## **Research paper-2: Pyramid Scene Parsing Network**

Source: <https://arxiv.org/pdf/1612.01105.pdf>

## **Summary**

Introduction:

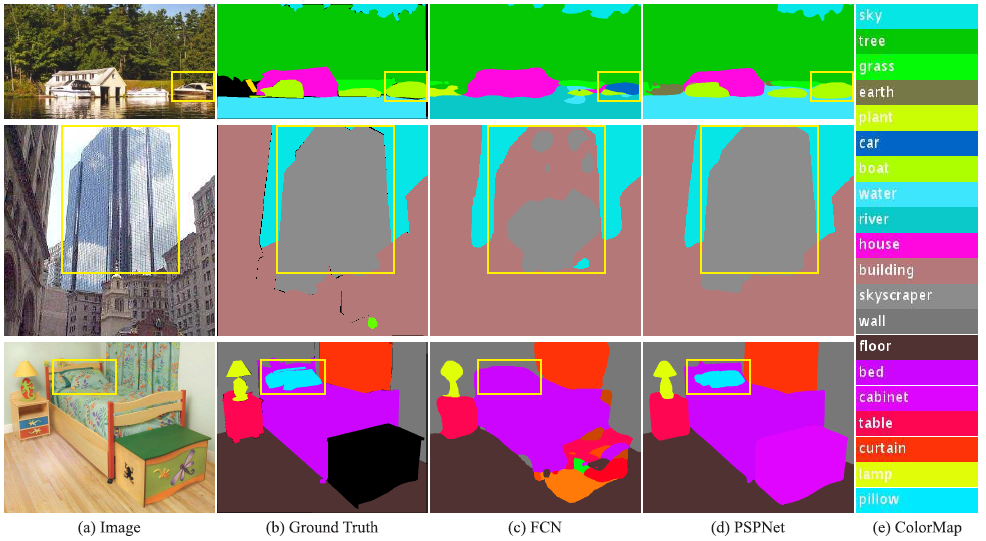
* Scene parsing is based on semantic segmentation, and it is a fundamental topic in computer vision. The goal is to assign each pixel in the image a category label where Scene parsing provides a complete understanding of the scene.
* The State-of-the-art scene parsing frameworks are mostly based on the fully convolutional network (FCN), these networks work well for computer vision tasks but still fallback for tasks that involve diverse scenes and unrestricted vocabulary.
* The fallback is due to the similar appearance of objects and viewing the image regardless of context prior, FCN based models lack the suitable strategy to utilize context and global scene category clues for scene understanding.
* The author proposes Pyramid Scene Parsing Network(PSPNet) where it takes the global
* context of the image into account for Scene Understanding.
* Scene parsing has potential applications in computer vision tasks like robot sensing, autonomous navigation, etc.

Related Work:

* In recent years there have been much research works on scene understanding like dilated convolutions, deconvolution network to learn the segmentation mask, Preprocessing using Conditional random field (CRF), etc. where some of them use context information.
* Most of the recent works include convolution networks that are replaced in place of the fully-connected layer that is used for classification.

Experimental Observations:

The experiment was performed with the ADE20K dataset which contains 150 object category labels (e.g., wall, sky, and tree) and 1,038 image-level scene descriptors for comparing some variations of FCN and PSPNET.



