Longitudinal Image Registration with Temporal-order and Subject-specificity Discrimination

Qianye Yang[1], Yunguan Fu[1,2], Francesco Giganti[3,4], Nooshin Ghavami[1,5], Qingchao Chen[6], J. Alison Noble[6], Tom Vercauteren[5], Dean Barratt[1], and Yipeng Hu[1,6]

[1] Centre for Medical Image Computing and Wellcome/EPSRC Centre for Interventional & Surgical Sciences, University College London, London, UK
[2] InstaDeep, London, UK
[3] Division of Surgery & Interventional Science, University College London, London, UK
[4] Department of Radiology, University College London Hospital NHS Foundation Trust, London, UK
[5] School of Biomedical Engineering & Imaging Sciences, Kings College London, London, UK
[6] Institute of Biomedical Engineering, University of Oxford, Oxford, UK
qianye.yang.19@ucl.ac.uk

Abstract. Morphological analysis of longitudinal MR images plays a key role in monitoring disease progression for prostate cancer patients, who are placed under an active surveillance program. In this paper, we describe a learning-based image registration algorithm to quantify changes on regions of interest between a pair of images from the same patient, acquired at two different time points. Combining intensity-based similarity and gland segmentation as weak supervision, the population-data-trained registration networks significantly lowered the target registration errors (TREs) on holdout patient data, compared with those before registration and those from an iterative registration algorithm. Furthermore, this work provides a quantitative analysis on several longitudinal-data-sampling strategies and, in turn, we propose a novel regularisation method based on maximum mean discrepancy, between differently-sampled training image pairs. Based on 216 3D MR images from 86 patients, we report a mean TRE of 5.6 mm and show statistically significant differences between the different training data sampling strategies.

Keywords: Medical image registration · Longitudinal data · Maximum mean discrepancy.

1 Introduction

Multiparametric MR (mpMR) imaging has gained increasing acceptance as a noninvasive diagnostic tool for detecting and staging prostate cancer [10]. Active

surveillance recruits patients with low-grade cancers that exhibit low-to-medium risk to long-term outcome [10], where mpMR imaging has been adopted to follow regions within the prostate glands and to recognise or even predict the disease progression [7]. As outlined by the panel of experts who drafted the PRECISE criteria for serial MR reporting in patients on active surveillance for prostate cancer [10], assessing radiological changes of morphological MR features is a key component when reporting longitudinal MR images. For individual regions of pathological interest, these morphological features include volume, shape, boundary, extension to neighbouring anatomical structures and the degree of conspicurity in these features. Pairwise medical image registration quantifies the morphological correspondence between two images, potentially providing an automated computational tool to measure these changes, without time-consuming and observer-biased manual reporting. This voxel-level analysis is particularly useful in developing imaging biomarkers, when the ground-truth of the disease progression are still under debate, as in this application, and cannot be reliably used to train an end-to-end progression classifier.

Registration algorithms designed for longitudinal images have been proposed for several other applications [14,5,8], such as those utilising temporal regularisation [13] when applied to a data set acquired at three or more time points. Most algorithms are still based on or derived from the basic pairwise methodologies. In this work, registration of a pair of longitudinal prostate MR images from the same patient is investigated. Classical algorithms iteratively optimise a spatial transformation between two given images without using population data. For example, a fixed set of parameters in an iterative registration algorithm may work well for one patient, but unlikely to be optimal for other patients. Substantial inter-patient variation leads to the lack of common intensity patterns or structures between different prostates. Ad hoc benign foci, varying zonal anatomy and highly patient-specific pathology are frequently observed in MR images, especially within the prostate gland. This is particularly problematic for classical iterative registration algorithms, when the regions of interest, smaller and non-metastasis tumours, are confined within the varying prostate glands, such as those from the active surveillance patient cohort considered in this study. We provide such an example in the presented results using an iterative registration algorithm.

