Autofocus Layer for Semantic Segmentation

Yao Qin[1,2(✉)], Konstantinos Kamnitsas[1,3], Siddharth Ancha[1,4], Jay Nanavati[1], Garrison Cottrell[2], Antonio Criminisi[1], and Aditya Nori[1]

[1] Microsoft Research, Cambridge, UK
[2] University of California, San Diego, USA
[3] Imperial College London, London, UK
[4] Carnegie Mellon University, Pittsburgh, USA
yaq007@eng.ucsd.edu

Abstract. We propose the autofocus convolutional layer for semantic segmentation with the objective of enhancing the capabilities of neural networks for multi-scale processing. Autofocus layers adaptively change the size of the effective receptive field based on the processed context to generate more powerful features. This is achieved by parallelising multiple convolutional layers with different dilation rates, combined by an attention mechanism that learns to focus on the optimal scales driven by context. By sharing the weights of the parallel convolutions we make the network scale-invariant, with only a modest increase in the number of parameters. The proposed autofocus layer can be easily integrated into existing networks to improve a model's representational power. We evaluate our mod els on the challenging tasks of multi-organ segmentation in pelvic CT and brain tumor segmentation in MRI and achieve very promising performance.

1 Introduction

Semantic segmentation is a fundamental problem in medical image analysis. Automatic segmentation systems can improve clinical pipelines, facilitating quantitative assessment of pathology, treatment planning and monitoring of disease progression. They can also facilitate large-scale research studies, by extracting measurements from magnetic resonance images (MRI) or computational tomography (CT) scans of large populations in an efficient and reproducible manner.

For high performance, segmentation algorithms are required to use multi-scale context [6], while still aiming for pixel-level accuracy. Multi-scale processing provides detailed cues, such as texture information of a structure, combined with contextual information, such as a structure's surroundings, which can facilitate decisions that are ambiguous when based only on local context. Note that such

---

a mechanism is also part of the human visual system, via foveal and peripheral vision.

A large volume of research has sought algorithms for effective multi-scale processing. An overview of traditional approaches can be found in [6]. Contemporary segmentation systems are often powered by convolutional neural networks (CNNs). The various network architectures proposed to effectively capture image context can be broadly grouped into three categories. The first type creates an image pyramid at multiple scales. The image is down-sampled and processed at different resolutions. Farabet *et al.* trained the same filters to perform on all such versions of an image to achieve scale invariance [5]. In contrast, DeepMedic [9] proposed learning dedicated pathways for several scales, to enable 3D CNNs to extract more patterns from a larger context in a computationally efficient manner. The second type uses an encoder that gradually down-samples to capture more context, followed by a decoder that learns to upsample the segmentations, combining multi-scale context using skip connections [11]. Later extensions include U-net [15], which used a larger decoder to learn upsampling features instead of segmentations as in [11]. Learning to upsample with a decoder, however, increases model complexity and computational requirements, when downsampling may not be even necessary. Finally, driven by this idea, [3,16] proposed dilated convolutions to process greater context without ever downsampling the feature maps. Taking it further, DeepLab [3] introduced the module Atrous Spatial Pyramid Pooling (ASPP), where dilated convolutions with varying rates are applied in parallel to capture multi-scale information. The activations from all scales are naively fused via summation or concatenation.

We propose the *autofocus layer*, a novel module that enhances the multi-scale processing of CNNs by learning to select the 'appropriate' scale for identifying different objects in an image. Our work on autofocus shares similarities with ASPP in that we also use parallel dilated convolutional filters to capture both local and more global context. The crucial difference is that instead of naively aggregating features from all scales, the autofocus layer adaptively chooses the optimal scale to focus on in a data-driven, learned manner. In particular, our autofocus module uses an attention mechanism [1] to indicate the importance of each scale when processing different locations of an image (Fig. 1). The computed attention maps, one per scale, serve as filters for the patterns extracted at that scale. Autofocus also enhances interpretability of a network as the attention maps reveal how it locally 'zooms in or out' to segment different context. Compared to the use of attention in [4], our solution is modular and independent of architecture.