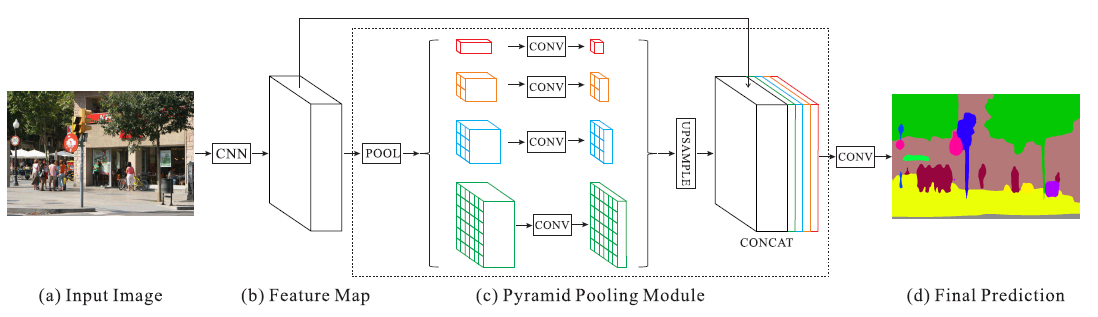
* **Mismatched Relationship:** The Lack of the ability to collect contextual information increases the chance of misclassification, Context relationship is important for complex scene understanding. For example, FCN predicts the boat in the yellow box as a “car” based on its appearance. But the common knowledge is that a car is seldom over a river.
* **Confusion Categories:** There was confusion in the labels category for objects that had similar appearances that are confusing in classification. Examples of building and skyscraper. FCN predicts a part of the skyscraper as a part of the building.
* **Inconspicuous Classes:** Some Small-size objects can have greater importance like a Traffic light, signboard, etc but it is hard to get them right for FCN based networks as they overlook global context and different subregion of the scene example the pillow is very similar to the bed sheet in terms of color and texture which is misclassified.

Pyramid Scene Parsing Network

The Pyramid Scene Parsing Network(Pspnet) consists of a pyramid pooling module as an effective global context prior which represents global context information and also considers sub-region context which is helpful for accurate prediction.

The pspnet is described in below image as below

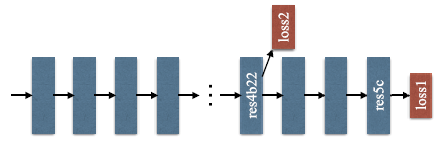
1. Input image
2. Feature map
3. Pyramid pooling module
4. Final Prediction



implementation Overview:

1. **Input image:**
   1. This is the input image with some resolution
2. **Feature map**:
   1. Feature map can be obtained by using Restnet with dilated convolutions aka atrous convolution which is the same as Deep\_lab network strategy.
   2. The final output size of the feature map should be ⅛ size of the input image
3. **Pyramid parsing module:**
   1. Global average pooling and Sub average pooling is applied to feature maps to get different sub-region representations as Sub-regions: Red:1x1, Orange:2x2, Blue:3x3, Green:6x6.
   2. 1x1 convolution is applied to reduce the context of feature map to 1/number of the subregion that is ¼.
   3. Upsampling is performed on subregions to get the size same as original feature map and finally, all feature maps are concatenated to get the final feature representation, which carries both local and global context information.
4. **Final Prediction:** The final representation is fed into a convolution layer to get the final per-pixel prediction.

Auxiliary loss with Pspnet:



Dilated RestNet101 with Auxiliary loss

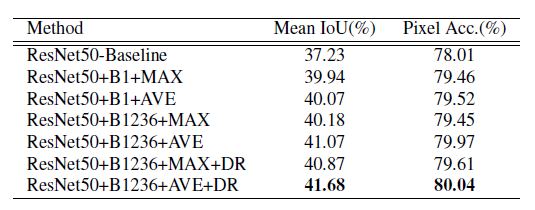
Consider a restnet101 that is used to obtain a feature map as in the figure above

* The pspnet with auxiliary loss has an extra auxiliary loss apart from the main softmax loss to train the network.
* The Auxiliary loss is applied after the fourth stage, i.e., the res4b22 of residue block.
* The auxiliary loss helps to optimize the learning process and the main loss takes the most responsibility and the final loss is the sum of the weighted auxiliary loss and main loss.
* In the testing phase, the auxiliary loss is not considered for loss calculation of final predictions.

Ablation Study for PSPNet:

The ADE20K dataset is used for some experimentation of image scene parsing and more about implementation detail in a research paper.

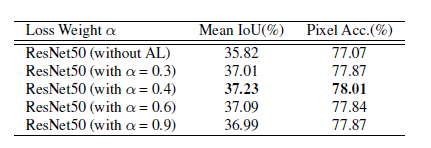
* Used “poly” learning rate policy and with the base learning rate set to 0.01 and power to 0.9.
* Used Data augmentation with random mirror and random resize between 0.5 and 2.
* Used the Batch normalization layer and the “batch size” set to 16.



