Medical Image Analysis 54 (2019) 1–9

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

OBELISK-Net: Fewer layers to solve 3D multi-organ segmentation with sparse deformable convolutions

Mattias P. Heinrich[a,*], Ozan Oktay[b], Nassim Bouteldja[a]

[a] Institute of Medical Informatics, University of Lübeck, Germany
[b] Biomedical Image Analysis Group, Imperial College London, UK

ARTICLE INFO

Article history:
Received 15 August 2018
Revised 10 January 2019
Accepted 12 February 2019
Available online 13 February 2019

Keywords:
Image segmentation
Deep learning
Sparse kernels
Deformable convolutions

ABSTRACT

Deep networks have set the state-of-the-art in most image analysis tasks by replacing handcrafted features with learned convolution filters within end-to-end trainable architectures. Still, the specifications of a convolutional network are subject to much manual design – the shape and size of the receptive field for convolutional operations is a very sensitive part that has to be tuned for different image analysis applications. 3D fully-convolutional multi-scale architectures with skip-connection that excel at semantic segmentation and landmark localisation have huge memory requirements and rely on large annotated datasets - an important limitation for wider adaptation in medical image analysis.

We propose a novel and effective method based on trainable 3D convolution kernels that learns both filter coefficients and spatial filter offsets in a continuous space based on the principle of differentiable image interpolation first introduced for spatial transformer network. A deep network that incorporates this *one binary extremely large and inflecting sparse kernel* (OBELISK) filter requires fewer trainable parameters and less memory while achieving high quality results compared to fully-convolutional U-Net architectures on two challenging 3D CT multi-organ segmentation tasks.

Extensive validation experiments indicate that the performance of sparse deformable convolutions is due to their ability to capture large spatial context with few expressive filter parameters and that network depth is not always necessary to learn complex shape and appearance features. A combination with conventional CNNs further improves the delineation of small organs with large shape variations and the fast inference time using flexible image sampling may offer new potential use cases for deep networks in computer-assisted, image-guided interventions.

© 2019 Published by Elsevier B.V.

1. Introduction and motivation

A series of recent research papers have demonstrated that convolutional encoder-decoder networks excel at object delineation tasks. This is very important for medical image understanding and analysis, since segmenting different organs, anatomies and pathologies lies at the core of computer-assistance for diagnosis and interventions (Litjens et al., 2017). While the performance of automated algorithms has rapidly and steadily increased since the advent of deep learning, there is limited knowledge of why certain architectural choices lead to empirically observed improvements. Furthermore, while feature engineering has been superseded by learned convolution filters, researchers are tending towards network engineering (Xie et al., 2017). This means, for a certain task it is often necessary to explore a large number of architectures and adapt the number of layers, their kernel sizes and dilation parameters as well as the residual connections or concatenations of different scales of feature maps. In this work, we aim to offer an alternative strategy for deep learning architectures, which can help to reduce these empirical choices, restrict the number of trainable parameters and provide the possibility to learn robust models from few labelled scans.

Empirical evidence has shown that increasing the number of convolutional neural network layers leads to improved performance on large-scale image classification with fine-grained object recognition of 1000 classes and feature invariance can be theoretically linked to network depth (Wiatowski and Bölcskei, 2018). Typical state-of-the-art network architectures that may exceed human-level perception on ImageNet classification rely on more than a hundred of convolutional layers to learn feature and object part hierarchies, which leads to substantial demand of memory and compute resources already for a global 2D image classification. The ResNet-152 (He et al., 2016), for example, has more than 50 million trainable parameters, has to be trained on a GPU array for more

* Corresponding author.
E-mail address: heinrich@imi.uni-luebeck.de (M.P. Heinrich).

https://doi.org/10.1016/j.media.2019.02.006
1361-8415/© 2019 Published by Elsevier B.V.

than a week and requires more than 20 billion floating point operations for a single classification. Adapting, such an architecture to perform **patch-wise segmentation** in a sliding window manner to 3D medical volumes would lead to very long inference times of at least 15 minutes (when considering a coarse grid of 72^3 voxels and only 2D slices) making its direct adaptation undesirable. To address computational issues of patch-based classification so called **fully-convolutional networks** (FCNs) have been proposed (Long et al., 2015; Ronneberger et al., 2015) that use an intrinsic multi-scale approach within an encoder-decoder architecture and residual or skip connections to obtain a good trade-off between accuracy and computational demand, while still relying on dozens of convolutional layers. However, we note that most object and medical segmentation tasks deal with fewer than a dozen classes of anatomy, which raises the question whether very deep networks with many spatial convolution layers are indeed required to obtain high-quality results.

In the context of the limited amount of available medical data for training convolutional architectures, solutions that restrict the number of free parameters will also play an important role to extend the use of deep learning to new medical domains. In order to capture enough spatial context without increasing the number of trainable parameters too much, dilated convolutions have been proposed for classification and segmentation (Yu and Koltun, 2015; Yu et al., 2017), which introduce sparsity into kernels. This all leads to another open issue of recent attention that is to determine the best choice of the size, dilation, structure and number of convolution filters, which are required to accurately detect and delineate objects for a given medical image analysis task (cf. Gibson et al., 2018). Learnt kernels do not have to be regularly structured, this opens a new research direction that motivates our investigation of spatially-adaptive convolution filters.

In this work, we present an alternative concept to dilated convolutions in which both the **spatial filter offsets and coefficients** of a **large and sparse convolutional kernel** are learned in a continuous, differentiable space. We strongly believe that these spatially deformable kernels that automatically adapt their filter layout are an extremely important clue to deepen the understanding of the processes within convolutional networks and improve the applicability to medical volumetric data. By learning the spatial filter offsets, the network can decide on its own how sparse and how large the receptive field has to be for a given task. In our initial experiments with the OBELISK (one binary extremely large and inflecting sparse kernel) approach in (Heinrich et al., 2018), we could already demonstrate that a filter with a very narrow random initialisation that is in the range of few pixels, can quickly enlarge its spatial spread to **capture the relevant regional contextual information** in a completely data-driven manner. It matches the performance for multi-organ segmentation of carefully designed multi-scale architectures, e.g. (Milletari et al., 2016; Ronneberger et al., 2015), using only a single large kernel with fewer trainable weights followed by 1×1 convolutions (Lin et al., 2013) as classifier. Since the classification output can be computed very efficiently and independently for any spatial location, our approach offers an alternative to patch-based and fully-convolutional approaches.

This paper is outlined as follows. In the next section, we review selected related work in the context of semantic segmentation of medical scans, including multi-atlas label fusion and deep learning approaches, discuss the importance of spatial context and highlight our novel contributions to the field of deep learning by exploring deformable convolution kernels. In Section 3, we explain the concept of trainable spatial filter layouts in detail and extend our single kernel approach to multi-layer, multi-channel deformable architectures. In formulating a modular version of OBELISK that can be applied with a regular grid sampling, we enable new hybrid architectures that can benefit from the complementary strengths of fully-convolutional and deformable filters. The different methods are evaluated using experimental validation on two different 3D CT multi-organ segmentation tasks in Section 4.

