# Joint optic disc and cup segmentation based on multi-module U-shaped network

Zhu, Qianlong, Luo, Gaohui, Chen, Xinjian, Shi, Fei, Pan, Lingjiao, et al.

Qianlong Zhu, Gaohui Luo, Xinjian Chen, Fei Shi, Lingjiao Pan, Weifang Zhu, "Joint optic disc and cup segmentation based on multi-module U-shaped network," Proc. SPIE 11596, Medical Imaging 2021: Image Processing, 115961W (15 February 2021); doi: 10.1117/12.2580204

SPIE.

Event: SPIE Medical Imaging, 2021, Online Only

# Joint optic disc and cup segmentation based on multi-module U-shaped network

Qianlong Zhu[1,#], Gaohui Luo[1,#], Xinjian Chen[1,3], Fei Shi[1,2], Lingjiao Pan[4], Weifang Zhu[1,2,*]
[1] School of Electronics and Information Engineering, Soochow University, Suzhou, Jiangsu Province, 215006, China
[2] Collaborative Innovation Center of IoT Industrialization and Intelligent Production, Minjiang University, Fuzhou 350108, China
[3] State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou 215123, China
[4] School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou, Jiangsu Province, 213000, China
[#]indicates these authors contributed equally to this work.

## ABSTRACT

Glaucoma is a leading cause of irreversible blindness. Accurate optic disc (OD) and optic cup (OC) segmentation in fundus images is beneficial to glaucoma screening and diagnosis. Recently, convolutional neural networks have demonstrated promising progress in OD and OC joint segmentation in fundus images. However, the segmentation of OC is a challenge due to the low contrast and blurred boundary. In this paper, we propose an improved U-shape based network to jointly segment OD and OC. There are three main contributions: (1) The efficient channel attention (ECA) blocks are embedded into our proposed network to avoid dimensionality reduction and capture cross-channel interaction in an efficient way. (2) A multiplexed dilation convolution (MDC) module is proposed to extract more target features with various sizes and preserve more spatial information. (3) Three global context extraction (GCE) modules are used in our network. By introducing multiple GCE modules between encoder and decoder, the global semantic information flow from high-level stages can be gradually guided to different stages. The method proposed in this paper was tested on 240 fundus images. Compared with U-Net, Attention U-Net, Seg-Net and FCNs, the OD and OC's mean Dice similarity coefficient of the proposed method can reach 96.20% and 90.00% respectively, which are better than the above networks.

**Keywords:** Glaucoma, optic disc and cup segmentation, efficient channel attention, dilation convolution, global context extraction

## 1. INTRODUCTION

Glaucoma is one of the leading causes of irreversible vision loss. Screening and diagnosis at early stage can facilitate the treatment and reduce the risk of vision loss. With the recent advancements in optical fundus imaging, objective and quantitative glaucoma assessments based on the morphology of optic disc (OD) and optic cup (OC), and the cup-to-disc ratio (CDR) become available [1]. Accurately segmenting OD and OC in fundus image via automatic solutions would prompt the large scale glaucoma screening.

As shown in Fig.1, OC appears as a bright yellowish oval in fundus images, and OD is a darker one in turn. OC is contained in OD and the segmentation of OC region in fundus images is a challenge due to the low contrast and blurred boundary. To deal with these problems, we propose a multi-module U-shaped network for the joint segmentation of OD and OC.

*Corresponding author: Weifang Zhu, E-mail: wfzhu@suda.edu.cn

Medical Imaging 2021: Image Processing, edited by Ivana Išgum, Bennett A. Landman, Proc. of SPIE Vol. 11596, 115961W · © 2021 SPIE · CCC code: 1605-7422/21/$21 · doi: 10.1117/12.2580204

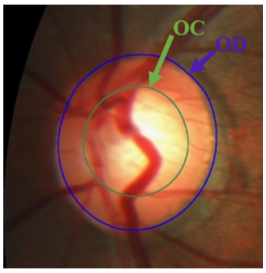


Figure 1. OD and OC in fundus image. OD: the region enclosed by the violet curve. OC: the region enclosed by the green curve

# 2. METHODS

## 2.1 Data Pre-processing

OD and OC are relatively small in fundus images, which are not conducive to the joint segmentation. A pre-trained Disc-aware Ensemble Network (DE-Net) [2] is used to crop the region of interest (ROI) patches (512 x 512) from the original fundus image (2124 x 2056). This data processing also can allow the model to focus on learning the most important pixel-wise information.