We extensively evaluate and compare our method with strong baselines on two tasks: multi-organ segmentation in pelvic CT and brain tumor segmentation in MRI. We show that thanks to its adaptive nature, the autofocus layer copes well with biological variability in the two tasks, improving performance of a well-established model. Despite its simplicity, our system is competitive with more elaborate pipelines, demonstrating the potential of the autofocus mechanism.

Additionally, autofocus can be easily incorporated into existing architectures by replacing a standard convolutional layer.

2   Method

2.1   Dilated Convolution

As they are fundamental to our work, we first present the basics of dilated convolutions [3,16] while introducing notation. The standard 3D dilated convolutional layer at depth l with dilation rate r can be represented as a mapping $\mathbf{Conv}_l^r : \mathbf{F}_{l-1} \to \mathbf{F}_l^r$, where $\mathbf{F}_{l-1} \in \mathbb{R}^{W' \times H' \times D' \times C'}$ and $\mathbf{F}_l^r \in \mathbb{R}^{W \times H \times D \times C}$ are input and output tensors with C channels (feature maps) of size $(W \times H \times D)$. For neurons in $\mathbf{F}_l^r$, the size $\phi_l^{\{x,y,z\}} \in \mathbb{N}^3$ of their receptive field on the input image can be controlled via r. For dilated convolution layers with kernel size $\boldsymbol{\theta}_l^{\{x,y,z\}} \in \mathbb{N}^3$, $\phi_l^{\{x,y,z\}}$ can be derived recursively as follows:

$$\phi_l^{\{x,y,z\}} = \phi_{l-1}^{\{x,y,z\}} + r_l(\boldsymbol{\theta}_l^{\{x,y,z\}} - \mathbf{1})\boldsymbol{\eta}_l^{\{x,y,z\}}, \tag{1}$$

Here $\boldsymbol{\eta}_l^{\{x,y,z\}} \in \mathbb{N}^3$ denotes the stride of the receptive field at layer l, which is a product of the strides of kernels in preceding layers. It can be observed from Eq. (1) that greater context can be captured by increasing dilation r_l but in less detail as the input signal is probed more sparsely. Thus greater r_l leads to a 'zoom out' behavior. Usually, the dilation rate r is a hyperparameter that is manually set and fixed for each layer. Standard convolution is a special case when $r = 1$. Below we describe the autofocus mechanism that adaptively chooses the optimal dilation rate for different areas of the input.
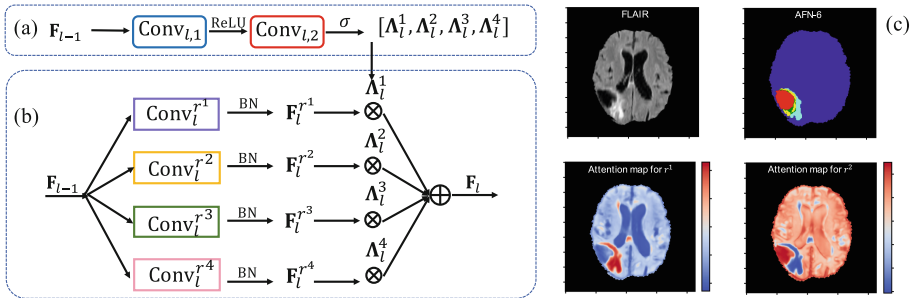

Fig. 1. An autofocus convolutional layer with the number of candidate dilation rates $K = 4$. (a) The attention model. (b) A weighted summation of activations from parallel dilated convolutions. (c) An example of attention maps for a small (r^1) and larger (r^2) dilation rate. The first row is the input and the segmentation result of Afn-6 (described in Sect. 2.3). The second row shows how the module 'zooms out' for more context when processing large or ambiguous structures.