2. Related work

To compute a semantic segmentation using labelled training data different concepts have been proposed in the past. Statistical shape models aim to separate the task into first representing deformable objects through a linear statistical model of surface point distributions and second finding the optimal placement of edges based on the local feature appearance (Heimann and Meinzer, 2009). Multi-atlas label fusion (MALF) methods (Rohlfing et al., 2004) use a deformable registration algorithm, that may itself consist of various strategies for feature extraction, transformation model and optimisation, to estimate one-to-one alignment of all training images and the unseen test scan to propagate segmentation information. Registration errors are unavoidable, but when using label fusion to locally select the most trustworthy transformed labels based on the alignment quality of the grayvalue scans (cf. Wang et al., 2013) very high accuracies can be reached, in particular when using discrete optimisation (Xu et al., 2016), that have sometimes even outperformed deep learning approaches (Heinrich, 2015; Heinrich and Oster, 2017).

Fully-convolutional prediction: Deep learning has gained much interest for image segmentation, starting with patch-based classification, which uses a common feed-forward architecture to predict a single class label for a pixel at the centre of the input patch. A large receptive field is necessary to capture enough spatial context, which is typically realised through either strided convolutions (Springenberg et al., 2015) or pooling layers. Fully convolutional, encoder-decoder architectures (FCNs) were subsequently proposed to address the immense computational demand of patch-based classification and reduce the amount of redundant computations (Long et al., 2015), which, however, led to several problems, including: reduced batch variability; the need for extra layers to recover lost details; and very large memory demand.

FCNs are only computationally efficient when trained with large image patches (or the whole image) in parallel, which means the variability in each batch is severely reduced. While this may not be a severe problem in computer vision where thousands of labelled images are used for training, it becomes problematic in the medical context, where usually only dozens of labelled 3D scans are available. The training with stochastic gradient descent is therefore slower and may require substantial data augmentation when using these large nearly uniform batches of pixels belonging to the same image. Due to the downsampling in the encoder of an FCN, there is a severe loss of local detail for dense prediction tasks, i.e. semantic segmentation or localisation, and upsampling or fractionally strided layers in conjunction with skip-connections have to be employed to increase the resolution and restore details. An alternative strategy to increase the receptive field without losing spatial detail are dilated or atrous convolutions (Yu and Koltun, 2015; Wolterink et al., 2016; Chen et al., 2018) that introduce zeros into kernels (while keeping the regular grid of a conventional filter kernel) in order to design larger kernels without increasing the number of free parameters. This way the stride of feature maps can remain as small as possible at the cost of larger memory demand.

Large Spatial Context: The key to understanding the requirement of numerous convolution layers for semantic segmentation lies in the limited ability of small kernels to capture enough spatial context. Due to local ambiguities it is impossible to classify a voxel based on its immediate neighbourhood. Commonly used convolutional kernels have a user-defined layout of sampling locations that is usually restricted to a regular 3×3 (or $3 \times 3 \times 3$) grid and

only filter coefficients are automatically learned. In order to capture enough regional context, several of these small 3×3 kernels are concatenated to increase the receptive field of each voxel and further supplemented by strided pooling operations that reduce resolution. Recent works on the *valid* (Zhou et al., 2014) or *effective receptive field* (Luo et al., 2016) and spatially sparse convolutions (Graham, 2014) have highlighted the fact that this straightforward filter concatenation does not necessarily lead to equally informative large contextual information aggregation for every part of the image or each object.

Earlier work in face recognition and person re-identification (Taigman et al., 2014) as well as more recent research in semantic segmentation (Peng et al., 2017) have promoted the use of **globally connected** layers (also called locally fully-connected without weight sharing) to capture the entire context of the whole image and alleviate the negative impact of effectively small receptive field sizes. (Peng et al., 2017) propose to use a separable convolution to address the imminent challenge of an exploding number of parameters. However, they still require multiple scales and network paths as well as additional boundary refinement modules to improve on state-of-the-art in semantic segmentation of natural images.

Training of sparse kernels: An alternative strategy to increase global context that will be developed in our work and does not resort to huge fully-connected layers focusses on learning the right degree of sparsity. Sparse spatial feature extractors have played an important role in computer vision for years, e.g. in descriptors such as DAISY (Tola et al., 2010) and BRIEF (Calonder et al., 2010). Similarly, random box-offset features have been succesfully used in depth image segmentation (Shotton et al., 2011) and localisation in medical scans (Criminisi et al., 2009; Heinrich and Blendowski, 2016). In these cases, the spatial layout of pixels for which intensities were obtained to derive feature representations were randomly drawn and only their combination in a classifier was later trained by supervision using random forests. The first uses of sparse features with large spatial (but still randomly chosen) offsets within filter kernels of end-to-end deep CNN architectures were proposed as local binary convolutions in Juefei-Xu et al. (2017) for 2D image classification and for 3D medical image segmentation as BRIEFnet in Heinrich and Oktay (2017). The concept was further extended to trainable deformable convolutions kernels in Dai et al. (2017), which train an additional network module that predicts a class-dependent nonlinear transformation of filter kernels to improve classification and segmentation accuracy by e.g. trying to adapt the kernel to the size of an object. This approach heavily depends on the concept of differentiable image interpolation that was first used in deep learning for spatial transformer networks (Jaderberg et al., 2015).

Contributions

In our conference submission at MIDL 2018 (Heinrich et al., 2018), we first introduced the concept of learnable sparse 3D kernels that can capture a very large receptive field, replace several classical grid-based filter layers and reduce the number of trainable parameters to avoid overfitting for small datasets. The details of this *one binary extremely large and inflecting sparse kernel* (OBELISK) approach will be briefly reviewed in Section 3. Despite its advantages: including fast convergence on small datasets, low-memory consumption and ease of the design of deep convolutional networks for semantic segmentation, our initial work had some limitations. First, only a single OBELISK convolution was employed followed by a few 1×1 filters, which may limit the abstraction of features through the depth of a network. Second, the spatial sampling locations were randomly and sparsely chosen across the image domain, which prevents the use of subsequent grid-based convolution kernels. Third, the initial network made no use of the popular encoder-decoder architecture that can learn both spatial context aggregation for object recognition as well as a structured spatial prediction of an output segmentation. In this work, we address these limitations and provide additional experimental evidence that supports the benefits of learning spatial filter offsets for convolutions.