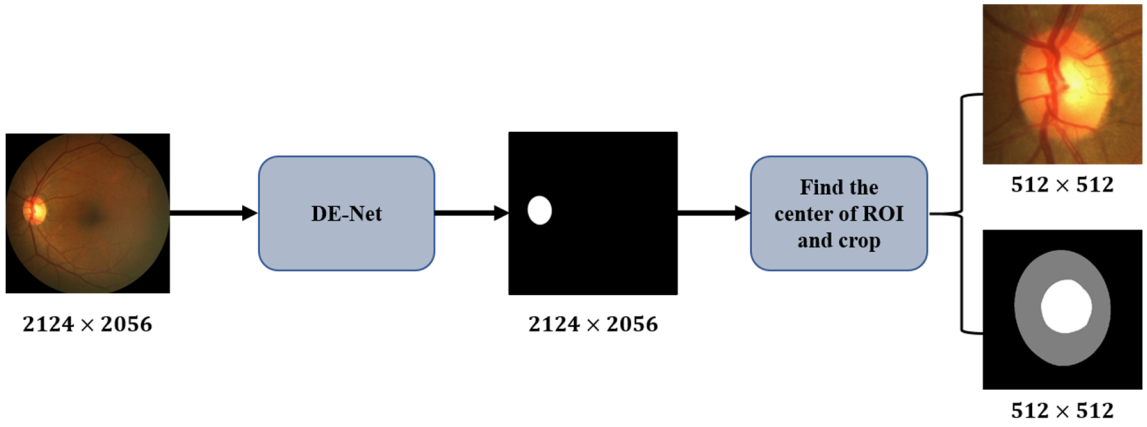


Figure 2. Data pre-processing for training, validation and test dataset.

## 2.2 Overall structure of the proposed network

In recent years, deep networks based on U-Net [3] have been widely used in medical image segmentation. In original U-Net architecture, each block of encoder consists of two 3x3 convolution layers and one max pooling layer. In the proposed network, we replace it with the pre-trained ResNet-34 [4] in the feature encoder module. And for compatibility purpose, the average pooling layer and fully connection layers are removed. Fig.3 illustrates the overall structure of the proposed network. We use the improved U-Net as the backbone of our network. The efficient channel attention (ECA) blocks [5] are embedded into the encoder and decoder path to avoid dimensionality reduction and capture cross-channel interaction in an efficient way. We integrate the proposed multiplexed dilation convolution (MDC) module with our proposed network, which can capture wider and deeper semantic features by infusing five cascade branches with multi-scale dilation convolutions. Besides, in order to solve the problem which the original skip-connection in the U-Net will introduce irrelevantly clutters and have semantic gap due to the mismatch of receptive fields, global context extraction (GCE) modules are proposed and embedded into the skip-connection of the backbone. The GCE modules combine multi-layer global context information to reconstruct skip-connection and provide global information guidance flow for the decoder.

Specifically, each layer's skip-connection consists of both local context information from this layer and global context information from higher-level layers.

