In this work, we revisit atrous convolution, a powerful tool to explicitly adjust filter’s field-of-view as well as control the resolution of feature responses computed by Deep Convolutional Neural Networks, in the application of semantic image segmentation. To handle the problem of segmenting objects at multiple scales, we design modules which employ atrous convolution in cascade or in parallel to capture multi-scale context by adopting multiple atrous rates. Furthermore, we propose to augment our previously proposed Atrous Spatial Pyramid Pooling module, which probes convolutional features at multiple scales, with image-level features encoding global context and further boost performance. We also elaborate on implementation details and share our experience on training our system. The proposed ‘DeepLabv3’ system significantly improves over our previous DeepLab versions without DenseCRF post-processing and attains comparable performance with other state-of-art models on the PASCAL VOC 2012 semantic image segmentation benchmark.
found important to be trained as well. We experiment with laying out the modules in cascade or in parallel (specifically, Atrous Spatial Pyramid Pooling (ASPP) method [10]). We discuss an important practical issue when applying a $3 \times 3$ atrous convolution with an extremely large rate, which fails to capture long range information due to image boundary effects, effectively simply degenerating to $1 \times 1$ convolution, and propose to incorporate image-level features into the ASPP module. Furthermore, we elaborate on implementation details and share experience on training the proposed models, including a simple yet effective bootstrapping method for handling rare and finely annotated objects. In the end, our proposed model, ‘DeepLabv3’ improves over our previous works [9, 10] and attains performance of 85.7% on the PASCAL VOC 2012 test set without DenseCRF post-processing.

2. Related Work

It has been shown that global features or contextual interactions [29, 67, 38, 42, 23, 80] are beneficial in correctly classifying pixels for semantic segmentation. In this work, we discuss four types of Fully Convolutional Networks (FCNs) [66, 53] (see Fig. 2 for illustration) that exploit context information for semantic segmentation [26, 14, 54, 8, 85, 48, 57, 65, 78].

Image pyramid: The same model, typically with shared weights, is applied to multi-scale inputs. Feature responses from the small scale inputs encode the long-range context, while the large scale inputs preserve the small object details. Typical examples include Farabet et al. [20] who transform the input image through a Laplacian pyramid, feed each scale input to a DCNN and merge the feature maps from all the scales. [18, 61] apply multi-scale inputs sequentially from coarse-to-fine, while [48, 11, 10] directly resize the input for several scales and fuse the features from all the scales. The main drawback of this type of models is that it does not scale well for larger/deeper DCNNs (e.g., networks like [28, 82, 77]) due to limited GPU memory and thus it is usually applied during the inference stage [15].

Encoder-decoder: This model consists of two parts: (a) the encoder where the spatial dimension of feature maps is gradually reduced and thus longer range information is more easily captured in the deeper encoder output, and (b) the decoder where object details and spatial dimension are gradually recovered. For example, [53, 56] employ deconvolution [83] to learn the upsampling of low resolution feature responses. SegNet [3] reuses the pooling indices from the encoder and learn extra convolutional layers to densify the feature responses, while U-Net [63] adds skip connections from the encoder features to the corresponding decoder activations, and [21] employs a Laplacian pyramid reconstruction network. More recently, RefineNet [47] and [62, 60, 35] have demonstrated the effectiveness of models based on encoder-decoder structure on several semantic segmentation benchmarks. This type of model is also explored in the context of object detection [49, 68].

Context module: This model contains extra modules laid out in cascade to encode long-range context. One effective method is to incorporate DenseCRF [40] (with efficient high-dimensional filtering algorithms [2]) to DCNNs [9, 10]. Furthermore, [85, 48, 65] propose to jointly train both the CRF and DCNN components, while [52, 81] employ several extra convolutional layers on top of the belief maps of DCNNs (belief maps are the final DCNN feature maps that contain output channels equal to the number of predicted classes) to capture context information. Recently, [37] proposes to learn a general and sparse high-dimensional convolution (bilateral convolution), and [73, 7] combine Gaussian Conditional Random Fields and DCNNs for semantic segmentation.