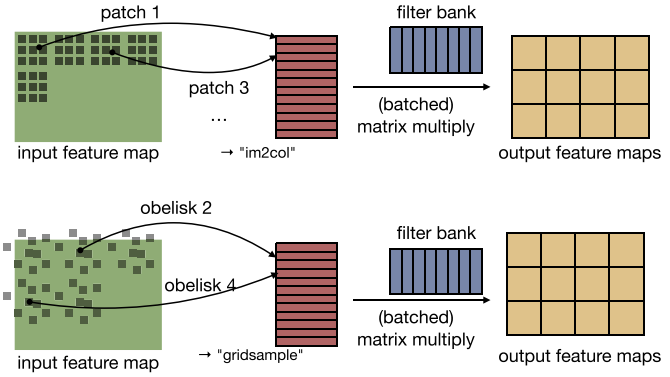

Fig. 1. Implementation of conventional convolutions in deep networks using the `im2col` operator to extract overlapping patches followed by a matrix multiplication with a filter bank and reshaping to the expected feature map dimensions. The deformable convolutions in OBELISK follow a similar principle, but replace the rectangular patch-extraction with a continuously sampled spatial filter offset layout (here 9 2D coordinates), which is added to the feature map coordinates (here a 3×5 grid). When performing bilinear interpolation using the `gridsample` operator a 15×9 matrix is computed. The subsequent hardware-optimised single-precision matrix multiplication of a filter bank is equivalently effective in terms of computation time, but may capture much more spatial context within a single layer and with few trainable parameters.

3. Method

In this section, we will explain the concept of sparse deformable convolutions based on differentiable image sampling, its relations to the `im2col` operator, our new extensions to multi-layer and hybrid architectures and the benefits and perspectives for 3D medical image analysis.

3.1. OBELISK: One binary extremely large and inflecting sparse kernel

The key to our novel convolutional architecture is a large layer that learns both spatial filter offsets and coefficients automatically from the given data. By inflecting (spatially adapting) the local offsets directly in a data-driven way, the manual design of convolutional architectures can be omitted and the best setting for the given problem can be automatically learned. While this is inspired by spatial transformer networks (Jaderberg et al., 2015) and deformable convolutions (Dai et al., 2017), our approach is more general in that the learned kernel is not dependent on a separately estimated class or object geometry prediction. Instead, we aim to learn a generic kernel that is applicable without spatial transformation and can replace multiple small filter kernels at different scales.

An intuitive way of understanding deformable convolutions is to consider the two step approach that is used in the implementation of conventional convolution kernels for benefiting from highly efficient CUDA code for batched matrix multiplications. Fig. 1 demonstrates the principle of a standard convolution layer that relies on the `im2col` operator to extract overlapping rectangular patches. Considering an analogous interpretation of our proposed OBELISK kernel in comparison, it becomes easy to see that only this regular patch extraction step is replaced using the `gridsample` operator. This change enables the use of continuous valued and therefore trainable spatial filter offsets that are

added to the feature map coordinates and cause little additional computational overhead (see Section 5 for experiments with training times).

Our learned kernel can be very sparse and substantially increase the receptive field and spatial context aggregation, an important aspect for medical image analysis as shown in previous works (Criminisi et al., 2010; Heinrich and Oktay, 2017). In turn, it enables us to train a network with as few as 130′000 free parameters that achieves remarkably accurate predictions, uses very little memory ($<$700 MBytes for our baseline model) and is very fast to train.

Differentiable image sampling: Consider a classical 2D convolution operation for a kernel with 25 elements (forming a 5×5 filter) and dilation factor of d (Wolterink et al., 2016; Yu and Koltun, 2015). The spatial filter offsets are statically defined as $(s_x, s_y) = \{-2d, d, 0, +d, +2d\}^2 \in \mathbb{Z}^{5\times5}$. Let $I(x, y)$ be the value of an input at location (x, y) and $W \in \mathbb{R}^{5\times5}$ the continuous valued and trainable filter coefficients. The output $F(x, y)$ can be calculated as:

$$F(x,y) = \sum_i \sum_j W(i,j) \cdot I(x + s_x(i,j), y + s_y(i,j)) \tag{1}$$

Since, both the pointwise multiplication and the sum operation are differentiable, we can easily find the derivate of a convolution operation with respect to the weights W and the input I. Let us now consider the continuous valued spatial filter offsets $S_x \in \mathbb{R}^{5\times5}$ and $S_y \in \mathbb{R}^{5\times5}$. To obtain the convolution output for inputs on a discrete grid, we need to perform bilinear interpolation:

$$\begin{aligned} F(x,y) = \sum_i \sum_j W(i,j) \cdot (& w_1 I(\lfloor x + S_x(i,j) \rfloor, \lfloor y + S_y(i,j) \rfloor) \\ + w_2 I(\lceil x + S_x(i,j) \rceil, & \lfloor y + S_y(i,j) \rfloor) + w_3 I(\lfloor x + S_x(i,j) \rfloor, \\ & \lceil y + S_y(i,j) \rceil) + w_4 I(\lceil x + S_x(i,j) \rceil, \lceil y + S_y(i,j) \rceil)) \end{aligned} \tag{2}$$

with the following bilinear coefficients $w_1, \ldots, w_4$:

$$\begin{aligned} w_{xu} &= (\lfloor x + S_x(i,j) + 1 \rfloor - (x + S_x(i,j))) \\ w_{yu} &= (\lfloor y + S_y(i,j) + 1 \rfloor - (y + S_y(i,j))) \\ w_{xd} &= (x + S_x(i,j) - \lfloor x + S_x(i,j) \rfloor) \\ w_{yd} &= (y + S_y(i,j) - \lfloor y + S_y(i,j) \rfloor) \\ w_1 &= w_{xu} \cdot w_{yu}; \; w_2 = w_{xd} \cdot w_{yu} \\ w_3 &= w_{xu} \cdot w_{yd}; \; w_4 = w_{xd} \cdot w_{yd} \end{aligned} \tag{3}$$

Again all operations (multiplications, min/max for floor/ceil, and additions) are differentiable. We can therefore obtain the derivatives with respect to the filter coefficients W, their spatial filter offsets S_x, S_y and the input if necessary. We employ our approach mainly for 3D applications and the extensions of Eqs. (2) and (3) to trilinear interpolation using 8 positions and interpolation weights is straightforward. We refer to learning one spatial 3D offset per filter coefficient as *unary variant* and propose a *binary* extension that learns two offsets for each filter coefficient. In this case the two interpolated values from the preceding layer (the image input) are first subtracted and the outcome is multiplied with their joint filter coefficient. Pairing and subtracting two values within the receptive field is motivated by the success of so-called pixel or box difference features used in many vision tasks (Criminisi et al., 2010; Heinrich and Oktay, 2017). The advantage of such sampling layouts have been investigated extensively in Calonder et al. (2010). Note, that many filters motivated by the visual system comprise positive and negative dipole like layouts. Employing pairwise subtractions may also reduce the sensitivity of a model on absolute intensity values in the training data and it has been shown to be a useful approach to achieve better model generalisation as in the case of random forest models.