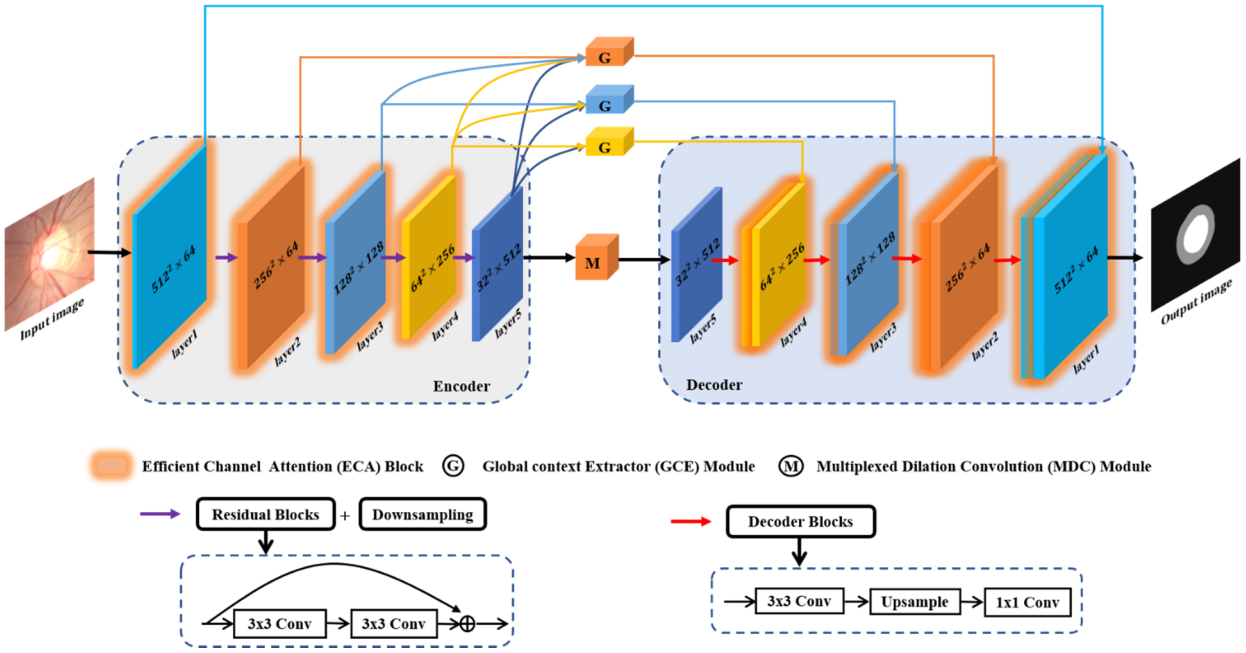


Figure 3. Overall structure of the proposed network

## 2.3 Global context extraction (GCE) module

In the GCE module, the skip-connection is reconstructed by combining the feature map of this layer with the feature maps of all higher-level stages. For example, Fig.4 shows the GCE module on Layer4. First, features of all stages are mapped into the same channel space as Layer4 by a regular 3×3 convolution. Next, the generated feature maps are upsampled to the same size as Layer4 and concatenated. In order to extract global context information from different levels of feature maps, the feature map F1 will be operated in parallel. On the one hand, two parallel separable convolutions with different dilation rates (1, 2) are performed on F1 and concatenated, and then a regular 1x1 convolution is used to obtain the feature map F2. On the other hand, we use the global context (GC) block [6] to extract global contextual information from F1. Then the feature map F3 is obtained after concatenation and a regular 1x1 convolution. Finally, the final feature map is obtained by concatenating F2 and F3 and then through a regular 1x1 convolution.
In summary, each GCE module in different stages can be summarized as:

$$GCE_k = \left(C_{i=k}^{i=5}\left(D^{2^{i-k}}\left(C_{i=k}^{i=5}(F_k \otimes 2^{i-j})\right)\right)\right) © \left(GC\left(C_{i=k}^{i=5}(F_k \otimes 2^{i-j})\right)\right) \quad (1)$$

Where $GCE_k$ denotes the output of GCE module inserted in the $k^{th}$ stage, $F_k$ denotes the feature map of the $k^{th}$ stage in the encoder, GC denotes the global context block, $\otimes 2^{i-k}$ represents the upsampling operation with rate of $2^{i-k}$, $C$ represents parallel operation, $D^{2^{i-k}}$ represents the separable convolutions with dilation rate of $2^{i-k}$, and © represents concatenation operation.