Spatial pyramid pooling: This model employs spatial pyramid pooling [24, 43] to capture context at several ranges. The image-level features are exploited in ParseNet [51] for global context information. DeepLabv2 [10] proposes atrous spatial pyramid pooling (ASPP), where parallel atrous convolution layers with different rates capture multi-scale information. Recently, Pyramid Scene Parsing Net (PSP) [84] performs spatial pooling at several grid scales and demonstrates outstanding performance on several semantic segmentation benchmarks. There are other methods based on LSTM
We then discuss the proposed modules with atrous convolution as a context module and tool for spatial pyramid pooling. Our proposed framework is general in the sense that it could be applied to any network. To be concrete, we duplicate several copies of the original last block in ResNet [28] and arrange them in cascade, and also revisit the ASPP module [10] which contains several atrous convolutions in parallel. Note that our cascaded modules are applied directly on the feature maps instead of belief maps. For the proposed modules, we experimentally find it important to train with batch normalization [34]. To further capture global context, we propose to augment ASPP with image-level features, similar to [51, 84].

**Atrous convolution:** Models based on atrous convolution have been actively explored for semantic segmentation. For example, [76] experiments with the effect of modifying atrous rates for capturing long-range information, [75] adopts hybrid atrous rates within the last two blocks of ResNet, while [17] further proposes to learn the deformable convolution which samples the input features with learned offset, generalizing atrous convolution. To further improve segmentation model accuracy, [74] exploits image captions, [36] utilizes video motion, and [39] incorporates depth information. Besides, atrous convolution has been applied to object detection by [58, 16, 33].

### 3. Methods

In this section, we review how atrous convolution is applied to extract dense features for semantic segmentation. We then discuss the proposed modules with atrous convolution modules employed in cascade or in parallel.

#### 3.1. Atrous Convolution for Dense Feature Extraction

Deep Convolutional Neural Networks (DCNNs) [44] deployed in fully convolutional fashion [66, 53] have shown to be effective for the task of semantic segmentation. However, the repeated combination of max-pooling and striding at consecutive layers of these networks significantly reduces the spatial resolution of the resulting feature maps, typically by a factor of 32 across each direction in recent DCNNs [41, 69, 28]. Deconvolutional layers (or transposed convolution) [33, 53, 56, 3, 63, 60] have been employed to recover the spatial resolution. Instead, we advocate the use of ‘atrous convolution’, originally developed for the efficient computation of the undecimated wavelet transform in the “algorithmé á trous” scheme of [32] and used before in the DCNN context by [22, 66, 58].

Consider two-dimensional signals, for each location $i$ on the output $y$ and a filter $w$, atrous convolution is applied over the input feature map $x$:

$$y[i] = \sum_k x[i + r \cdot k]w[k]$$  \hspace{1cm} (1)

where the atrous rate $r$ corresponds to the stride with which we sample the input signal, which is equivalent to convolving the input $x$ with upsampled filters produced by inserting $r - 1$ zeros between two consecutive filter values along each spatial dimension (hence the name atrous convolution where the French word trous means holes in English). Standard convolution is a special case for rate $r = 1$, and atrous convolution allows us to adaptively modify filter’s field-of-view by changing the rate value. See Fig. 1 for illustration.

Atrous convolution also allows us to explicitly control how densely to compute feature responses in fully convolutional networks. Here, we denote by `output_stride` the ratio of input image spatial resolution to final output resolution. For the DCNNs [41, 69, 28] deployed for the task of image classification, the final feature responses (before fully connected layers or global pooling) is 32 times smaller than the input image dimension, and thus `output_stride = 32`. If one would like to double the spatial density of computed feature responses in the DCNNs (i.e., `output_stride = 16`), the stride of last pooling or convolutional layer that decreases resolution is set to 1 to avoid signal decimation. Then, all subsequent convolutional layers are replaced with atrous convolutional layers having rate $r = 2$. This allows us to extract denser feature responses without requiring learning any extra parameters. Please refer to [10] for more details.