In this paper, we propose an alternative method that uses recently-introduced deep-learning-based non-iterative registration for this application. Based on the results on holdout patients, we argue that learning population statistics in patient-specific intensity patterns [2] and weak segmentation labels [6] can overcome the difficulties due to the large inter-patient variation, for aligning intra-patient prostate MR images. In order to efficiently and effectively utilise the often limited longitudinal data for training the registration network, we compare several methods to sample the time-ordered image pairs from the same patients and those from different patients, and propose a new regularisation method to discriminate time-forward image pairs versus time-backward image pairs and/or subject-specific image pairs versus inter-subject image pairs.
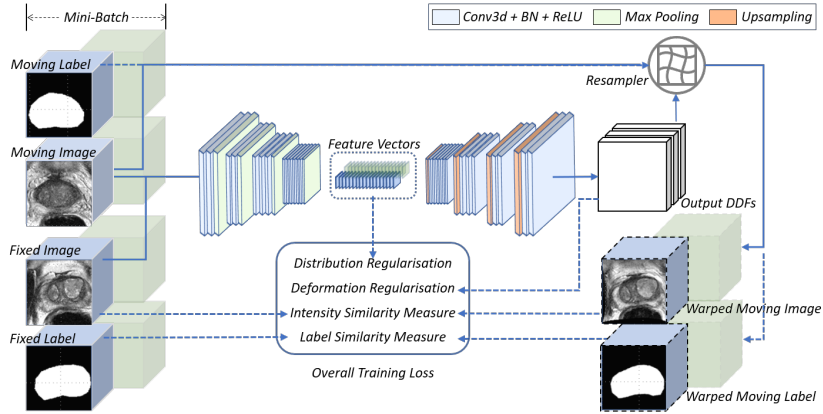
Fig. 1. Demonstration of the proposed image registration framework, with the dotted lines indicating the data flow only at the training.

We summarise the contributions in this work: 1) We developed an end-to-end deep-learning registration method tested on longitudinal images. To our knowledge, this is the first study for longitudinal MR Image registration for prostate cancer patients; 2) We present a quantitative analysis on longitudinal data sampling strategies for registration network training, with and without the new regularisation method; and 3) We report a set of rigorous results based on real clinical longitudinal data, demonstrating statistically significant differences between these sampling methods. This provides practically useful evidences for further development of the registration tools in longitudinal image analysis for active surveillance patients.

2   Methods

2.1   Learning-based image registration

In this work, the pairwise registration paradigm based on deep learning is adopted for registering longitudinal MR image pairs. Denote $\{(x_n^A, x_n^B)\}, n = 1...N$, as a set of paired images to register, x_n^A and x_n^B being the moving- and fixed images, respectively. For each pair n in the set, let a pair of corresponding prostate gland anatomical segmentation, represented with binary masks, be (l_n^A, l_n^B). During training a registration network $f(\theta)$ with network parameters θ, the nth pair of images is fed into the network to predict a dense displacement field (DDF) $\mu_n^{(\theta)}$. An overview of the network training is illustrated in Fig. 1.

The training loss is comprised of three measures of the goodness-of-predicted-DDF: 1) an intensity-based similarity measure between the DDF-warped moving image $x_n^A \circ \mu_n^{(\theta)}$ and the fixed image x_n^B, the sum-of-square differences (SSD) in intensity values being used in this study; 2) an overlap measure between the

DDF-warped moving gland segmentation $l_n^A \circ \mu_n^{(\theta)}$ and the fixed gland segmentation l_n^B based on a multi-scale Dice [6]; and 3) a deformation regularisation to penalise non-smooth DDFs using bending energy [12]. With a minibatch gradient descent optimisation, the network weights θ are optimised by minimising the overall loss function $J(\theta)$:

$$J(\theta) = \frac{1}{N} \sum_{n=1}^{N} (-\alpha \cdot Dice(l_n^B, l_n^A \circ \mu_n^{(\theta)}) + \beta \cdot SSD(x_n^B, x_n^A \circ \mu_n^{(\theta)}) + \gamma \cdot \Omega_{bending}(\mu_n^{(\theta)}))$$

(1)

where, α, β and γ are three hyper-parameters controlling the weights between the weak supervision, the intensity similarity and the deformation regulariser. These unsupervised and weakly-supervised losses were selected based on limited experiments on a validation data set, among other options, such as cross-correlation, Jaccard and DDF gradient norms. The three weights in Eq. 1 are co-dependent with the learning rate and potentially can be reduced to two. Therefore, the fine-tuning of these hyper-parameters warrants further investigation in future studies.

