Enhancing OCT Signal by Fusion of GANs: Improving Statistical Power of Glaucoma Clinical Trials

Georgios Lazaridis[1,3,4], Marco Lorenzi[2], Sebastien Ourselin[3], and David Garway-Heath[4,5]

[1] Centre for Medical Image Computing, University College London, London, UK
[2] Université Côte d'Azur, Inria Sophia Antipolis, Epione Research Project, France
[3] School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK
[4] NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust, London, UK
[5] Institute of Ophthalmology, University College London, London, UK
G.Lazaridis@cs.ucl.ac.uk

Abstract. Accurately monitoring the efficacy of disease-modifying drugs in glaucoma therapy is of critical importance. Albeit high resolution spectral-domain optical coherence tomography (SDOCT) is now in widespread clinical use, past landmark glaucoma clinical trials have used time-domain optical coherence tomography (TDOCT), which leads, however, to poor statistical power due to low signal-to-noise characteristics. Here, we propose a probabilistic ensemble model for improving the statistical power of imaging-based clinical trials. TDOCT are converted to synthesized SDOCT images and segmented via Bayesian fusion of an ensemble of generative adversarial networks (GANs). The proposed model integrates super resolution (SR) and multi-atlas segmentation (MAS) in a principled way. Experiments on the UK Glaucoma Treatment Study (UKGTS) show that the model successfully combines the strengths of both techniques (improved image quality of SR and effective label propagation of MAS), and produces a significantly better separation between treatment arms than conventional segmentation of TDOCT.

1 Introduction

Glaucoma is the leading cause of irreversible blindness. Evaluating the progression rate of the pathology is crucial in order to assess the risk of functional impairment and to establish sound treatment strategies [1]. Clinically, optical coherence tomography (OCT) is used as a surrogate measure to evaluate retinal ganglion cell loss by measuring retinal nerve fibre layer (RNFL) thickness around the optic nerve head (ONH), whereas standard automated perimetry (SAP) is employed to assess the status of the visual field (VF) [1].

Glaucoma research has produced several clinical trials, trying to monitor the

disease progression and the efficacy of disease-modifying drugs. Up until the introduction of high-resolution spectral-domain OCT (SDOCT), trials relied on time-domain OCT (TDOCT), characterized by lower quality acquisitions and signal-to-noise (SNR) ratio. Thus, structural measurements in past studies provided low statistical power in detecting significant treatment effects. Such an example is the UK Glaucoma Treatment Study (UKGTS) [1]. The UKGTS is the only glaucoma study to assess the vision-preserving efficacy of one disease-modifying drug with both VF and OCT outcome. Nonetheless, TDOCT information could not be effectively combined with VF outcomes to improve detection of a treatment effect. Improving the quality of image-related anatomical measurements is therefore imperative for increasing statistical power in clinical trials.

While prospective studies seek to modify the statistical power determinants [2], retrospective analyses aim to maximize effect size in order to gain insight on the efficacy of disease-modifying drugs. For instance, optimal spatial image smoothing [3] prior to analysis can improve statistical power to detect group differences. In [4], it has been proposed to use reference images to guide statistical analysis of a new dataset through transfer learning, and to select only relevant voxels in novel studies. When image segmentation is required, multi-atlas segmentation (MAS) [5] is successful in leveraging diverse reference image information, by propagating atlas labels to novel image coordinates.

Meanwhile, various methods for super resolution (SR) using convolutional neural networks (CNNs), such as generative adversarial networks (GANs), have been proposed to transform image quality and appearance [6,7,8,9,10]. In medical imaging, GANs have been successfully employed to address the ill-posed nature of cross-modal synthesis. For example, in [6,7,8], GANs have been proposed to predict computed tomography (CT) and positron emission tomography (PET) images from magnetic resonance imaging (MRI). Concerning signal enhancement as well, in [9] and [10], synthesis was achieved at different resolution scales and by enforcing cycle-consistency, albeit not focusing on medical applications. These works may, however, present important limitations for SR in medical imaging. First, due to the restricted view of GANs spatial window, preservation of spatial smoothness and anatomical features in predictions is not always guaranteed. Second, single GAN predictions are characterized by spatial and intensity variability. Therefore, in order to extract robust anatomical quantifications from the output of GANs, principled schemes accounting for prediction uncertainty must be developed. This requires, for instance, probabilistic modeling of the uncertainty of the underlying signal distributions on distinct image parts, to preserve anatomical structures and account for spatial coherency.