#### 3.2. Going Deeper with Atrous Convolution

We first explore designing modules with atrous convolution laid out in cascade. To be concrete, we duplicate several copies of the last ResNet block, denoted as block4 in Fig. 3, and arrange them in cascade. There are three convolutions in each block, and the last convolution contains stride 2 except the one in last block, similar to original ResNet. The motivation behind this model is that the introduced striding makes it easy to capture long range information in the deeper blocks. For example, the whole image feature could be summarized in the last small resolution feature map, as illustrated in Fig. 3 (a). However, we discover that the consecutive striding is harmful for semantic segmentation (see Tab. 1 in Sec. 4) since detail information is decimated, and thus we apply atrous convolution with rates determined by the desired `output_stride` value, as shown in Fig. 3 (b) where `output_stride = 16`.

In this proposed model, we experiment with cascaded ResNet blocks up to block7 (i.e., extra block5, block6, block7 as replicas of block4), which has `output_stride = 256` if no atrous convolution is applied.
3.2.1 Multigrid

Motivated by multigrid methods which employ a hierarchy of grids of different sizes [4, 72, 5, 59] and following [75, 17], we adopt different atrous rates within block4 to block7 in the proposed model. In particular, we define as Multi_Grid = (r1, r2, r3) the unit rates for the three convolutional layers within block4 to block7. The final atrous rate for the convolutional layer is equal to the multiplication of the unit rate and the corresponding rate. For example, when output_stride = 16 and Multi_Grid = (1, 2, 4), the three convolutions will have rates = 2 × (1, 2, 4) = (2, 4, 8) in the block4, respectively.

3.3. Atrous Spatial Pyramid Pooling

We revisit the Atrous Spatial Pyramid Pooling proposed in [10], where four parallel atrous convolutions with different atrous rates are applied on top of the feature map. ASPP is inspired by the success of spatial pyramid pooling [24, 43, 27] which showed that it is effective to resample features at different scales for accurately and efficiently classifying regions of an arbitrary scale. Different from [10], we include batch normalization within ASPP.

ASPP with different atrous rates effectively captures multi-scale information. However, we discover that as the sampling rate becomes larger, the number of valid filter weights (i.e., the weights that are applied to the valid feature region, instead of padded zeros) becomes smaller. This effect is illustrated in Fig. 4 when applying a 3 × 3 filter to a 65 × 65 feature map with different atrous rates. In the extreme case where the rate value is close to the feature map size, the 3 × 3 filter, instead of capturing the whole image context, degenerates to a simple 1 × 1 filter since only the center filter weight is effective.

To overcome this problem and incorporate global context information to the model, we adopt image-level features, similar to [51, 84]. Specifically, we apply global average pooling on the last feature map of the model, feed the resulting image-level features to a 1 × 1 convolution with 256 filters (and batch normalization [34]), and then bilinearly upsample the feature to the desired spatial dimension. In the end, our improved ASPP consists of (a) one 1 × 1 convolution and three 3 × 3 convolutions with rates = (6, 12, 18) when output_stride = 16 (all with 256 filters and batch normalization), and (b) the image-level features, as shown in Fig. 5. Note that the rates are doubled when output_stride = 8. The resulting features from all the branches are then concatenated and pass through another 1 × 1 convolution (also with 256 filters and batch normalization) before the final 1 × 1 convolution which generates the final logits.

4. Experimental Evaluation

We adapt the ImageNet-pretrained [64] ResNet [28] to the semantic segmentation by applying atrous convolution to extract dense features. Recall that output_stride is defined as the ratio of input image spatial resolution to final out-
put resolution. For example, when \( output\_stride = 8 \), the last two blocks (block3 and block4 in our notation) in the original ResNet contains atrous convolution with \( rate = 2 \) and \( rate = 4 \) respectively. Our implementation is built on TensorFlow [1].

We evaluate the proposed models on the PASCAL VOC 2012 semantic segmentation benchmark [19] which contains 20 foreground object classes and one background class. The original dataset contains 1, 464 \((train)\), 1, 449 \((val)\), and 1, 456 \((test)\) pixel-level labeled images for training, validation, and testing, respectively. The dataset is augmented by the extra annotations provided by [25], resulting in 10, 582 \((trainaug)\) training images. The performance is measured in terms of pixel intersection-over-union (IOU) averaged across the 21 classes.