Implementation details: The spatial filter offsets S_x, S_y, S_z are initialised with normally distributed random numbers and zero mean just as their filter coefficients are. Fig. 2 (left) shows an example of the spatial distribution of spatial filter offsets. Similar to previous work (e.g. Yu and Koltun, 2015), we empirically found that for very large receptive fields a sufficient throughput of local information is necessary. Therefore, smaller values for the standard deviation σ of the normal distribution are preferable. We used $\sigma = 0.05$ – the image coordinates range from -1 to +1, but $\sigma = 0.02$ to $\sigma = 0.1$ gave almost identical results. The network will automatically learn to increase the receptive field if necessary to as much as half of the image domain. We use zero-padding for the `gridsample` operation, so that no gradient occurs for samples outside the image domain, which implicitly helps at avoiding infeasible offsets.

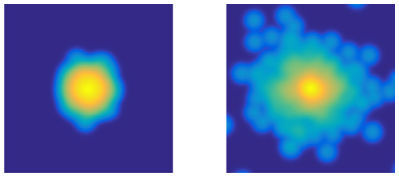

Fig. 2. Example of spatial distribution of spatial filter offsets (S_x, S_y are shown, S_z omitted) – shown with logarithmic colormap – at the start of the training of the OBELISK layer (left) and after optimisation for 50 epochs (right). The size of the extent of the figures corresponds to half of the image domain. It is evident that our data-driven approach yielded a larger receptive field, but still contains many filter coefficients close to the centre.

Subsequently, one could completely avoid any subsequent spatial convolutions (following OBELISK) as initially proposed in Heinrich et al. (2018) and simply use 1×1 convolutions (Lin et al., 2013) (with batch norm and ReLU) that form a multi-layer perceptron shared among all locations for dense pixelwise predictions.

3.2. Integrating OBELISK layers into conventional networks

In order to integrate deformable 3D convolutions into fully-convolutional or multi-layer networks, the OBELISK filter has to be applied using a regular (but potential coarse) spatial sampling grid as shown in the example of Fig 1. We can thus implement OBELISK as a modular layer that performs an efficient aggregation of spatial context, a concept that can be readily integrated into commonly used U-Net or V-Net architectures and gain from complementary advantages of pixel-wise convolution operators. The original OBELISK layer required a single-channel 3D input image (the CT scan smoothed with two small kernels) and used a very large number of trainable offsets to capture enough information for the subsequent classification layers. For improved generality, we will now describe OBELISK for multichannel inputs (for ease of description explained for 2D) and define a more general deformable convolution operation as follows. Given a dense input tensor (feature maps) of size $B \times C_{in} \times H \times W$ with a spatial sampling coordinate tensor of size $B \times 1 \times S_{sp} \times 2$, a deformable offset tensor of size $1 \times K \times 1 \times 2$ and a weight tensor of size $1 \times C_{out} \times C_{in} \cdot K$ are required. Here, B is the batch size, C_{in} the number of input channels, C_{out} the number of output channels, H and W the spatial dimensions of the input feature map, S_{sp} the number of spatial output locations, K the number of kernel elements.

When using OBELISK layers within a fully-convolutional architecture the spatial sampling has to be chosen so that S_{sp} equals a regular grid for further processing. Using the `gridsample` operation the input is sampled to a new size of $B \times C_{in} \times K \times S_{sp}$ similar to the process of `im2col` for classical convolutions (see also Fig. 1). The implementation of all network variants and our training routines are publicly available at https://github.com/mattiaspaul/OBELISK.

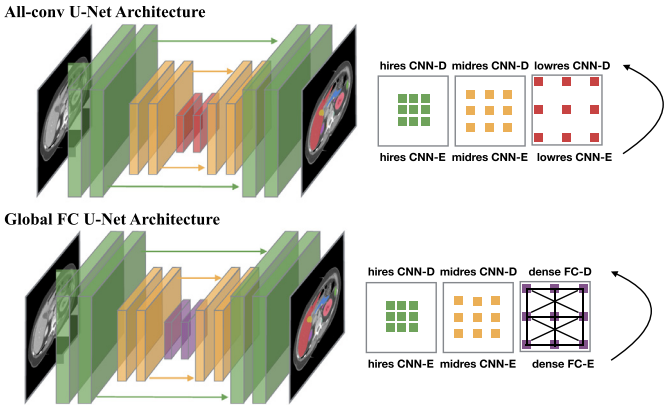

Fig. 3. Visual comparison of classical fully-convolutional multi-scale U-Net architecture and its globally fully-connected variant (not all layers used in our experiments are shown). The densely connected layers in the bottleneck have a significant amount of trainable parameters, but enable us to capture global context.

Table 1
Description of baseline U-Net model (**all-conv 880k**). Spatial dimensions are given for an exemplary input size of 144^3 voxels. All layers are 3D convolutions, followed by batch normalisation and leaky ReLU activation. Outgoing and incoming skip connections are noted in the last column. #L defines the number of output classes. Upsampling is performed using trilinear interpolation.

Layer	(Out)-Size	Kernel/Stride	# Channels	Skip
Input	144^3		1	
#1	144^3	$3 \times 3 \times 3$	$1 \rightarrow 5$	$\rightarrow$ #13
#2	72^3	$3 \times 3 \times 3$ /2	$5 \rightarrow 16$	
#3	72^3	$3 \times 3 \times 3$	$16 \rightarrow 16$	$\rightarrow$ #12
#4	36^3	$3 \times 3 \times 3$ /2	$16 \rightarrow 32$	
#5	36^3	$3 \times 3 \times 3$	$32 \rightarrow 32$	$\rightarrow$ #11
#6	18^3	$3 \times 3 \times 3$ /2	$32 \rightarrow 64$	
#7	18^3	$3 \times 3 \times 3$	$64 \rightarrow 64$	$\rightarrow$ #10
#8	9^3	$3 \times 3 \times 3$ /2	$80 \rightarrow 80$	
#9	9^3	$3 \times 3 \times 3$	$80 \rightarrow 80$	
	18^3	Upsample	80	
#10	18^3	$3 \times 3 \times 3$	$144 \rightarrow 64$	#7 $\rightarrow$
	36^3	Upsample	64	
#11	36^3	$3 \times 3 \times 3$	$96 \rightarrow 32$	#5 $\rightarrow$
	72^3	Upsample	32	
#12	72^3	$3 \times 3 \times 3$	$48 \rightarrow 9$	#3 $\rightarrow$
	144^3	Upsample	14	
#13	144^3	$3 \times 3 \times 3$	$14 \rightarrow$ #L	#1 $\rightarrow$
#14	144^3	$3 \times 3 \times 3$	#L $\rightarrow$ #L	

3.3. Network architectures

We aim to gain a better understanding into the working principle of deformable convolution kernels that can learn spatial filter offsets. We therefore conducted a comparison between sparse deformable (OBELISK), fully-convolutional and globally fully-connected architectures (U-Net), which are described in the following.