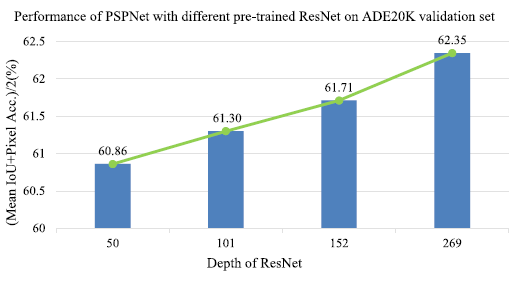
PSPNet with different settings

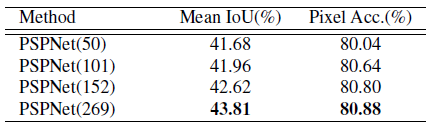
* The baseline is ResNet50-based FCN with a dilated network.
* ‘B1’ and ‘B1236’ denote pooled feature maps of bin sizes 1x1 and 1x1; 2x2; 3x3; 6x6 respectively.
* ‘MAX’ and ‘AVE’ represent max pooling and average pooling operations individually. ‘DR’ means that dimension reduction is taken after pooling.
* Average pooling works better than max-pooling in all settings and With dimension reduction, the performance is further enhanced.
* Restnet50+B1236+Ave+DR outperform all models with Mean IOU of 41.68 in the various setting.

Ablation Study with Auxiliary Loss:



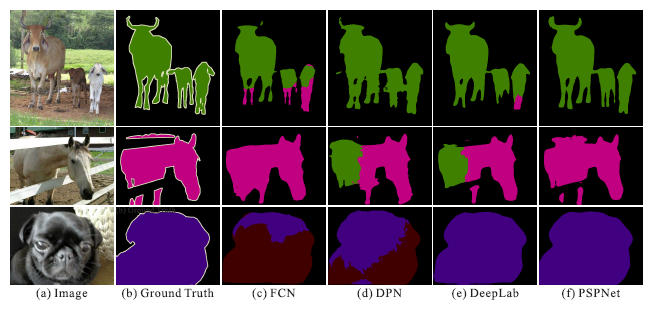
* The baseline uses ResNet50- based FCN with a dilated network.
* ‘AL’ denotes the auxiliary loss,’α’ denotes the weight where auxiliary loss helps to optimize the learning process and the auxiliary loss weight is between 0 and 1.
* ResNet50 with α=0.4 yields the best performance when compared to all models.





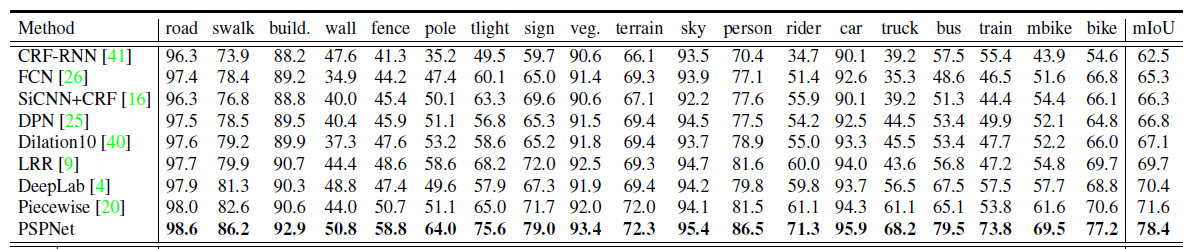
* The performance improves as the depth of the ResNet increases as shown above.

Visual comparison with PASCAL VOC Data:



Visual comparison on PASCAL VOC (a) Image. (b) Ground Truth. (c) FCN. (d) DPN. (e) DeepLab.

Comparison with Cityscapes Data:



The above is the Per-class MIOU as shown where Pspnet outperforms all other networks with the best results.

**Conclusion:**

* PspNet is an effective network for complex scene understanding with a global pyramid pooling feature that provides additional contextual information
* All the failure cases that are discovered when FCN methods are applied to scene parsing are totally eliminated by PspNet.
* Pspnet holds the first place in ImageNet scene parsing challenge 2016, PASCAL VOC 2012 benchmark, and Cityscapes benchmark.
* PspNet outperforms every other model that one can imagine for image segmentation tasks and it has created a new benchmark to find new better algorithms for scene parsing.