### 4.1. Training Protocol

In this subsection, we discuss details of our training protocol.

**Learning rate policy:** Similar to [51, 10], we employ a “poly” learning rate policy where the initial learning rate is multiplied by \( 1 - \frac{iter}{max\_iter} \)^power with power = 0.9.

**Crop size:** Following the original training protocol [9, 10], patches are cropped from the image during training. For atrous convolution with large rates to be effective, large crop size is required, otherwise, the filter weights with large atrous rate are mostly applied to the padded zero region. We thus employ crop size to be 513 during both training and test on PASCAL VOC 2012 dataset.

**Batch normalization:** Our added modules on top of ResNet all include batch normalization parameters [34], which we found important to be trained as well. Since large batch size is required to train batch normalization parameters, we employ \( output\_stride = 16 \) and compute the batch normalization statistics with a batch size of 16. The batch normalization parameters are trained with decay = 0.9997. After training on the \( trainaug \) set with 30K iterations and initial learning rate = 0.007, we then freeze batch normalization parameters, employ \( output\_stride = 8 \), and train on the official PASCAL VOC 2012 \( trainval \) set for another 30K iterations and smaller base learning rate = 0.001. Note that atrous convolution allows us to control \( output\_stride \) value at different training stages without requiring learning extra model parameters. Also note that training with \( output\_stride = 16 \) is several times faster than \( output\_stride = 8 \) since the intermediate feature maps are spatially four times smaller, but at a sacrifice of accuracy since \( output\_stride = 16 \) provides coarser feature maps.

**Upsampling logits:** In our previous works [9, 10], the target groundtruths are downsampled by 8 during training when \( output\_stride = 8 \). We find it important to keep the groundtruths intact and instead upsample the final logits, since downsampling the groundtruths removes the fine annotations resulting in no back-propagation of details.

**Data augmentation:** We apply data augmentation by randomly scaling the input images (from 0.5 to 2.0) and randomly left-right flipping during training.

### 4.2. Going Deeper with Atrous Convolution

We first experiment with building more blocks with atrous convolution in cascade.

**ResNet-50:** In Tab. 1, we experiment with the effect of \( output\_stride \) when employing ResNet-50 with block7 \((i.e., \ extra\ block5, block6, and block7)\). As shown in the table, in the case of \( output\_stride = 256 \) \((i.e., \ no\ atrous\ convolution\ at\ all)\), the performance is much worse than the others due to the severe signal decimation. When \( output\_stride \) gets larger and apply atrous convolution correspondingly, the performance improves from 20.29% to 75.18%, showing that atrous convolution is essential when building more blocks cascadedly for semantic segmentation.

**ResNet-50 vs. ResNet-101:** We replace ResNet-50 with deeper network ResNet-101 and change the number of cascaded blocks. As shown in Tab. 2, the performance improves

| \( output\_stride \) | 8 | 16 | 32 | 64 | 128 | 256 |
|---------------------|---|----|----|----|-----|-----|
| mIOU                | 75.18 | 73.88 | 70.06 | 59.99 | 42.34 | 20.29 |

Table 1. Going deeper with atrous convolution when employing ResNet-50 with block7 and different \( output\_stride \). Adopting \( output\_stride = 8 \) leads to better performance at the cost of more memory usage.
### Table 2. Going deeper with atrous convolution when employing ResNet-50 and ResNet-101 with different number of cascaded blocks at output_stride = 16. Network structures ‘block4’, ‘block5’, ‘block6’, and ‘block7’ add extra 0, 1, 2, 3 cascaded modules respectively. The performance is generally improved by adopting more cascaded blocks.

| Network   | block4 | block5 | block6 | block7 |
|-----------|--------|--------|--------|--------|
| ResNet-50 | 64.81  | 72.14  | 74.29  | 73.88  |
| ResNet-101| 68.39  | 73.21  | 75.34  | 75.76  |

| Network   | block4 | block5 | block6 | block7 |
|-----------|--------|--------|--------|--------|
| ResNet-50 | 70.23  | 75.67  | 76.09  | 76.66  |
| ResNet-101| 73.14  | 75.78  | 75.96  | 76.11  |

| Network   | block4 | block5 | block6 | block7 |
|-----------|--------|--------|--------|--------|
| ResNet-50 | 73.45  | 75.74  | 75.85  | 76.02  |
| ResNet-101| 71.45  | 74.30  | 74.70  | 74.62  |