2.2   Autofocus Convolutional Layer

Unambiguously classifying different objects in an image is likely to require different combinations of local and global information. For example, large structures may be better segmented by processing a large receptive field ϕ_l at the expense of fine details, while small objects may require focusing on high resolution local information. Consequently, architectures that statically define multi-scale processing may be suboptimal. Our adaptive solution, the autofocus module, is summarized in Fig. 1 and formalized in the following.

Given activations of the previous layer $\mathbf{F}_{l-1}$, we capture multi-scale information by processing it in parallel via K convolutional layers with different dilation rates r^k. They produce K tensors $\mathbf{F}_l^{r^k}$ (Fig. 1(b)), each set to have same number of channels C. They detect patterns at K different scales which we merge in a data-driven manner by introducing a soft attention mechanism [1].

Within the module we construct a small attention network (Fig. 1(a)) that processes $\mathbf{F}_{l-1}$. In this work it consists of two convolutional layers. The first, $\mathbf{Conv}_{l,1}$, applies $3 \times 3 \times 3$ kernels, produces half the number of channels than those in $\mathbf{F}_{l-1}$ (empirically chosen) and is followed by a ReLU activation function f. The second, $\mathbf{Conv}_{l,2}$, applies $1 \times 1 \times 1$ filters and produces a tensor with K channels, one per scale. It is followed by an element-wise softmax σ that normalizes the K activations for each voxel to add up to one. Let this normalized output be $\mathbf{\Lambda}_l = [\mathbf{\Lambda}_l^1, \mathbf{\Lambda}_l^2, \cdots, \mathbf{\Lambda}_l^K] \in \mathbb{R}^{W \times H \times D \times K}$. Formally:

$$\mathbf{\Lambda}_l = \sigma(\text{Conv}_{l,2}(f(\text{Conv}_{l,1}(\mathbf{F}_{l-1})))) \tag{2}$$

In the above, $\mathbf{\Lambda}_l^k \in \mathbb{R}^{W \times H \times D}$ is an attention map that corresponds to the k-th scale. For any specific spatial location (voxel), the corresponding K values from the K attention maps $\mathbf{\Lambda}_l^k$ can be interpreted as how much focus to put on each scale. Thus the final output of the autofocus layer is computed by fusing the outputs from the parallel dilated convolutions as follows:

$$\mathbf{F}_l = \sum_{k=1}^{K} \mathbf{\Lambda}_l^k \cdot \mathbf{F}_l^{r^k} \tag{3}$$

where $\cdot$ is an element-wise multiplication. Note that the attention weights $\mathbf{\Lambda}_l^k$ are shared across all channels of tensor $\mathbf{F}_l^{r^k}$ for scale k. Since the attention maps are predicted by a fully convolutional network, different attention is predicted for each voxel, driven by the image context for the optimal choice of scale (Fig. 1(c)).

The increase in representational power offered by each autofocus layer naturally comes with computational requirements as the module is based in parallelism of K dilated convolutional layers. Therefore an appropriate balance should be sought, which we investigate in Sect. 3 with very promising results.

Scale Invariance: The size of some anatomical structures such as bones and organs may vary, while the overall appearance is rather similar. For others, size
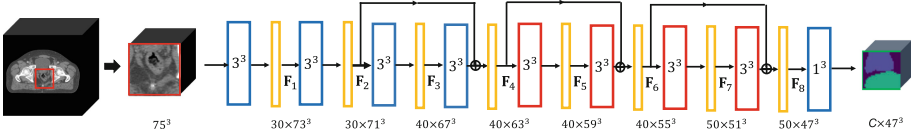
Fig. 2. The AFNET-4 model. Layers 1–2 are standard convolutions and 3–4 are dilated with rate 2. Layers 4–8 are autofocus layers, denoted with red. All layers except the classification layer use 3^3 kernels. Yellow rectangles represent ReLU layers. Residual connections are used. Number and size of feature maps shown as (number $\times$ size).

may correlate with appearance. For instance, the texture of large developed tumors differs from early-stage small tumors. This suggests that scale invariance could be leveraged to regularize learning but must be done appropriately. We make the parallel filters in an autofocus layer share parameters. This makes the number of trainable parameters independent of K, with only the attention module adding parameters over a standard convolution. As a result, each parallel filter seeks patterns with similar appearance but of different sizes. Hence, the network is *adaptively* scale-invariant – the attention mechanism chooses the scale in a data-driven manner, unlike Farabet *et al.* [5], whose network learns shared filters between different scales but naively concatenates all their responses.

