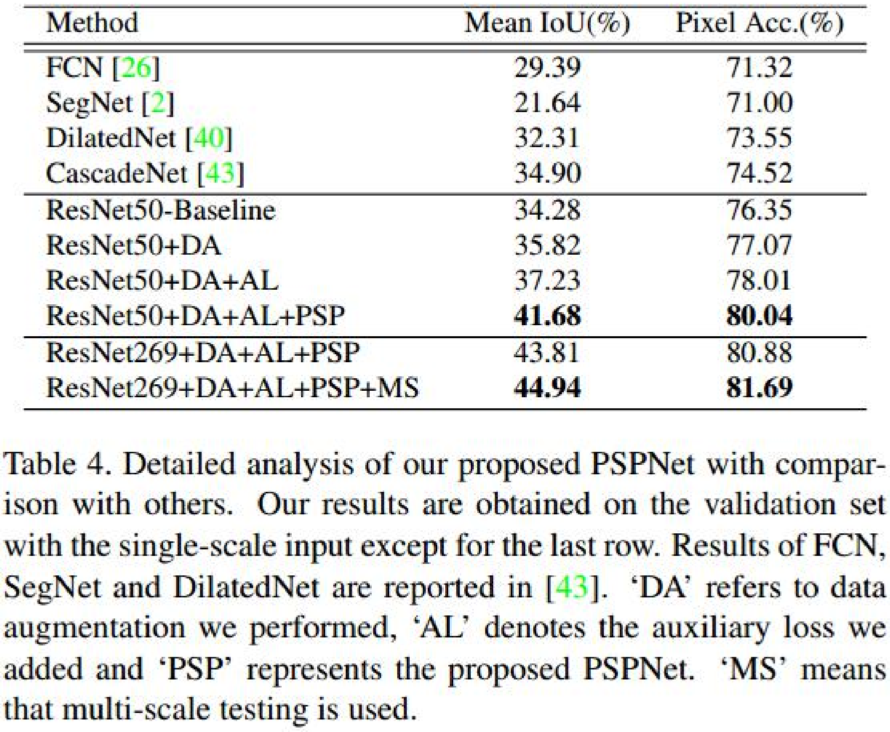
| Method | Mean IoU(%) | Pixel Acc.(%) |
|---|---|---|
| ResNet50-Baseline | 37.23 | 78.01 |
| ResNet50+B1+MAX | 39.94 | 79.46 |
| ResNet50+B1+AVE | 40.07 | 79.52 |
| ResNet50+B1236+MAX | 40.18 | 79.45 |
| ResNet50+B1236+AVE | 41.07 | 79.97 |
| ResNet50+B1236+MAX+DR | 40.87 | 79.61 |
| ResNet50+B1236+AVE+DR | **41.68** | **80.04** |

Table 1. Investigation of PSPNet with different settings. Baseline is ResNet50-based FCN with dilated network. 'B1' and 'B1236' denote pooled feature maps of bin sizes $\{1 \times 1\}$ and $\{1 \times 1, 2 \times 2, 3 \times 3, 6 \times 6\}$ respectively. 'MAX' and 'AVE' represent max pooling and average pooling operations individually. 'DR' means that dimension reduction is taken after pooling. The results are tested on the validation set with the single-scale input.



| Loss Weight $\alpha$ | Mean IoU(%) | Pixel Acc.(%) |
|---|---|---|
| ResNet50 (without AL) | 35.82 | 77.07 |
| ResNet50 (with $\alpha = 0.3$) | 37.01 | 77.87 |
| ResNet50 (with $\alpha = 0.4$) | **37.23** | **78.01** |
| ResNet50 (with $\alpha = 0.6$) | 37.09 | 77.84 |
| ResNet50 (with $\alpha = 0.9$) | 36.99 | 77.87 |

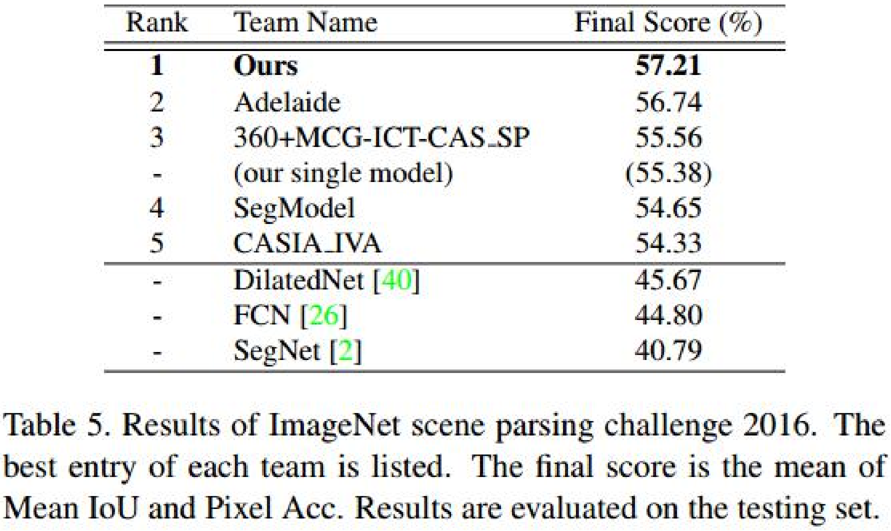
Table 2. Setting an appropriate loss weight $\alpha$ in the auxiliary branch is important. 'AL' denotes the auxiliary loss. Baseline is ResNet50-based FCN with dilated network. Empirically, $\alpha = 0.4$ yields the best performance. The results are tested on the validation set with the single-scale input.



| Method | Mean IoU(%) | Pixel Acc.(%) |
|---|---|---|
| FCN [26] | 29.39 | 71.32 |
| SegNet [2] | 21.64 | 71.00 |
| DilatedNet [40] | 32.31 | 73.55 |
| CascadeNet [43] | 34.90 | 74.52 |
| ResNet50-Baseline | 34.28 | 76.35 |
| ResNet50+DA | 35.82 | 77.07 |
| ResNet50+DA+AL | 37.23 | 78.01 |
| ResNet50+DA+AL+PSP | **41.68** | **80.04** |
| ResNet269+DA+AL+PSP | 43.81 | 80.88 |
| ResNet269+DA+AL+PSP+MS | **44.94** | **81.69** |

Table 4. Detailed analysis of our proposed PSPNet with comparison with others. Our results are obtained on the validation set with the single-scale input except for the last row. Results of FCN, SegNet and DilatedNet are reported in [43]. 'DA' refers to data augmentation we performed, 'AL' denotes the auxiliary loss we added and 'PSP' represents the proposed PSPNet. 'MS' means that multi-scale testing is used.

## Results in Challenge



| Rank | Team Name | Final Score (%) |
|---|---|---|
| **1** | **Ours** | **57.21** |
| 2 | Adelaide | 56.74 |
| 3 | 360+MCG-ICT-CAS_SP | 55.56 |
| - | (our single model) | (55.38) |
| 4 | SegModel | 54.65 |
| 5 | CASIA_IVA | 54.33 |
| - | DilatedNet [40] | 45.67 |
| - | FCN [26] | 44.80 |
| - | SegNet [2] | 40.79 |

Table 5. Results of ImageNet scene parsing challenge 2016. The best entry of each team is listed. The final score is the mean of Mean IoU and Pixel Acc. Results are evaluated on the testing set.