## 2.4 Multiplexed dilation convolution (MDC) module

Compared with the normal convolution, the dilation convolution [7] increases the receptive field without the loss of information due to pooling, so that the output of each convolution contains a wider range of feature information. So we propose the MDC module which is based on the dilation convolution. As shown in Fig.5, MDC has five cascade branches with the gradual increment of the dilation convolutions from 1 to 1, 2, 4, 8, whose receptive fields are 3x3, 5x5, 9x9, 17x17

and 33x33 respectively. The MDC module can extract objective features with different receptive fields by combining different dilation rates.

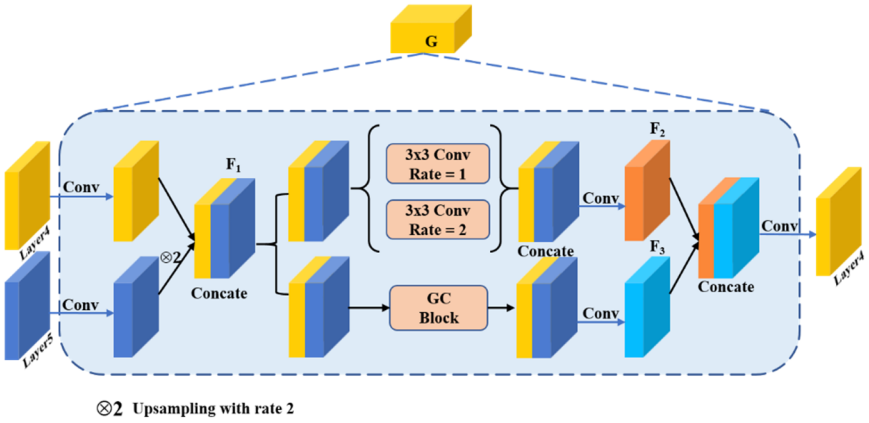


Figure 4. Structure of global context extraction (GCE) module

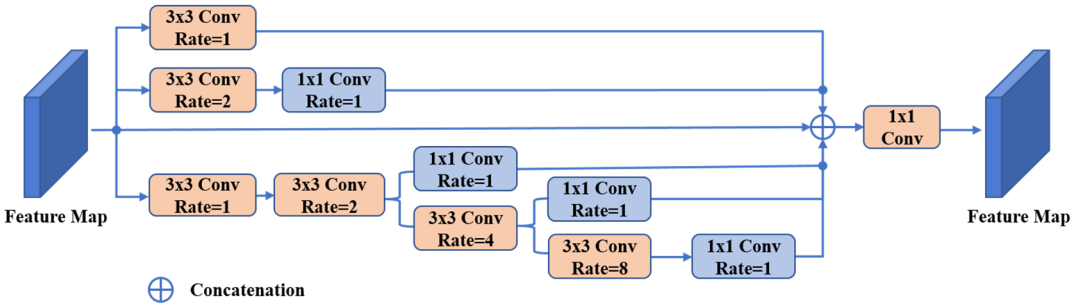


Figure 5. Structure of multiplexed dilation convolution (MDC) module

## 2.5 Loss function

To effectively solve the data imbalance problem in the training process, the combination of the Dice loss and the binary cross-entropy (BCE) loss is adopted as the joint loss, which can be defined as follows:

$$L_{Total} = L_{Dice} + L_{BCE} \tag{2}$$

$$L_{Dice} = 1 - \frac{2\sum_i^N \bar{y}_{(k,i)} y_{(k,i)} + \varepsilon}{\sum_i^N \bar{y}_{(k,i)} + \sum_i^N y_{(k,i)} + \varepsilon} \tag{3}$$

$$L_{BCE} = -\frac{1}{N}\sum_i^N \left( y_{(k,i)} \log \bar{y}_{(k,i)} + (1 - y_{(k,i)}) \log(1 - \bar{y}_{(k,i)}) \right) \tag{4}$$

Where $N$ indicates the batch size, $\bar{y}_i \in [0,1]$ and $y_i \in [0,1]$ denote the predicted probability and ground truth label respectively. $\varepsilon$ is a small smoothing factor.