2.3 Autofocus Neural Networks

The proposed autofocus layer can be integrated into existing architectures to improve their multi-scale processing capabilities by replacing standard or dilated convolutions. To demonstrate this, we chose DeepMedic (DM) [9] with residual connections [8] as a starting point. DM uses different pathways with high and low resolution inputs for multi-scale processing. Instead, we keep only its high-resolution pathway and seek to empower it with our method. First, we enhance it with standard dilated convolutions with rate 2 in its last 6 hidden layers to enlarge its receptive field, arriving at the BASIC model that serves as another baseline. We now define a family of AFNETs by converting the last n hidden layers of BASIC to autofocus layers—denoted as "AFN-n", where $n \in \{1, \ldots, 6\}$. Figure 2 shows AFNET-4. The proposed AFNETs are trained end-to-end.

3 Evaluation

We extensively evaluate AFNETs on the tasks of multi-organ and brain tumor segmentation. Specifically, on both tasks we perform: (1) a study where we successively add autofocus to more layers of the BASIC network to explore its impact, and (2) comparison of AFNETs with baselines. Finally, (3) we evaluate on the public benchmark BRATS'15 and show that our method competes with state-of-the-art pipelines regardless its simplicity, showing its potential.

Baselines: We compare AFNETs with the previously defined BASIC model to show the contribution of autofocus layer over standard dilated convolutions. Similarly, we compare with DeepMedic [9], denoted as DM, to compare our adaptive multi-scale processing with the static multi-scale pathways. Finally, we place an ASPP module [3] on top of BASIC, comparison of which against AFN-1 shows contribution of the attention mechanism. ASPP-C and ASPP-S represent fusion of ASPP activations via concatenation and summation respectively. Source codes and pretrained models in PyTorch framework are online available at: https://github.com/yaq007/Autofocus-Layer.

3.1   ADD and UW Datasets of Pelvic CT Scans

Material: We use two databases of pelvic CT scans, collected from patients diagnosed with prostate cancer in different clinical centers. The first, referred to as ADD, contains 86 scans with varying number of 512×512 slices and $3\,\mathrm{mm}$ inter-slice spacing. UW consists of 34 scans of 512×512 slices with $1\,\mathrm{mm}$ inter-slice spacing. Expert oncologists manually delineated in all images the following structures: prostate gland, seminal vesicles (SV), bladder, rectum, left femur and right femur. Each scan is normalized so that its intensities have zero mean and unit variance. We also re-sample UW to the spacing of ADD. To produce a stringent test of the models' generalization, we train them for this multi-class problem using the ADD data and then evaluate them on UW data.

Configuration Details: BASIC, ASPP and AFN models were trained with the ADAM optimizer for 300 epochs to minimize the soft dice loss [13]. Each batch consists of 7 segments of size 75^3. The learning rate starts at 0.001 and is reduced to 0.0001 after 200 epochs. We use dilation rates 2, 6, 10 and 14 ($K = 4$) for both ASPP and the autofocus modules. It takes around 20 hours to train an AFNET with 2 NVIDIA TITAN X GPUs. Performance of DeepMedic was obtained by training the public software [9] with default parameters, but without augmentation and by sampling each class equally, similar to other methods.