This paper presents a novel method to improve the statistical power of clinical trials with low quality images. Our methodology leverages Bayesian fusion of GANs to infer morphological descriptors from low to high quality anatomical information. The transfer mapping is learned in an independent dataset and the proposed method is demonstrated on the UKGTS, enhancing the power of TDOCT via quality transfer from SDOCT. As a result, RNFL segmentations are improved and further refined via the effective label-propagation of MAS.

2 Materials and Methods

2.1 Data

We used two studies to validate and test our proposed methodology. For training and validation, we used the RAPID study: 82 glaucoma patients attended for up to 10 visits within a 3-month period, consisting of 4.902 TDOCT (StratusOCT, ZEISS) and 1.789 SDOCT (SpectralisOCT, Heidelberg Engineering) images. For testing, we used the UKGTS subset of participants with TDOCT imaging available [1]: 373 glaucoma patients, attended for up to 2 years. Eligible patients were assigned to treatment with Latanoprost 0.005% or placebo. The UGKTS consists solely of 78.415 TDOCT (StratusOCT, ZEISS) images.

2.2 Proposed Methodology

The definition of our framework requires to address a number of challenges. First, due to different acquisition protocols, the pairing between target SDOCT and predictor TDOCT training images is ill-defined. To solve this issue, we propose an automated method for target-predictor image pairing (Sec. 2.2.1). Second, OCT signal is characterized by diverse degrees of noise and spatial information, whereas RNFL segmentation is subject to variability due to the different attributes of the synthesized images. This problem is tackled in Sec. 2.2.2, where we present our method to obtain representations accounting for the different spatial coherence of OCT images. Finally, in Sec. 2.2.3 we identify a probabilistic consensus strategy for RNFL segmentations on the average synthesized image.

2.2.1 Training Pairs Generation

Although TDOCT and SDOCT images were acquired at each patient visit, there is not a correspondence between the two sets of predictor and target modalities. Our method finds a matching based on global and local image information represented by (i) the vessel profile given by the average retinal pigment epithelium (RPE) pixel intensity, (ii) the internal limiting membrane (ILM) contour and (iii) the average norm of the deformation fields between TDOCT and SDOCT images within a patient's longitudinal history. First, as the topography around the ONH undulates, we flatten all images using a pilot estimate of the hyperreflective RPE layer. Hence, images are aligned according to a fixed vertical RPE offset. We further exploit the RPE identification to detect the vessels, as they appear as shaded bands in the RPE. We then segment the ILM contour (upper high-contrast boundary on the dark-to-bright gradient image) and smooth it by Gaussian Process interpolation. Iterative closest point was used to evaluate the matching between the sets of features in (i) and (ii), and mutual information to evaluate the image registration in (iii). We evaluate the robustness of our pairing method on a benchmark of synthetic images with spatial variability, achieving 100% sensitivity (see Supplementary material). We note that a patient with N TDOCT and M SDOCT can theoretically produce a maximum of $N \times M$ images. Application to the RAPID dataset lead to 24.792 TDOCT and SDOCT pairs.