# 3. RESULTS

## 3.1 Datasets

The experimental fundus images are acquired from MICCAI 2018 REFUGE challenge dataset [8]. The original dataset contains 1,200 fundus images and the corresponding OD and OC ground truth. We randomly divide the 1,200 fundus images into training set (720), validation set (240) and test set (240). To increase the generalization of the model, we adopt online augmentation strategy including left and right flipping, up and down flipping, random rotation and additive Gaussian noise addition. For each round of training, 2-4 of these augmentation methods are used.

## 3.2 Parameter settings

The proposed network is realized on the Pytorch1.1.0 framework. In the training process, the stochastic gradient descent (SGD) algorithm with an initial learning rate of 0.01, momentum of 0.9 and weight decay of 0.0001 is used to optimize the network. The batch size is set to 4 and the number of epochs is 40.

## 3.3 Evaluation metrics

To quantitatively evaluate the segmentation performance, two common segmentation evaluation metrics including Dice similarity coefficient (DSC) and intersection over union (IoU) are used.

$$\mathrm{DSC} = \frac{2TP}{2TP+FP+FN} \tag{5}$$

$$\mathrm{IoU} = \frac{TP}{TP+FP+FN} \tag{6}$$

Where *TP* denotes the true positive, *FP* denotes the false positive, *FN* denotes the false negative.

## 3.4 Results

To evaluate the performance of our method, we perform comparison experiments with other networks, including U-Net, Attention U-Net [9], Seg-Net [10] and FCNs [11]. As can be clearly seen in Fig.6, the segmentation results of our proposed method are closer to the ground truth in terms of OD and OC segmentation boundaries and shapes. In order to further verify the effectiveness of GCE, MDC and ECA, we use the improved U-Net as our backbone and perform three ablation experiments. Table1 shows objective evaluation metrics of experimental results, including the mean and standard deviation of dice similarity coefficient (DSC), intersection over union (IoU). As shown in Table1, the proposed method performs better than U-Net, Seg-Net, Attention U-Net and FCNs in all evaluation metrics. The ablation experiments (backbone + ECA, backbone + MDC, backbone + GCE and our method) in Table1 show the necessity and effectiveness of the proposed ECA, MDC and GCE modules.

Table 1. The performance of segmentation with different evaluation metrics.

| Methods | Optic disc | | Optic cup | |
|---|---|---|---|---|
| | DSC (%) | IoU (%) | DSC (%) | IoU (%) |
| **U-Net** | 94.87±4.9 | 90.58±7.3 | 87.22±8.0 | 78.10±10.9 |
| **Seg-Net** | 95.22±3.4 | 91.06±5.5 | 87.45±9.5 | 78.76±12.7 |
| **Attention U-Net** | 95.25±3.9 | 91.15±6.0 | 87.63±8.3 | 78.81±11.2 |
| **FCNs** | 95.27±2.5 | 91.07±4.3 | 87.59±6.9 | 78.51±9.7 |
| **Backbone** | 95.33±2.6 | 91.18±4.3 | 88.32±5.7 | 79.52±8.5 |
| **Backbone + ECA** | 95.79±2.2 | 92.01±3.8 | 89.48±5.8 | 81.43±9.0 |
| **Backbone + MDC** | 95.90±2.6 | 92.23±4.4 | 89.25±5.9 | 81.07±9.0 |
| **Backbone + GCE** | 95.85±2.8 | 92.14±4.5 | 89.25±5.7 | 81.03±8.8 |
| **Our method** | **96.20±1.9** | **92.73±3.3** | **90.00±5.3** | **82.20±8.2** |