1. All-Conv U-Net: As baseline architecture, we adapt an all-convolutional (without pooling layers, but strided convolutions) 3D U-Net encoder (see Fig. 3) similarly parameterised as in Gibson et al. (2018), which contains 9 convolutional layers in the contracting path and 5 in the expanding path with in total approx. 880k trainable parameters (580k in the encoder part and 300k in the decoder). Details are listed in Table 1. Later, we also include a **leaner all-conv U-Net** version with only a quarter of trainable parameters. This is achieved by using grouped convolutions in the middle of the U with group size of 4 for layers #6, #7 and #11 and size 8 for layers #8 - #10.

2. Global FC U-Net: We discussed in Section 2 that much relevant global context can best be captured using large globally connected layers, which is the key motivation behind the sparse deformable kernels of OBELISK. In order to reflect this aspect in a U-Net encoder decoder architecture, we also experimented with a global FC U-Net variant that contains two fully-connected layers in the bottleneck (middle of U) (see Fig. 3) to encode a feature map of $9^3 \times 20$ to a 400-dimensional vector and then decode this back to a spatial feature map afterwards. These two layers alone contain 11.7 million parameters, but are able to capture rich shape information.

Table 2
Description of our **original OBELISK** model with 1×1-Dense-Net. The number of sampling positions S_{sp} can either be a coarse and regular 3D grid (for inference) or simply a small random subset (usually 512 voxels) for training. Layers #3 - #6 define the compact 1×1 dense-net classifier.

Layer	(Out) -Size	Kernel /Stride	# Channels /Groups	Skip
Input	144^3		1	
Avg	144^3	$3 \times 3 \times 3$	1	
Avg	144^3	$3 \times 3 \times 3$	1	
offsets		1024×3		
	S_{sp}	`gridsample`		$1 \rightarrow 1024$
#1	S_{sp}	$1 \times 1 \times 1$	$1024 \rightarrow 256$ /4	
#2	S_{sp}	$1 \times 1 \times 1$	$256 \rightarrow 128$	$\rightarrow$ #3-#6
#3	S_{sp}	$1 \times 1 \times 1$	$128 \rightarrow 32$	$\rightarrow$ #4-#6
#4	S_{sp}	$1 \times 1 \times 1$	$160 \rightarrow 32$	$\rightarrow$ #5-#6
#5	S_{sp}	$1 \times 1 \times 1$	$192 \rightarrow 32$	$\rightarrow$ #6
#6	S_{sp}	$1 \times 1 \times 1$	$224 \rightarrow 32$	
#7	S_{sp}	$1 \times 1 \times 1$	$32 \rightarrow$ #L	

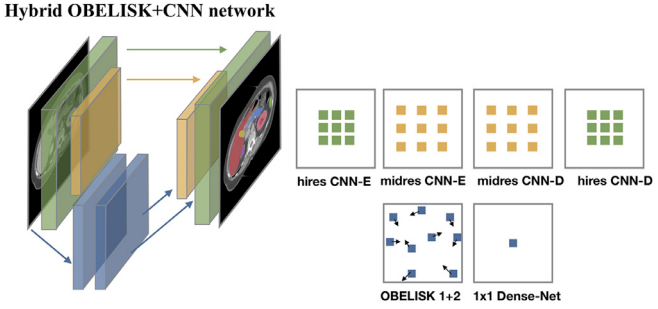

Fig. 4. Visual illustration of multi-layer Hybrid OBELISK-CNN architecture. Two High- and mid-resolution conventional CNN layers are followed by two OBELISK layers with spatially deformable kernels that use a regular coarse sample grid (in contrast to the original OBELISK network) for the resulting feature map and again only 1×1 filters that act as multi-layer perceptron for feature abstraction and classification. Each voxel can be computed independently and much fewer parameters are required that are shared across sampling locations (translational invariance). The spatial filter offsets of these kernels are continuously defined and end-to-end trainable.

3. Original OBELISK: Our baseline OBELISK model (**OBELISK+1×1-Dense**) of Heinrich et al. (2018) in contrast contains only a single large spatial convolution kernel that is followed by a 1×1 dense-net classifier (Huang et al., 2017) (see Table 2). The spatial sampling size does not have to lie on a regular grid and can be randomly sampled, unless we extend the network to contain multiple convolution kernels. Starting from a normally distributed random distribution with $\sigma = 0.05$, we learn 1024 paired spatial filter offsets (*binary variant*) and a filter bank with 256 output channels and another linear layer to reduce this to 128 (approx. 104k parameters). The input images are smoothed with two $3 \times 3 \times 3$ average pooling layers. The output of the OBELISK layer is processed with a 5 layer dense-net with approx. 26k parameters. No classical convolution is used.

4. Hybrid OBELISK+CNN: Finally, we also explore a hybrid architecture that combines the advantages of OBELISK with the edge-preserving filters learned by a shallow U-Net (see Fig. 4). For this network, we use two OBELISK layers – one with 512 spatial filter

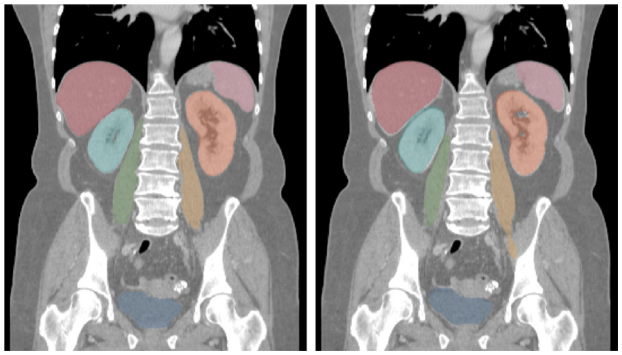

Fig. 5. Comparison of segmentation overlay for seven anatomical structures of VISCERAL dataset: liver, spleen, bladder, left kidney, right kidney, left psoas major muscle (pmm) and right pmm. Left: Coronal plane of original CT scan with ground truth segmentation. Right: Our originally proposed hybrid OBELISK network (with data augmentation). A clear segmentation of the urinary bladder and detailed delineation of psoas muscles as well as a clear good differentiation between liver and kidney are visible for our approach.

offsets and a very coarse regular sampling grid (spacing of eight w.r.t. the size of the input) and another one with 128 spatial filter offsets and a coarse grid (quarter size of each input dimension) – both followed by a fully-convolutional 1×1 dense-net. We resorted to the simpler (*unary variant*) where no offsets are paired or subtracted. These layers contain less than 50k parameters and each use a smoothed version of the original scans as input. Their output has 32 and 16 channels respectively and is concatenated within a shallow U-Net that follows the baseline U-Net model in Table 1, but with the most parameter-intensive layers #6 - #10 removed, so that only around 140k trainable parameters remain.