2.2.2 Ensemble GANs

To account for the specific anatomical geometry and signal properties in OCT images, we propose an adaptation of standard cycle-consistent GANs (cycleGANs) [10], to improve robustness and accuracy of the modality transfer. OCT images have a very specific geometry where the background, i.e. vitreous cavity, is clearly separated from the layers at the ILM. Thus, we used image stitching, exploiting the ILM identification, to separate background from layer signal. Moreover, cycleGANs require a fixed window on which spatial filters and mappings are learned. However, since OCT signal and noise properties are characterized by different spatial scales, a modality transfer method based on a fixed spatial window might not be able to capture all the necessary spatial information needed for synthesis. This reduces the chance for cross-modal distributions to share supports in latent space. To address this problem, we propose an ensemble of spatially coherent cycleGANs [10] to learn the TDOCT-to-SDOCT mapping and to translate a TDOCT into a synthesized SDOCT image. The scheme is the following. Each GAN is trained by employing a different spatial window size: 128×128, 256×256 and 512×512, learning a mapping from the observed TDOCT image $\boldsymbol{I}_{TD}$ and random noise vector $\boldsymbol{z}$, to the target SDOCT image $\boldsymbol{I}_{SD}$, $G\colon \{\boldsymbol{I}_{TD},\ \boldsymbol{z}\} \rightarrow \boldsymbol{I}_{SD}$. As a result, we train six GANs: three with background pairs and three with layer pairs. The synthesized backgrounds and layers are stitched back according to the window size, i.e. $\boldsymbol{I}_{128\times128}$, $\boldsymbol{I}_{256\times256}$, $\boldsymbol{I}_{512\times512}$ and the average synthesized stitched image $\bar{\boldsymbol{I}}$ is obtained. To preserve the morphological correlation between training pairs, cycleGANs were trained with windows centered at the same geometrical location in both pairs. Fig. 1 shows the proposed framework for OCT synthesis via the ensemble of GANs.

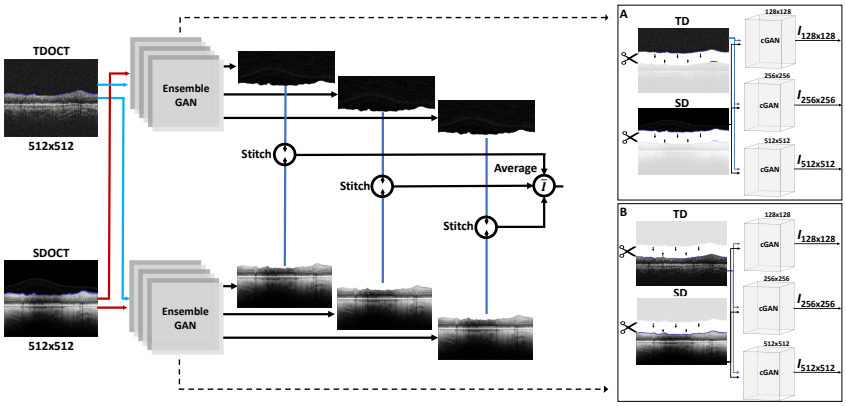
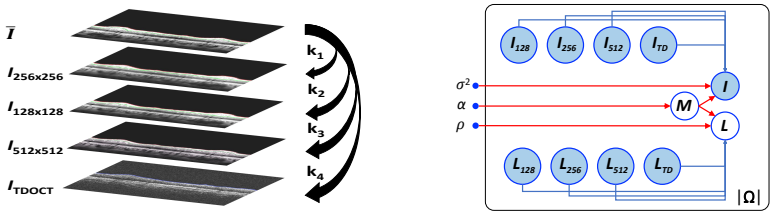


Fig. 1: SDOCT synthesis via ensemble of GANs. Three GANs are trained with backgrounds (box A) and three with layers (box B). Synthesized images are stitched back and the average synthesized stitched image is obtained. Separation of layers and background is illustrated with scissors.



(a) (b)

Fig. 2: (a) Stack of images, where k_1, k_2, k_3, k_4 are the distances between $\bar{\boldsymbol{I}}$ and $\boldsymbol{I}_{256\times256}$, $\boldsymbol{I}_{128\times128}$, $\boldsymbol{I}_{512\times512}$ and $\boldsymbol{I}_{\text{TDOCT}}$. (b) Graphical model representing the relationship between the model variables in MAS. Replications are illustrated with plates. Shaded variables are observed.