2.2   Training data distribution

Temporal-order and subject-specificity Fig. 2 illustrates example longitudinal images from individual subjects (prostate cancer patients) at multiple visits in order of time. Without loss of generality, we aim to model the morphological changes to quantify a chronological deformation field, i.e. from a baseline T2-weighted image to a follow-up, acquired at time points t_1 and t_2, $t_1 < t_2$, respectively. To train a registration network for this task, one can sample *task-specific training data*, i.e. intra-subject, time-forward image pairs. Given sufficient training data, i.e. the empirical training data distribution adequately representing the population data distribution, there is little reason to add other types of image pairs, i.e. time-backward or inter-subject pairs.
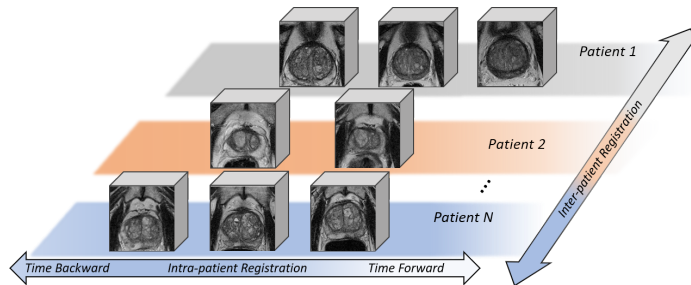

Fig. 2. Demonstration of a longitudinal image data set from active surveillance.

However, as is common in the field of medical image computing, the acquisition of training data is limited by various practical or clinical reasons. This leads to an empirical risk minimisation (ERM) on a sparse empirical training distribution [16]. ERM is known to be prone to overfitting [16]. Data augmentation strategies such as using affine/nonrigid transformation and the "mixup" [17], in geometric transformation and intensity spaces, respectively, have been applied to overcome overfitting. The former is also applied in this work.

Particularly interesting to longitudinal pairwise image registration, image pairs with reversed temporal-order, the time-backward pairs, and those with different subject-specificity, the inter-subject pairs, can be considered as augmenting data for this task to potentially improve generalisability. This is of practical importance in training longitudinal image registration networks and is the first hypothesis to test in this study.

Discriminating prior for regularising network training Furthermore, a data augmentation strategy becomes effective if the augmented empirical data distribution better represents population distribution or "aligns" empirical feature vectors with a potentially new population distribution in feature vector space [15]. Directly mixing the intra-/inter-subject and/or time-forward/-backward image pairs does not guarantee such alignment. To utilise the prior knowledge of the known temporal order and subject specificity, we test a new regularisation approach based on empirical maximum mean discrepancy (MMD) [4]. By penalising the divergence between the feature vectors generated from different empirical data distributions. As illustrated in Fig. 1, an MMD square is computed over the two sets of high-dimensional feature vectors from the network encoder, $\{v_i\}_{i=1}^I$ and $\{v_j\}_{j=1}^J$, generated from different types of image pairs:

$$\Omega_{mmd} = \frac{1}{I^2} \sum_{i \neq i'}^{I} k_{II}(v_i, v_{i'}) - \frac{2}{IJ} \sum_{i,j=1}^{I,J} k_{IJ}(v_i, v_j) + \frac{1}{J^2} \sum_{j \neq j'}^{J} k_{JJ}(v_j, v_{j'}), \quad (2)$$

where $k(\mu, \nu) = \exp(-||\mu - \nu||^2/2\sigma)$ is a Gaussian kernel with a parameter σ [3]. I and J are the sample numbers of the feature vectors, within a minibatch of size $N = I + J$. The weighted MMD regularising term in Eq. 2 is added to the original loss in Eq. 1:

$$J^*(\theta) = J(\theta) + \lambda \cdot \Omega_{mmd}(\{v_i\}, \{v_j\}). \quad (3)$$

With the new loss in Eq. 3, we test the second hypothesis that encoding the discrimination of temporal order and subject specificity can further improve the network generalisability.