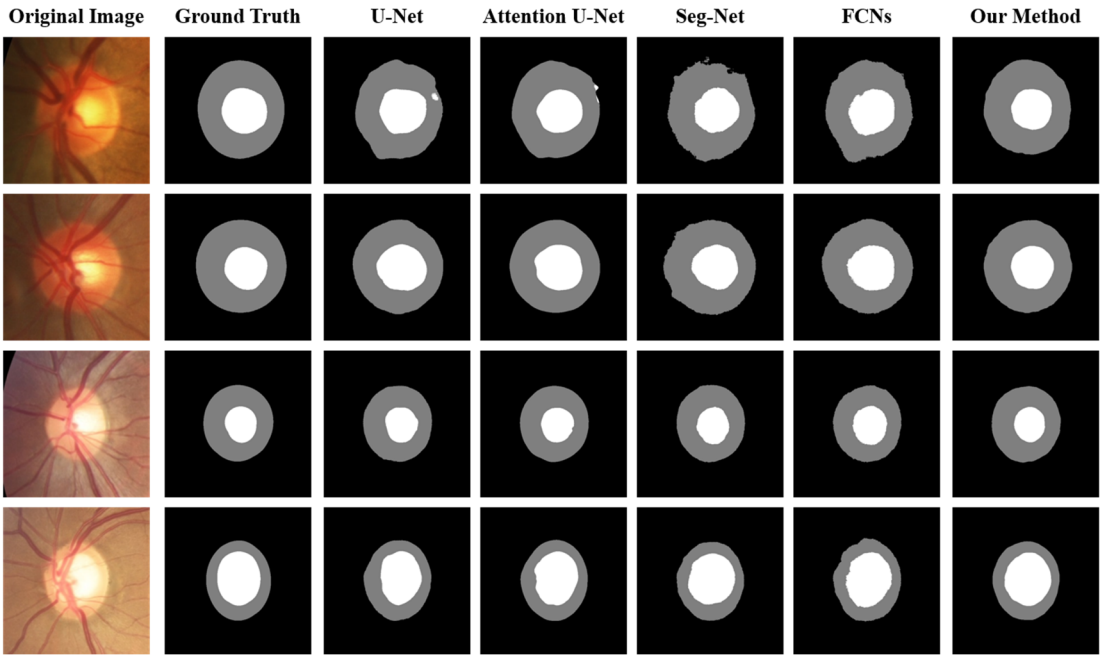


Figure 6. Examples of joint optic disc and cup segmentation results

## 4. CONCLUSIONS

In this paper, we propose a new multi-module U-shaped network for the joint segmentation of OD and OC in fundus images. To extract more objective features with different receptive fields and preserve more spatial information, we propose a multiplexed dilation convolution module. A global context extractor module is proposed to make the network pay more attention to the global context. The primary experiment results demonstrate the effectiveness of our proposed method.

## 5. ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0701700, in part by the National Nature Science Foundation of China under Grant 61622114 and 62001196, in part by the National Basic Research Program of China under Grant 2014CB748600, and in part by Collaborative Innovation Center of IoT Industrialization and Intelligent Production, Minjiang University under Grant IIC1702.

## REFERECENS

[1] M. C. V. Stella Mary, E. B. Rajsingh and G. R. Naik, "Retinal Fundus Image Analysis for Diagnosis of Glaucoma: A Comprehensive Survey," in IEEE Access, vol. 4, pp. 4327-4354, 2016.

[2] H. Fu et al., "Disc-Aware Ensemble Network for Glaucoma Screening From Fundus Image," in IEEE Transactions on Medical Imaging, vol. 37, no. 11, pp. 2493-2501, Nov. 2018.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation." in International Conference on Medical image computing and computer-assisted intervention, 2015, pp. 234-241: Springer.

[4] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778.

[5] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11534-11542.

[6] Y. Cao, J. Xu, S. Lin, F. Wei and H. Hu, "GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019, pp. 1971-1980.

[7] F. Yu, V. Koltun, "Multi-Scale Context Aggregation By Dilated Convolutions," arXiv preprint arXiv: 1511.07122, 2015.

[8] Available: https://refuge.grand-challenge.org/home/

[9] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz et al., "Attention U-Net: Learning Where to Look for the Pancreas," arXiv preprint arXiv:1804.03999, 2018.

[10] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 39,no. 12, pp. 2481-2495, 2017.

[11] E. Shelhamer, J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 640-651, 1 April 2017.