### Table 3. Employing multi-grid method for ResNet-101 with different number of cascaded blocks at output_stride = 16. The best model performance is shown in bold.

| Multi-Grid | block4 | block5 | block6 | block7 |
|------------|--------|--------|--------|--------|
| (1, 1, 1)  | 68.39  | 73.21  | 75.34  | 75.76  |
| (1, 2, 1)  | 70.23  | 75.67  | 76.09  | 76.66  |
| (1, 2, 3)  | 73.14  | 75.78  | 75.96  | 76.11  |
| (1, 2, 4)  | 73.45  | 75.74  | 75.85  | 76.02  |
| (2, 2, 2)  | 71.45  | 74.30  | 74.70  | 74.62  |

### Table 4. Inference strategy on the val set. MG: Multi-grid. OS: output_stride. MS: Multi-scale inputs during test. Flip: Add left-right flipped inputs.

| Multi-Grid | ASPP | Image Pooling | mIoU |
|------------|------|---------------|------|
| (1, 1, 1)  |       |               |      |
| (1, 2, 1)  |       |               |      |
| (1, 2, 3)  |       |               |      |
| (1, 2, 4)  |       |               |      |
| (2, 2, 2)  |       |               |      |

### Table 5. Atrous Spatial Pyramid Pooling with multi-grid method and image-level features at output_stride = 16.

| Method        | OS=16 | OS=8 | MS | mIoU |
|---------------|-------|------|----|------|
| MG(1, 2, 1)   |       |      |    | 75.36|
| ASPP(6, 12, 18) | ✓     |      |    | 75.93|
| Image Pooling | ✓     | ✓    |    | 76.58|

### Table 6. Inference strategy on the val set. MG: Multi-grid. ASPP: Atrous spatial pyramid pooling. OS: output_stride. MS: Multi-scale inputs during test. Flip: Add left-right flipped inputs.

| Method        | OS=16 | OS=8 | MS | mIoU |
|---------------|-------|------|----|------|
| MG(1, 2, 4)   |       |      |    | 77.21|
| ASPP(6, 12, 24) | ✓     |      |    | 78.51|
| Image Pooling | ✓     | ✓    |    | 79.45|

### 4.3. Atrous Spatial Pyramid Pooling

We then experiment with the Atrous Spatial Pyramid Pooling (ASPP) module with the main differences from [10] being that batch normalization parameters [34] are fine-tuned and image-level features are included.

**ASPP:** In Tab. 5, we experiment with the effect of incorporating multi-grid in block4 and image-level features to the improved ASPP module. We first fix ASPP = (6, 12, 18) (i.e., employ rates = (6, 12, 18) for the three parallel 3 × 3 convolution branches), and vary the multi-grid value. Employing Multi_Grid = (1, 2, 1) is better than Multi_Grid = (1, 1, 1), while further improvement is attained by adopting Multi_Grid = (1, 2, 4) in the context of ASPP = (6, 12, 18) (cf., the ‘block4’ column in Tab. 3). If we additionally employ another parallel branch with rate = 24 for longer range context, the performance drops slightly by 0.12%. On the other hand, augmenting the ASPP module with image-level feature is effective, reaching the final performance of 77.21%.

**Inference strategy on val set:** Similarly, we apply output_stride = 8 during inference once the model is trained. As shown in Tab. 6, employing output_stride = 8 brings 1.3% improvement over using output_stride = 16, adopting multi-scale inputs and adding left-right flipped images further improve the performance by 0.94% and 0.32%, respectively. The best model with ASPP attains the performance of 79.77%, better than the best model with cascaded atrous convolution modules (79.35%), and thus is selected as our final model for test set evaluation.