2.3   Validation

Gerneralisability on holdout data The patients and their data is randomly assigned into training, validation and holdout sets. The networks are developed

with training and validation sets, including hyper-parameter tuning. The generalisability is measured on the holdout set using three metrics: 1) the binary Dice similarity coefficient (DSC) between the fixed label l^B and the warped moving label $l^A \circ \mu$; 2) the MSE between the fixed image x^B and the wrapped moving image $x^A \circ \mu$; and 3) the centroid distance between the aligned prostate glands l^B and $l^A \circ \mu$. Results from paired t-tests with a significance level of $\alpha = 0.05$ are reported when comparing these metrics.

Registration performance Also on the holdout set, pairs of corresponding anatomical and pathological landmarks are manually identified on moving and fixed images, including patient-specific fluid-filled cysts, calcification and centroids of zonal boundaries. The target registration errors (TREs) between the corresponding pairs of landmarks, from the warped moving and those from the fixed images, are computed to quantify the registration performance. Other experiment details are provided in Section 3.

3   Experiments

3.1   Data and preprocessing

216 longitudinal prostate T2-weighted MR images were acquired from 86 patients at University College London Hospitals NHS Foundation. For each patient 2-4 images were available, with intervals between consecutive visits ranging from 11 to 28 months. All the image volumes were resampled to $0.7 \times 0.7 \times 0.7$ mm^3 isotropic voxels with a normalised intensity range of $[0, 1]$. For computational consideration, all images were also cropped from the image center to $128 \times 128 \times 102$ voxels, such that the prostate glands are preserved. The prostate glands were manually segmented for the weak supervision in training and for validation. The images were split into 70, 6 and 10 patients for training, validation and holdout sets.

3.2   Network training

A previously proposed DDF-predicting encoder-decoder architecture was used [6], which is an adapted 3D U-Net [11], with more densely connected skip layers and residual shortcuts. The network training was implemented with TensorFlow 2 [1] and made open source `https://github.com/DeepRegNet/DeepReg`. The Adam optimizer with an initial learning rate of 10^{-5} was used with the hyper-parameters α, β, γ and λ empirically set to 1, 1, 50 and 0.01, respectively, via qualitative assessment on the validation set. The networks were trained on Nvidia Tesla V100 GPUs with a minibatch of 4 sets of image-label data, each containing a pair of T2 MR images and a corresponding binary mask pair of prostate gland segmentation. Each network run for 272,000 iterations, approximately 64 hours.

3.3   Training image pair sampling

To test the first hypothesis in Section 2.2, three different training data sets were sampled, resulting in three networks: a network trained with only intra-subject, time-forward image pairs (IF); a second network (IF+IB) trained using all possible intra-subject pairs regardless of temporal order; and the third network (IT+IF+IB) with added inter-subject samples. All the networks were trained with the registration loss function in Eq. 1. Generalisability and TREs were computed on all the intra-subject, time-forward image pairs from the same holdout patient data.

To test the second hypothesis, two more networks were trained with the loss in Eq. 3, with respective training data sets IF+IB and IT+IF+IB. For intra-patient IF+IB pairs, two images were randomly sampled without replacement from a single patient. MMD may be sensitive to minibatch size and sample size imbalance [4], a two-stage sampling was adapted to ensure every minibatch samples 2 IF and 2 IB pairs during training the IF+IB network; and samples 2 IF/IB pairs and 2 IT pairs during training the IT+IF+IB network. For comparison purposes, the same sampling was adopted when the MMD was not used. When testing the IT+IF+IB network, with or without the MMD regulariser, results were computed on all the intra-patient pairs in the holdout set.