2.2.3 Multi-Atlas Segmentation

Once the average synthesized stitched image $\bar{\boldsymbol{I}}$ is obtained, the problem consists in finding a robust RNFL segmentation accounting for the signal variability introduced by the synthesis. We treat images as being in a stack where $\bar{\boldsymbol{I}}$ is used as test image, and, $\boldsymbol{I}_{128\times128}$, $\boldsymbol{I}_{256\times256}$, $\boldsymbol{I}_{512\times512}$, and the original $\boldsymbol{I}_{\text{TDOCT}}$ as atlases, here denoted by $\{I_n(\boldsymbol{x})\}_{n=1,\dots,4}$ (Fig. 2a). We want to propagate the atlas RNFL labels to the novel test image coordinates, where the segmentation of each pixel is decided through a label fusion approach. To account for the variability across atlases, we rely on a Bayesian model averaging technique, the graphical model of which is shown in Fig. 2b. Let $\{L_n(\boldsymbol{x})\}_{n=1,\dots,4}$ be segmentations corresponding to the atlases $\{I_n(\boldsymbol{x})\}$. We assume that these atlases are co-registered to the test image $\bar{I}(\boldsymbol{x})$, with unknown labels $L(\boldsymbol{x})$. A label fusion approach aims to estimate the label map $\boldsymbol{L}$ associated with $\bar{\boldsymbol{I}}$, given the registered atlases. We assume that the posterior probability of the segmentation p factorizes over pixels:

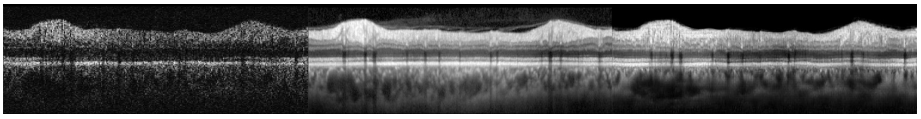
$$p(\boldsymbol{L}|\{\boldsymbol{I}_n\},\{\boldsymbol{L}_n\},\bar{\boldsymbol{I}}) = \prod_{\boldsymbol{x}\in\Omega} p_x(L(\boldsymbol{x})|\{\boldsymbol{I}_n\},\{\boldsymbol{L}_n\},\bar{\boldsymbol{I}}) \tag{1}$$

To model p_x, we choose the local label fusion model from [5], which relies on a latent discrete field $M(\boldsymbol{x})$ that indexes which atlas generates the test image and its segmentation at each location. The model further assumes that the image intensities $\bar{\boldsymbol{I}}$ and labels $\mathbf{L}$ are conditionally independent given the field $\boldsymbol{M}$. Following [5], we use a Gaussian likelihood term for the image intensities and a LogOdds model based on the signed distance transform for the labels. We use a prior for the field $\boldsymbol{M}$ that reflects lower reliability for the atlases associated with lower registration accuracy [11]. For each 2D location $\boldsymbol{x}$, the prior takes the form $p(M(\boldsymbol{x}) = n) \propto \exp(-k_n\alpha)$, where the coefficients k_n, $n = 1, 2, 3, 4$, are the distances between the test image $\bar{\boldsymbol{I}}$ and the atlases, while α is a parameter controlling the sharpness of the prior. Based on our experimental registration results, we empirically set the lowest distance value, $k_1 = 1$, for the atlas $\boldsymbol{I}_{256\times256}$, and increasing ones, $k_i = i$, for respectively the atlases $\boldsymbol{I}_{128\times128}$, $\boldsymbol{I}_{512\times512}$ and

$\boldsymbol{I}_{\text{TDOCT}}$. The posterior probability for the labels is finally [5]:

$$p(L(\boldsymbol{x})|\{\boldsymbol{I}_n\},\{\boldsymbol{L}_n\},\bar{\boldsymbol{I}}) = \frac{\sum_{n=1}^{N}\mathcal{N}\left(\bar{I}(\boldsymbol{x});I_n(\boldsymbol{x}),\sigma^2\right)e^{\rho D_x[L(x);\boldsymbol{L}_n]}e^{-k_n\alpha}}{\sum_{n=1}^{N}e^{-k_n\alpha}\mathcal{N}\left(\bar{I}(\boldsymbol{x});I_n(\boldsymbol{x}),\sigma^2\right)} \quad (2)$$

where $\mathcal{N}$ is the Gaussian probability density function; D_x is the signed distance transform evaluated at location x; and σ^2 and ρ are the likelihood parameters.



(a) TDOCT (b) SDOCT (c) Synthesized SDOCT

Fig. 3: OCT synthesis results via fusion of GANs. (a) and (b) illustrate a pair of TDOCT and SDOCT images. (c) Synthesized SDOCT from (a).