**Comparison with DeepLabv2:** Both our best cascaded
model (in Tab. 4) and ASPP model (in Tab. 6) (in both cases without DenseCRF post-processing or MS-COCO pre-training) already outperform DeepLabv2 (77.69% with DenseCRF and pretrained on MS-COCO in Tab. 4 of [10]) on the PASCAL VOC 2012 val set. The improvement mainly comes from including and fine-tuning batch normalization parameters [34] in the proposed models and having a better way to encode multi-scale context.

**Blocks + ASPP:** We have tried to combine both proposed models (i.e., cascaded modules and parallel modules). In particular, we employ atrous spatial pyramid pooling on top of the features after block5. However, we have not observed any improvement.

**Qualitative results:** We provide qualitative visual results of our best ASPP model in Fig. 6. As shown in the figure, our model is able to segment objects very well without any DenseCRF post-processing.

**Failure mode:** As shown in the bottom row of Fig. 6, our model has difficulty in segmenting (a) sofa vs. chair, (b) dining table and chair, and (c) rare view of objects in the training set.

**Pretrained on COCO:** For comparison with other state-of-art models, we further pretrain our best ASPP model on MS-COCO dataset [50]. From the MS-COCO trainval_minunis_minival set, we only select the images that have annotation regions larger than 1000 pixels and contain the classes defined in PASCAL VOC 2012, resulting in about 60K images for training. Besides, the MS-COCO classes not defined in PASCAL VOC 2012 are all treated as background class. After pretraining on MS-COCO dataset, our proposed model attains performance of 82.7% on val set when using output_stride = 8, multi-scale inputs and adding left-right flipped images during inference. We adopt smaller initial learning rate = 0.0001 and same training protocol as in Sec. 4.1 when fine-tuning on PASCAL VOC 2012 dataset.

**Test set result and an effective bootstrapping method:** We notice that PASCAL VOC 2012 dataset provides higher quality of annotations than the augmented dataset [25], especially for the bicycle class. We thus further fine tune our model on the official PASCAL VOC 2012 trainval set before evaluating on the test set. Specifically, our model is trained with output_stride = 8 (so that annotation details are kept) and the batch normalization parameters are frozen (see Sec. 4.1 for details). Besides, instead of performing pixel hard example mining as [76, 62], we resort to bootstrapping on hard images. In particular, we duplicate the images that contain hard classes (namely bicycle, chair, table, potted-plant, and sofa) in the training set. As shown in Fig. 7, the simple bootstrapping method is effective for segmenting the bicycle class. In the end, our ‘DeepLabv3’ achieves the performance of 85.7% on the test set without any DenseCRF post-processing, as shown in Tab. 7.

**Model pretrained on JFT-300M:** Motivated by the recent work of [70], we further employ the ResNet-101 model which has been pretrained on both ImageNet and the JFT-300M dataset [30, 12, 70], resulting in a performance of 86.9% on PASCAL VOC 2012 test set.

| Method                          | mIOU |
|---------------------------------|------|
| Adelaide_VeryDeep_FCN_VOC [76]  | 79.1 |
| LRR4x_ResNet-CRF [21]           | 79.3 |
| DeepLabv2-CRF [10]              | 79.7 |
| CentraleSupelec Deep G-CRF [7]  | 80.2 |
| HikSeg_COCO [71]                | 81.4 |
| Deep Layer Cascade (LC) [45]    | 82.7 |
| TuSimple [75]                   | 83.1 |
| Large_Kernel_Matters [60]       | 83.6 |
| Multipath-RefineNet [47]        | 84.2 |
| ResNet-38_MS_COCO [77]          | 84.9 |
| PSPNet [84]                     | 85.4 |
| IDW-CNN [74]                    | 86.3 |
| DeepLabv3                       | 85.7 |
| DeepLabv3-JFT                   | 86.9 |

**Table 7. Performance on PASCAL VOC 2012 test set.**

### 5. Conclusions

Our proposed model “DeepLabv3” employs atrous convolution with upsampled filters to extract dense feature maps and to capture long range context. Specifically, to encode multi-scale information, our proposed cascaded module gradually doubles the atrous rates while our proposed atrous spatial pyramid pooling module augmented with image-level features probes the features with filters at multiple sampling rates and effective field-of-views. Our experimental results show that the proposed model significantly improves over previous DeepLab versions and achieves comparable performance with other state-of-art models on the PASCAL VOC 2012 semantic image segmentation benchmark.