4. Experiments

We performed extensive experiments on two 3D CT datasets with different challenges. First, we explore the difficulties of a limited sized training dataset (10 VISCERAL training scans (Jimenez-del Toro et al., 2016)) with a set of moderately challenging abdominal structures as done in our conference paper (Heinrich et al., 2018) without specific preprocessing of the scans. Second, we perform additional validation experiments of different network architectures on a public multi-label dataset that is based on the TCIA CT pancreas dataset (with 43 scans) introduced in Roth et al. (2015), but has been extended by more manually labelled organs and uses the tightly cropped region of interests as done in Gibson et al. (2018). The TCIA data includes the labels: spleen (), pancreas (), kidney (), gallbladder (), esophagus (), liver (), stomach () and duodenum (), some of which are very challenging due to their small size, variable shape and difficult image contrast. In these new experiments, we want to find out whether our concept of using very shallow networks with only one (or very few) trainable convolutional layers is still viable for this more difficult semantic segmentation task.

The VISCERAL experiment is carried out with leave-one-out cross validation for the segmentation the following seven anatomical structures: liver (), spleen (), bladder (), left kidney (), right kidney (), left psoas major muscle (pmm,) and right pmm () (see Fig. 5). The only pre-processing applied was a resampling to isotropic voxel sizes of $3\,\mathrm{mm}^3$ and cropping to dimensions of $156 \times 115 \times 160$ voxels, without using any guidance information. In principle the OBELISK layer could learn to deal with different and anisotropic voxel spacings, but this has not been considered in our experiments. Note that such a rough cropping poses a much harder challenge than two-stage approaches (Roth et al., 2015; Zhou et al., 2017), regional CNNs (Larsson et al., 2017). For the original OBELISK approach the memory requirements are very low (roughly 700 MByte) and independent of image dimensions. All network layers are trained with the same learning rate of 0.002 (using Adam) for 300 epochs and a mini-batch size of 1 or 3 for the hybrid OBELISK and all U-Net approaches respectively. For the original OBELISK architecture, we trained with a batch size of 3 images with 192 spatial samples each, and 64 iterations (of random mini-batches) each for 50 epochs. Online hard example mining (Shrivastava et al., 2016) was used with quantiles of 75% for networks with OBELISK layers yet its influence on the final outcome was marginal when using data augmentation. The U-Nets were trained with fully-convolutional batches of 3 images each and for each conventional convolution a small leakage parameter for ReLUs was used.

Table 3
The quantitative evaluation of a leave-one-out cross-validation for 10 scans of the VISCERAL dataset demonstrates the high-quality in terms of Dice accuracy and reduction of parameters (ten-fold) of our proposed OBELISK approach compared to the state-of-the-art U-Net architecture. Incorporating The memory requirement for our originally proposed OBELISK network is ≈700 MByte, while a U-Net requires at least 2′500 MByte. ++ defines affine augmentation during training (in OBELISK this can be realised using a simple matrix multiplication of offsets without any additional image manipulation). All OBELISK variants employ online hard example mining, which simply back-propagates only the top 1/4 fraction of individual loss terms during training.

Method	#params.	Dice
Unary OBELISK + 1×1-Dense	126k	72.32%
Rand. Offsets + 1×1-Dense	124k	71.67%
Original OBELISK	130k	76.68%
Original OBELISK ++	130k	80.61%
Hybrid-OBELISK+CNN ++	230k	**82.27%**
all-conv U-Net	880k	72.84%
all-conv U-Net ++	880k	79.95%
global-FCN U-Net	13′000k	72.85%
global-FCN U-Net ++	13′000k	**81.67%**

In our new TCIA experiments we applied the same preprocessing as proposed by Gibson et al. (2018), which involves a tight manual bounding box cropping and resampling to common dimension of 144^3 voxels (which results in irregular voxel spacings). It was discussed in Gibson et al. (2018) that the manual bounding box selection was an important aspect and it was argued that single-pass feed-forward U-Nets without attention gates (Oktay et al., 2018) are dependent on accurate initialisation. Due to the larger database size of 43 scans, we employ a four-fold cross-validation with either 32 or 33 training images and 10 or 11 test scans. The training of each network was performed for 300 epochs and took around 90 minutes on a GTX 1080 Ti per model. We experimented with different schemes for dealing with class imbalance including the conventional Dice loss (Milletari et al., 2016) and a hinge Dice loss (Gibson et al., 2018), but found the differences to be negligible to a classical cross-entropy loss after 300 epochs for all U-Net architectures. For OBELISK we employ online hard example mining as before together with a cross-entropy loss weighted by the square root of the inverse label frequency (the weights for foreground are averaged to 1 and the background is weighted with 0.5). Randomly applied affine image transformations are used throughout as the only type of data augmentation unless otherwise mentioned.

5. Results and discussion

Table 3 lists the numerical evaluation that shows already remarkably good results of a single layer OBELISK network with

Table 4
Quantitative evaluation of four-fold cross-validation for 43 scans of the TCIA multi-label dataset. All models are trained for 300 epochs with affine data augmentation and evaluated without post-processing on the resampled scans with dimensions of 144^3 voxels. Individual label accuracies are listed for the eight foreground labels with Dice overlap scores according to the colour scheme described above.

Method #params and avg. Dice score							
■	■	■	■	■	■	■	■
Original OBELISK (130k parameters)						avg. 74.72%	
92.3	61.7	92.1	66.8	59.2	95.0	82.9	47.8
Hybrid-OBELISK+CNN (230k parameters)						**avg. 79.03%**	
94.4	70.2	94.2	75.3	63.3	95.4	86.8	53.8
All-convolutional U-Net (880k parameters)						avg. 77.77%	
95.0	66.9	94.6	67.8	59.4	95.8	87.4	55.4
Leaner all-conv. U-Net (220k parameters)						avg. 74.33%	
93.8	61.6	93.3	64.4	51.2	95.2	85.2	49.9
Global FCN U-Net (16 million parameters)						**avg. 78.49%**	
94.5	68.8	94.3	73.5	59.9	95.6	87.0	54.3

a dense-net classifier without using augmentation of 76.7% Dice score (averaged over all 10 folds and 7 labels). Our preliminary experiments with the original OBELISK from (Heinrich et al., 2018) also demonstrate the benefits of learning offsets (compared to simply using random initialisation – "Rand. Offsets" – with a tuned standard deviation) and pairing two offsets to one coefficient with improvements of 4–5% points each. Considering our new experiments using a hybrid OBELISK architecture with CNNs a further gain of 1.5% is reached, despite using the simpler *unary* variant with only one 3D offset per filter coefficient. In general, we observed that at least around 500 spatial filter offsets have to be learned to obtain a high quality prediction, which is comparable in size to a either classical 8^3 kernel for scalar inputs or a 3^3 kernel with a 24-channel input feature map.

Comparing the performance of OBELISK to U-Net architectures, we observe substantial advantages of nearly 4% without data augmentation. These differences are slightly reduced when including affine transforms of the input images during training, yet, an advantage of 3.5% compared to a large all-convolutional U-Net remains and our hybrid OBELISK network performs still on par with a much bigger FCN U-Net that contains a fully-connected layer (82.27% and 81.67% respectively) - the leaner all-conv U-Net variant did not yield sufficiently accurate results for the VISCERAL data.