3 Experiments and Results

3.1 Experimental Setup

We compared our method with respect to the results obtained with each single GAN used in our pipeline (Fig. 1), to a label fusion strategy on the GANs output, and to the original images provided by the StratusOCT machine. Testing on UKGTS was instead performed by quantifying the statistical power relative to the measurements obtained with our method, as compared with those derived from the StratusOCT, following the same evaluation protocol from prior image-to-image translation studies [10]. To quantify the quality of the synthesized SDOCT images, we segmented their RNFL and compared the resulting average RNFL thickness with the original SDOCT average RNFL thickness. The intuition is that if we can produce realistic SDOCT images, an off-the-shelf segmentation model should output the same RNFL thickness obtained with the original data. We adopt the layer segmentation model of Mayer et al. [12]. For label fusion, as atlases, we used the segmented RNFL sections of the synthesized SDOCT and the original TDOCT RNFL segmentation. For the test image, we used the average synthesized stitched image in which we registered the retinal layers of the atlases. We used the method from [13] for non-rigid registration of OCT layers, and computed predictions for the final RNFL labels with Eq. 2. The parameters were kept constant for all experiments: $\sigma^2 = 625$, $\rho = 30\mu m^{-1}$, $\alpha = 1mm^{-1}$. Decaying weights were set depending on the agreement measured when evaluating GANs performance individually. We used 9-Block Resnet models as generators, and 70×70 PatchGANs as the two discriminators [10]. All experiments were performed on a NVIDIA Titan X (12GB) GPU.

Table 1: Limits of agreement, mean difference, correlation of all methods versus ground truth, and mean SD of the first three visits difference for both eyes.

Method	GAN			Label Fusion		StratusOCT
	128x128	256x256	512x512	Direct	Proposed	
95% LOA	[22.53, -18.7]	[16.9, -14.2]	[23.34, -19.35]	[11.72, -9.72]	[**8.11**, **-6.73**]	[26.64, -22.95]
Mean Diff.	1.92	1.44	1.99	1.00	**0.69**	1.84
Pearson r	0.79	0.85	0.71	0.89	**0.92**	0.76
Mean SD	2.27	1.87	3.01	1.33	**1.29**	2.67

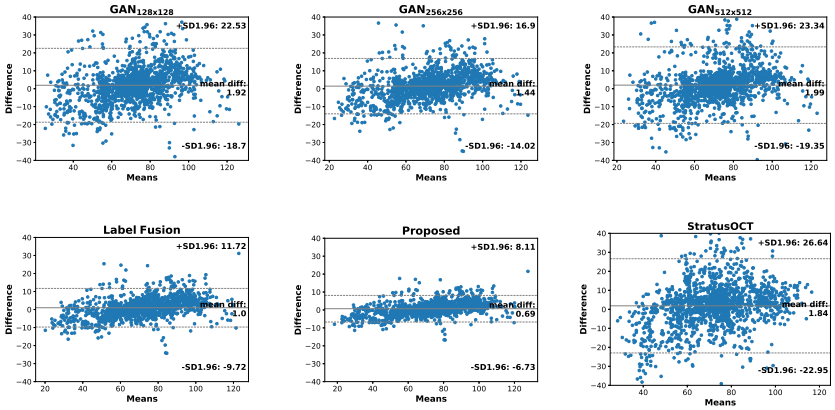


Fig. 4: Bland-Altman plots on the agreement between all methods versus ground truth on RAPID. The proposed method leads to significantly better agreement.

3.2 Results

Table 1 shows the 95% limits of agreement (LOA), mean difference, correlation and the mean standard deviation (SD) of the difference for three visits across all subjects of the RAPID study. $\text{GAN}_{256\times256}$ yields better scores compared to $\text{GAN}_{128\times128}$ and $\text{GAN}_{512\times512}$. Label fusion, without image stitching, on the average synthesized image outperforms the individual output of GANs, while a further improvement is obtained by integrating image stitching. These results suggest that combining the synthesized images of each GAN enables us to take advantage of the strengths of all architectures. Fig. 4 illustrates the compatibility of the measurements with respect to the ground truth SDOCT segmentation in Bland-Altman plots. Our approach not only manages to produce a RNFL segmentation close to the ground truth, but also reduces the variability in the measurements. We applied our method to the TDOCT images available from the UKGTS and subsequently segmented the newly synthesized SDOCT images. Table 2 shows the results of our method compared to the original StratusOCT. We appreciate a statistically significant improvement in the separation between

treatment and placebo groups ($p = 0.0017$), leading to sensibly lower sample size in power analysis.