3.2   Brain Tumor Segmentation Data (BRATS 2015)

Material: The training database of BRATS'15 [12] consists of multi-modal MR scans of 274 cases, along with corresponding annotations of the tumors. We normalize each scan so that intensities belonging to the brain have zero mean and unit variance. For our ablation study, we train all models on the same 193 subjects and evaluate their performance on 54 subjects. The subsets were chosen randomly, including both high and low grade gliomas. Results on the remaining 23 cases aren't reported as they were used for configuration during development. Following standard protocol, we report performance for segmenting the *whole* tumor, *core* and *enhancing* tumor. Finally, to compare with other methods, we train AFNET-6 on all 274 images, segment the 110 test cases of BRATS'15 (no annotations publicly available) and submit predictions for online evaluation.

Table 1. Performance on multi-organ segmentation problem of baseline models and AFN on Uw database, after being trained on ADD. Absolute dice scores are shown.

	BASIC	ASPP-S	ASPP-C	DM	AFN-1	AFN-2	AFN-3	AFN-4	AFN-5	AFN-6
Prostate	50.94	55.83	58.67	69.66	63.36	75.43	**76.66**	75.30	73.81	76.15
Bladder	72.83	81.53	80.43	93.54	88.76	92.38	93.56	**94.49**	94.28	93.32
Rectum	64.39	67.30	67.04	70.74	71.46	76.20	78.45	**79.80**	78.96	78.82
SV	53.97	50.11	59.37	56.75	61.68	65.03	**65.12**	64.83	60.87	63.24
LFemur	91.60	94.13	93.81	94.68	93.24	**95.18**	93.52	94.59	93.42	95.16
RFemur	91.65	94.34	92.78	94.63	91.93	94.38	91.61	94.60	94.16	**95.75**
Mean Dice	70.90	74.98	75.35	78.89	78.41	83.10	83.15	**83.94**	82.58	83.74

Table 2. Ablation study on BRATS'15 training database via cross-validation on 54 random held-out cases. Dice scores shown in format mean(standard deviation).

	BASIC	ASPP-S	ASPP-C	DM	AFN-1	AFN-2	AFN-3	AFN-4	AFN-5	AFN-6
Whole	87.90	87.83	87.90	88.93	88.03	88.63	88.42	88.88	89.19	**89.30**
	(8.57)	(8.36)	(8.08)	(7.05)	(8.29)	(8.02)	(9.28)	(8.31)	(7.87)	(**8.00**)
Core	72.61	74.08	73.58	71.42	73.93	73.43	73.79	73.91	**74.32**	74.11
	(26.39)	(24.16)	(24.57)	(27.48)	(25.04)	(24.90)	(26.00)	(25.49)	(**24.80**)	(24.19)
Enh	74.37	73.06	73.47	**76.08**	73.17	73.94	74.21	74.08	74.48	75.62
	(25.23)	(26.93)	(25.36)	(**25.12**)	(26.98)	(25.83)	(27.58)	(25.70)	(26.56)	(25.02)

Configuration Details: Settings are similar to Kamnitsas *et al.* [9] for a fair comparison. For each method in Table 2 we report the average of three runs with different seeds.

3.3 Results

Ablation Study: Results from the ablation study on the cervical CT database and the BRATS database are summarized in Tables 1 and 2 respectively. We observe the following: (a) Building AFN-1 by converting the last layer of BASIC to autofocus improves performance, while (b) the gains surpass those by the popular ASPP for most classes of the tasks. It is important to note that ASPP adds multiple parallel convolutional layers without sharing weights between them. This incurs a large increase in the number of parameters, and is therefore partly the reason for improvements of ASPP over BASIC (see Table 3). (c) Converting more layers of the BASIC baseline to autofocus layers tends to improve performance. An exception is AFN-4 vs. AFN-5/6 on the Uw dataset. We speculate that

Table 3. Number of trainable parameters in convolutional kernels of different models.

Models	BASIC	ASPP-S	ASPP-C	DM	AFN-1	AFN-6
Params	315,725	967,330	478,435	662,555	349,904	450,209

Table 4. Dice scores achieved by state-of-the-art methods on BRATS'15 test database. † are semi-automatic. * used CNN ensembles and more extensive augmentation.