Results on TCIA43 database: When comparing the Dice results of the hybrid OBELISK network for anatomical structures that are common in both datasets, we see a clear correlation of training dataset size and accuracy for these larger organs. The scores for liver ■, spleen (■) and left kidney (■) are 95.39%; 94.35% and 94.21% for the large TCIA dataset and 90.6%, 79.5% and 85.7% for the VISCERAL data with only 10 scans, a difference of on average more than 9% points. Detailed quantitative results for TCIA43 are listed in Table 4. The inference time for either the all-convolutional U-Net, the global FCN U-Net and the hybrid OBELISK+CNN network on a GPU is 87 milliseconds (and mainly dominated by the high-resolution CNN layers). While both Hybrid OBELISK+CNN and global FCN U-Net perform similarly well on Dice accuracy (79.03% and 78.49% respectively), the best model capacity to accuracy trade-off (also with respect to a leaner U-Net model) is clearly reached by using our proposed deformable convolution layers. Looking at individual organ segmentation labels (please see Fig. 6 for visual results), OBELISK reaches the best scores for pancreas (■) of 70.2%, which is 3.4% better than the all-convolutional U-Net and around 4% improvement for the esophagus (■). A leaner U-Net that has a comparable parameter count (of 220k) to our hybrid OBELISK model performs 4.5% worse on average. Further improvements may be obtained by increasing the parameter count, depth and number of training epochs of all architectures and at the same time introducing dropout regularisation to prevent overfitting and enable an intrinsic model ensembling.

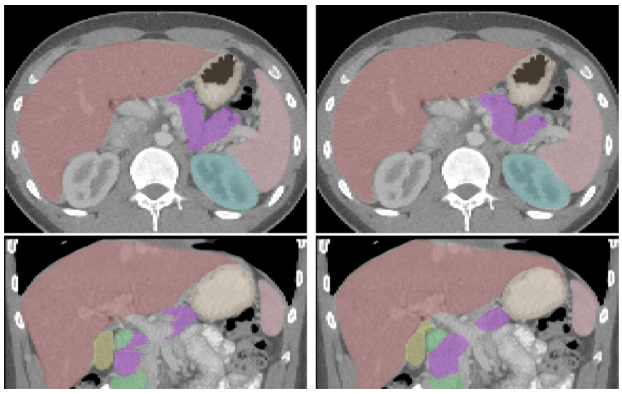

Fig. 6. Segmentation result for TCIA dataset shown as overlay for axial slice (top row) and coronal plane (bottom row). Left: the ground truth segmentation for liver (■), spleen (■), pancreas (■), stomach (■), gallbladder (■), duodenum (■), and kidney (■) are shown. The automatic segmentation of a multi-layer OBELISK network combined with a shallow U-Net and only 230k trainable parameters demonstrates highly accurate predictions of these challenging structures. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Comparison to state-of-the-art: Even though we have not used any post-processing such as edge-preserving smoothing (Heinrich and Blendowski, 2016), more aggressive data augmentation (including elastic deformations and contrast variations) or curvature flow smoothing (Gibson et al., 2018)), our new approach approaches state-of-the-art accuracies for two challenging 3D CT segmentation tasks. Despite using only 9 training images, we outperform keypoint-transfer (Wachinger et al., 2015) (Dice score for six labels of 78%) and multi-atlas label fusion (Dice score of 80% according to Wachinger et al. (2015)), on the VISCERAL dataset. Our approach reaches an average Dice of 83.7% for these six labels. However, the winner of the original challenge has recently published an improved performance of multi-scale multi-atlas registration together with graph-cut regularisation with a score of 88% (Kéchichian et al., 2018). For the TCIA multilabel dataset combined with another database of abdominal CTs from a recent MICCAI challenge, Gibson et al. (2018) reached an average (mean) Dice of 81.5% (2.5% better than our approach), but used a twice as large training dataset and additional post-processing. We furthermore note that one of the best deep network that was specifically tuned for pancreas segmentation reaches 71.8% Dice overlap for this challenging organ, which is very close to our Dice score of 70.2%. A more recent approach, however, achieved 81.3% by combining multiple steps and different networks for localisation and segmentations (Roth et al., 2018).

Efficiency considerations: As can be seen from the quantitative segmentation accuracies, an OBELISK layer with very few parameters can replace conventional fixed-grid kernels while often resulting in superior representation learning power, but might lead to more complex numerical operations due to irregular memory access patterns and extra computations for trilinear interpolation. We have made empirical comparisons that reveal this performance gap to be unexpectedly small. Given an input of 144^3 voxels with 8 feature channels, two strided convolutions with kernel sizes of $3 \times 3 \times 3$ and first 32 and then 64 output channels require approx. 10 ms for forward and backward pass (approx. 31 billion operations and 25% GPU efficiency), while an OBELISK with 128 learnable spatial filter offsets followed by an 1×1 convolution ($8 \cdot 128 \times 64$ channels) evaluated on the same dense spatial grid of 36^3 output locations takes twice as long (approx. 20 ms for about 10 billion operations including backpropagation). This is remark-

able given the fact that spatial convolutions are heavily optimised within the CuDNN framework. In addition, the OBELISK approach is not restricted to regular and fixed (or even integer) grids for spatial sampling locations and becomes much more efficient than conventional CNNs when considering a sparse sampling. When training a network where the prediction directly follows the sparse filters, dozens of small and sparse mini-batches with only few hundreds locations can be employed to speed up training times substantially. But even in our proposed hybrid network architecture the deformable convolutions require less than 10% (40 ms) of the time taken to pass a single volume forwards and backwards during training (total time 500 milliseconds) and the majority of computation time is spent for the spatial convolutions of the shallow U-Net.

6. Conclusion and outlook

We have presented a novel convolutional architecture called OBELISK that enables accurate dense predictions, demonstrated on the example of 3D multi-organ segmentation, using few extremely large and sparse filter kernels that not only learn filter coefficients but also spatial filter offsets. It aims at unifying aspects of previous multi-resolution, cascaded dilatation and deformable convolution approaches into a simple framework that consists of OBELISK layers and subsequent 1×1 convolutions as well as conventional convolutions. Our concept goes beyond a fixed pixel grid for filter coefficients in convolutions for extracting information from images. This will support future methods to deal more naturally with anisotropic highdimensional data, which is important in 3D medical imaging and also for temporal signals (2D+t). Furthermore, the use of sparse and adaptive sampling of spatial locations for predictions could significantly reduce computation times for inference in time-sensitive interventional medical tasks with computer-assisted image-guidance and for the coarse-to-fine localisation of small anatomies.