Table 2: Comparison of rate of RNFL change between our method and Stratus OCT in the UKGTS. Significant difference between treatment and placebo progression rates ($p < 0.05$, Mann–Whitney U test) is indicated with (*). Sample size for 80% power with $p = 0.05$.

Method	StratusOCT		Proposed	
	Treatment	Placebo	Treatment	Placebo
Mean (SD) (μm/visit)	0.0344 (1.964)	-0.0733 (2.066)	-0.0760 (1.5019)	-0.341 (1.8027)
Diff. in mean rate (95% CI)	0.107 (-0.358 to 0.574)		0.265* (-0.118 to 0.648)	
Sample size	5495		616	

4 Discussion and Conclusion

We presented a probabilistic ensemble model for enhancing the statistical power of clinical trials with RNFL thickness change outcome derived from TDOCT. Our approach is based on image synthesis and semi-automated segmentation of synthesized SDOCT images, integrating label fusion with image stitching and deep learning to further improve statistical separation between treatment groups. The proposed methodology appears robust and flexible both in terms of architecture and label fusion. Future work will focus on modifying a regularization scheme to improve conditioning on RNFL and on integrating, in parallel, multiple resolution scales.

Acknowledgements. Thiw work was supported by the EPSRC (CDT in Medical Imaging, EP/L016478/1) and Santen Pharmaceutical Co., Ltd.

References

1. Garway-Heath, D.F., Crabb, D.P., et al.: Latanoprost for Open-Angle Glaucoma (UKGTS): A Randomised, Multicentre, Placebo-Controlled Trial. The Lancet **385**(9975), 1295–1304 (2015)
2. Button, K., Ioannidis, J., et al.: Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. Nature reviews. Neuroscience **14** (04 2013)
3. Zhang, T., Davatzikos, C.: Optimally-Discriminative Voxel-Based Analysis. In: MICCAI. pp. 257–265. Springer (2010)
4. Schwartz, Y., Varoquaux, G., Pallier, C., Pinel, P., Poline, J.B., Thirion, B.: Improving Accuracy and Power with Transfer Learning Using a Meta-analytic Database. In: MICCAI. pp. 248–255. Springer (2012)

5. Sabuncu, M.R., Yeo, B.T.T., Van Leemput, K., Fischl, B., Golland, P.: A Generative Model for Image Segmentation Based on Label Fusion. IEEE Trans. Med. Imaging **29**(10), 1714–1729 (Oct 2010)
6. Nie, D., Trullo, R., et al.: Medical image synthesis with context-aware Generative Adversarial Networks. In: MICCAI. pp. 417–425. Springer (2017)
7. Wolterink, J.M., et al.: Deep MR to CT Synthesis Using Unpaired Data. In: MICCAI. pp. 14–23. Springer (2017)
8. Ben-Cohen, A., Klang, E., Raskin, S.P., Amitai, M.M., Greenspan, H.: Virtual PET Images from CT Data Using Deep Convolutional Networks: Initial Results. In: MICCAI. pp. 49—57. Springer (2017)
9. Wang, T.C., Liu, M.Y., et al.: High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. 2018 IEEE CVPR pp. 8798–8807 (Jun 2018)
10. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. IEEE ICCV pp. 2242–2251 (2017). https://doi.org/10.1109/ICCV.2017.244
11. Atzeni, A., Jansen, M., Ourselin, S., Iglesias, J.E.: A Probabilistic Model Combining Deep Learning and Multi-atlas Segmentation for Semi-automated Labelling of Histology. In: MICCAI. pp. 219–227. Springer (2018)
12. Mayer, M.A., Hornegger, J., Mardin, C.Y., Tornow, R.P.: Retinal Nerve Fiber Layer Segmentation on FD-OCT Scans of Normal Subjects and Glaucoma Patients. Biomed. Opt. Express **1**(5), 1358–1383 (Dec 2010). https://doi.org/10.1364/BOE.1.001358
13. Du, X., Gong, L., et al.: Non-rigid Registration of Retinal OCT Images Using Conditional Correlation Ratio. In: MICCAI OMIA. pp. 159–167. Springer (2017)