Models	AFN-6	peres1[†][*][14]	bakas1[†][2]	kamnk1/DM [9]	kayab1* [10]	isenf1* [7]
Whole	84%	83%	81%	84%	85%	85%
Core	69%	72%	63%	67%	72%	74%
Enh	63%	60%	58%	63%	61%	64%
Cases	110/110	53/110	53/110	110/110	110/110	110/110

this is due to randomness in training and suboptimal optimization. (d) Empowering the high-resolution pathway of DeepMedic with adaptive autofocus quickly outperforms the gains from the static second pathway on pelvic scan and brain tumor segmentation except for the enhancing tumor. We speculate that gains are more profound in the former task due to the greater variation in the size of structures, where the adaptive nature of autofocus shines. Finally we note that by sharing weights across scales, AFNETs have small number of trainable parameters, shown in Table 3, which could enable rapid learning from little data, which is however left for future work. On the downside, the multiple scales on each autofocus layer increase memory and computation requirements.

Comparison with State-of-the-Art on BRATS'15: Performance on test data of BRATS'15 obtained via the online evaluation platform is shown on Table 4, along with other top published methods. AFN-6 compares favorably to the semi-automatic methods that topped the BRATS'15 challenge [2,14], as well as DeepMedic with the second static lower-resolution pathway. Note that in [14] high and low grade gliomas were separated by visual inspection and then passed to an appropriately specialized CNN, giving them an advantage over other methods. Our model is only surpassed by the pipelines of [7,10], who both used ensembles of CNNs with deep supervision and more aggressive data augmentation. The promising performance obtained by our simple method indicates the potential of the autofocus layer, which can be adopted in more elaborate systems.

4   Conclusion

We proposed an autofocus convolutional layer for segmentation of biomedical images. An autofocus layer can adapt the network's receptive field at different spatial locations in a data-driven manner. Our extensive evaluation of AFNETs shows that they cope well with biological variability in different tasks and generalize well on both MR and CT images. We have shown that the autofocus convolutional layer can be integrated into existing network architectures to substantially increase their representational power with only a small increase in model parameters. In addition, the interpretability of autofocus layers can leverage understanding of deep learning systems. Investigating the potential of autofocus modules in regression problems would be interesting future work.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR (2015)
2. Bakas, S., et al.: GLISTRboost: combining multimodal MRI segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. In: Crimi, A., Menze, B., Maier, O., Reyes, M., Handels, H. (eds.) BrainLes 2015. LNCS, vol. 9556, pp. 144–155. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30858-6_13
3. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv preprint arXiv:1606.00915 (2016)
4. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: CVPR (2016)
5. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE Trans. PAMI **35**(8), 1915–1929 (2013)
6. Galleguillos, C., Belongie, S.: Context based object categorization: a critical survey. Comput. Vis. Image Underst. **114**(6), 712–722 (2010)
7. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: Brain tumor segmentation and radiomics survival prediction: contribution to the BRATS 2017 challenge. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 287–297. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_25
8. Kamnitsas, K., et al.: Deepmedic for brain tumor segmentation. In: Crimi, A., Menze, B., Maier, O., Reyes, M., Winzeck, S., Handels, H. (eds.) BrainLes 2016. LNCS, vol. 10154, pp. 138–149. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-55524-9_14
9. Kamnitsas, K., et al.: Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation. MedIA **36**, 61–78 (2017)
10. Kayalibay, B., Jensen, G., van der Smagt, P.: CNN-based segmentation of medical imaging data. arXiv preprint arXiv:1701.03056 (2017)
11. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 (2015)
12. Menze, B.H.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015)
13. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
14. Pereira, S., Pinto, A., Alves, V., Silva, C.A.: Brain tumor segmentation using convolutional neural networks in MRI images. IEEE TMI **35**, 1240–1251 (2016)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: ICLR (2016)