Our extensive experimental comparisons confirm that a single OBELISK layer can indeed “solve nearly everything” (as mentioned in the title of our original conference paper Heinrich et al., 2018), however, further improvements can be obtained by combining multiple deformable convolution layers together with conventional CNNs within the same network. We show that the flexibility of trainable spatial filter offsets is valuable in comparison to fixed dilation kernels and they may reduce the importance of skip-connections found in U-Net architectures for dense predictions. The positive impact of sparse sampling in training was smaller than originally expected and appears to only speed-up the initial learning rate.

In conclusion, we believe that the deformable convolutions of OBELISK are a promising and powerful drop-in feature that can be easily integrated into existing shallow U-Nets and provides huge potential for using it as a pre-trained low-parametric shape-encoder. It also offers benefits of reduced model complexity and increased speed during inference.

Conflicts of interest

We have no conflicts of interest to declare.

Acknowledgements

We would like to thank Nvidia for donating a Titan Xp GPU used in this research.

References

Calonder, M., Lepetit, V., Strecha, C., Fua, P., 2010. BRIEF: binary robust independent elementary features. In: European Conference on Computer Vision. Springer, pp. 778–792.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 40 (4), 834–848.

Criminisi, A., Shotton, J., Bucciarelli, S., 2009. Decision forests with long-range spatial context for organ localization in CT volumes. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 69–80.

Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E., 2010. Regression forests for efficient anatomy detection and localization in CT studies. In: International MICCAI Workshop on Medical Computer Vision. Springer, pp. 106–117.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 764–773.

Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018. Automatic multi-organ segmentation on abdominal CT with dense v-networks. IEEE Trans. Med. Imag. 37 (8), 1822–1834.

Graham, B., 2014. Spatially-sparse convolutional neural networks. arXiv:1409.6070.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

Heimann, T., Meinzer, H.-P., 2009. Statistical shape models for 3D medical image segmentation: a review. Med Image Anal 13 (4), 543–563.

Heinrich, M.P., 2015. Multi-organ segmentation using deeds, self-similarity context and joint fusion. MICCAI Challenge Workshop on Multiatlas Segmentation Beyond the Cranial Vault.

Heinrich, M.P., Blendowski, M., 2016. Multi-organ segmentation using vantage point forests and binary context features. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016. Springer, pp. 598–606.

Heinrich, M.P., Oktay, O., 2017. BRIEFnet: deep pancreas segmentation using binary sparse convolutions. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 329–337.

Heinrich, M.P., Oktay, O., Bouteldja, N., 2018. OBELISK - One kernel to solve nearly everything: unified 3D binary convolutions for image analysis. Medical Imaging with Deep Learning 1, 1–9.

Heinrich, M.P., Oster, J., 2017. MRI whole heart segmentation using discrete nonlinear registration and fast non-local fusion. In: International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, pp. 233–241.

Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L., 2017. Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1, p. 3.

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 2017–2025.

Juefei-Xu, F., Boddeti, V.N., Savvides, M., 2017. Local binary convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1.

Kéchichian, R., Valette, S., Desvignes, M., 2018. Automatic multiorgan segmentation via multiscale registration and graph cut. IEEE Trans Med Imaging 37 (12), 2739–2749.

Larsson, M., Zhang, Y., Kahl, F., 2017. Robust abdominal organ segmentation using regional convolutional neural networks. In: Scandinavian Conference on Image Analysis. Springer, pp. 41–52.

Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv:1312.4400.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440.

Luo, W., Li, Y., Urtasun, R., Zemel, R., 2016. Understanding the effective receptive field in deep convolutional neural networks. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 4898–4906.

Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Fourth International Conference on 3D Vision (3DV). IEEE, pp. 565–571.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al., 2018. Attention U-Net: learning where to look for the pancreas. arXiv:1804.03999.

Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J., 2017. Large kernel matters: improve semantic segmentation by global convolutional network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 1743–1751.

Rohlfing, T., Brandt, R., Menzel, R., Maurer, C., 2004. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 21 (4), 1428–1442.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 234–241.

Roth, H.R., Lu, L., Farag, A., Shin, H.-C., Liu, J., Turkbey, E.B., Summers, R.M., 2015. DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 556–564.

Roth, H.R., Lu, L., Lay, N., Harrison, A.P., Farag, A., Sohn, A., Summers, R.M. , 2018. Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation. Med. Image. Anal. 45, 94–107.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A., 2011. Real-time human pose recognition in parts from single depth images. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. Ieee, pp. 1297–1304.

Shrivastava, A., Gupta, A., Girshick, R., 2016. Training region-based object detectors with online hard example mining. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 761–769.

Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2015. Striving for simplicity: the all convolutional net. ICLR-2015 workshops.

Taigman, Y., Yang, M., Ranzato, M., Wolf, L., 2014. Deepface: closing the gap to human-level performance in face verification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1701–1708.

Tola, E., Lepetit, V., Fua, P., 2010. DAISY: an efficient dense descriptor applied to wide-baseline stereo. IEEE Trans. Pattern Anal. Mach. Intell. 32 (5), 815–830.

Jimenez-del Toro, O., Müller, H., Krenn, M., Gruenberg, K., Taha, A.A., Winterstein, M., Eggel, I., Foncubierta-Rodríguez, A., Goksel, O., Jakab, A., et al., 2016. Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: visceral anatomy benchmarks. IEEE Trans. Med. Imag. 35 (11), 2459–2475.

Wachinger, C., Toews, M., Langs, G., Wells, W., Golland, P., 2015. Keypoint transfer segmentation. In: Ourselin, S., Alexander, D., Westin, C., Cardoso, M. (Eds.), IPMI 2015. LNCS, vol. 9123. Springer, Heidelberg, pp. 233–245.

Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A., 2013. Multi-atlas segmentation with joint label fusion. IEEE Trans. Pattern Anal. Mach. Intell. 35 (3), 611–623.

Wiatowski, T., Bölcskei, H., 2018. A mathematical theory of deep convolutional neural networks for feature extraction. IEEE Trans. Inf. Theor. 64 (3), 1845–1866.

Wolterink, J.M., Leiner, T., Viergever, M.A., Išgum, I., 2016. Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease. In: Reconstruction, Segmentation, and Analysis of Medical Images. Springer, pp. 95–102.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 5987–5995.

Xu, Z., Lee, C.P., Heinrich, M.P., Modat, M., Rueckert, D., Ourselin, S., Abramson, R.G., Landman, B.A., 2016. Evaluation of six registration methods for the human abdomen on clinically acquired CT. IEEE Trans. Biomed. Eng. 63 (8), 1563–1572.

Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122.

Yu, F., Koltun, V., Funkhouser, T.A., 2017. Dilated residual networks.. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2, p. 3.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2014. Object detectors emerge in deep scene CNNs. arXiv:1412.6856.

Zhou, Y., Xie, L., Shen, W., Wang, Y., Fishman, E.K., Yuille, A.L., 2017. A fixed-point model for pancreas segmentation in abdominal CT scans. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 693–701.