**Acknowledgments** We would like to acknowledge valuable discussions with Zbigniew Wojna, the help from Chen Sun, and the support from Google Mobile Vision team.

**References**

[1] M. Abadi, A. Agarwal, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv*:1603.04467, 2016.

[2] A. Adams, J. Baek, and M. A. Davis. Fast high-dimensional filtering using the permutohedral lattice. In *Eurographics*, 2010.

[3] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv*:1511.00561, 2015.
Figure 6. Visualization results on the val set when employing our best ASPP model. The last row shows a failure mode.
Figure 7. Bootstrapping on hard images improves segmentation accuracy for rare and finely annotated classes such as bicycle.

[4] A. Brandt. Multi-level adaptive solutions to boundary-value problems. *Mathematics of computation*, 31(138):333–390, 1977.

[5] W. L. Briggs, V. E. Henson, and S. F. McCormick. *A multigrid tutorial*. SIAM, 2000.

[6] W. Byeon, T. M. Breuel, F. Raue, and M. Liwicki. Scene labeling with lstm recurrent neural networks. In *CVPR*, 2015.

[7] S. Chandra and I. Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep Gaussian CRFs. *arXiv:1603.08358*, 2016.

[8] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *CVPR*, 2016.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected cnfs. In *ICLR*, 2015.

[10] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected cnfs. *arXiv:1606.00915*, 2016.

[11] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *CVPR*, 2016.

[12] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *arXiv:1610.02357*, 2016.

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[14] J. Dai, K. He, and J. Sun. Convolutional feature masking for joint object and stuff segmentation. *arXiv:1412.1283*, 2014.

[15] J. Dai, K. He, and J. Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.

[16] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. *arXiv:1605.06409*, 2016.

[17] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. *arXiv:1703.06211*, 2017.

[18] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *arXiv:1411.4734*, 2014.

[19] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge a retrospective. *IJCV*, 2014.

[20] C. Farabet, C.ouprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *PAMI*, 2013.

[21] G. Ghiasi and C. C. Fowlkes. Laplacian reconstruction and refinement for semantic segmentation. *arXiv:1605.02264*, 2016.

[22] A. Giusti, D. Ciresan, J. Masci, L. Gambardella, and J. Schmidhuber. Fast image scanning with deep max-pooling convolutional neural networks. In *ICIP*, 2013.

[23] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, IEEE, 2009.

[24] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005.

[25] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

[26] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015.

[27] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014.

[28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv:1512.03385*, 2015.

[29] X. He, R. S. Zemel, and M. Carreira-Perpindn. Multiscale conditional random fields for image labeling. In *CVPR*, 2004.

[30] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. In *NIPS*, 2014.

[31] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[32] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian. A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets: Time-Frequency Methods and Phase Space*, pages 289–297. 1989.

[33] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*, 2017.

[34] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv:1502.03167*, 2015.

[35] M. A. Islam, M. Rochan, N. D. Bruce, and Y. Wang. Gated recurrent neural networks. In *ICLR*, 2015.

[36] S. D. Jain, B. Xiong, and K. Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *CVPR*, 2017.

[37] V. Jampani, M. Kiefel, and P. V. Gehler. Learning sparse high dimensional filters: Image filtering, dense crfs and bilateral crfs. In *ICCV*, 2013.

[38] S. Kong and C. Fowlkes. Recurrent scene parsing with perspective understanding in the loop. *arXiv:1705.07238*, 2017.
[80] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In CVPR, 2012.

[81] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In ICLR, 2016.

[82] S. Zagoruyko and N. Komodakis. Wide residual networks. arXiv:1605.07146, 2016.

[83] M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In ICCV, 2011.

[84] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. arXiv:1612.01105, 2016.

[85] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In ICCV, 2015.

[86] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In CVPR, 2017.