3.4   Comparison of networks with an iterative algorithm

To test an iterative intensity-based registration, the widely-adopted nonrigid method using B-splines was tested on the same holdout images. For comparison purposes, the sum-of-square difference in intensity values was used as similarity measure with all other parameters set to default in the NiftyReg [9] package. The reported registration performance was to demonstrate its feasibility. Although the default configuration is unlikely to perform optimally, tuning this method is considered out of scope for the current study.

3.5   Results

Sampling Strategies Networks with different training data sampling methods, described in Section 3.3, are summarised in Table 1. Adding time-backward and inter-subject image pairs in the training significantly improved the performance, both in network generalisability and registration performance. For example, the TREs decreased from 6.456 ± 6.822 mm to 5.801 ± 7.104 mm (p-value=0.0004), when adding time-backward data, to 5.482 ± 5.589 mm (p-value=0.0332), when further inter-subject data was added in training. The same conclusion was also observed in DSCs and gland CDs.

Regularisation effect Table 2 summarises the comparison between networks trained with MMD regularisation (Eq. 3) and without (Eq. 1). Although improved results were consistently observed in expected DSCs and TREs, statistical significance was not found on holdout set. However, we report a statistically significant improvement on the validation set, due to the introduction

Table 1. Registration performance with NiftyReg and networks with different sampling strategies. *See text for details including the explnation of the inferior NiftyReg result.

Methods	DSC	CD	MSE	TRE
NiftyReg*	0.270±0.304	22.869±11.761mm	0.041±0.019	21.147±15.841mm
w/o registration	0.701±0.097	8.842±4.067mm	0.051±0.090	10.736±7.718mm
IF	0.861±0.042	2.910±1.756mm	0.049±0.097	6.456±6.822mm
IF+IB	0.870±0.033	2.257±1.503mm	0.043±0.096	5.801±7.104mm
IT+IF+IB	0.885±0.024	2.132±0.951mm	0.053±0.014	5.482±5.589mm

of the MMD regulariser, e.g. p-values=0.016 in lowered MSEs. This was probably limited by the small holdout set and the under-optimised hyperparameters specifically for the MMD regulariser in current study.

Table 2. The comparison between networks trained with and without MMD regularisation. See text for more details.

Method	Test	MMD	DSC	CD	MSE	TRE
IF+IB	IF	×	0.870±0.033	2.257±1.503mm	0.043±0.096	5.801±7.104mm
IF+IB	IF	√	0.876±0.027	2.300±1.007mm	0.042±0.094	5.847±6.360mm
IT+IF+IB	IF+IB	×	0.890±0.019	1.928±0.797mm	0.048±0.010	5.638±6.021mm
IT+IF+IB	IF+IB	√	0.893±0.023	1.527±0.832mm	0.049±0.010	6.048±6.721mm

Registration performance As shown in Table 1, the proposed registration network, with any training data sampling strategies, produced significantly lower TREs on holdout data than the TREs before registration or that from default NiftyReg algorithms. With the overall inferior NiftyReg results summarised in the table, we report that the best-performed registration from NiftyReg achieved a comparable DSC of 0.81 and a TRE of 4.75 mm, better than the network-predicted. This provides an example of highly variable registration performance from an iterative algorithm, frequently encountered in our experiment. However, a comprehensive comparison is still needed to draw further conclusions. On average, the inference time was 0.76 seconds for the registration network, compared with 50.4 seconds for NiftyReg.

A case study for longitudinal analysis of prostate cancer Fig. 3 qualitatively illustrates a 60-year-old man from active surveillance with a baseline and three follow-up visits, subject to a biopsy of Gleason 3+3. The yellow arrows indicate the evolution of a marked adenomatous area within the left transitional zone. The registration network was able to track the changes of the suspicious regions between consecutive visits, in an automated, unbiased and consistent manner over 3 years. Ongoing research is focused on analysis using the registration-quantified changes.

4   Conclusion

For the first time, we have developed a deep-learning-based image registration method and validated the network using clinical longitudinal data from prostate
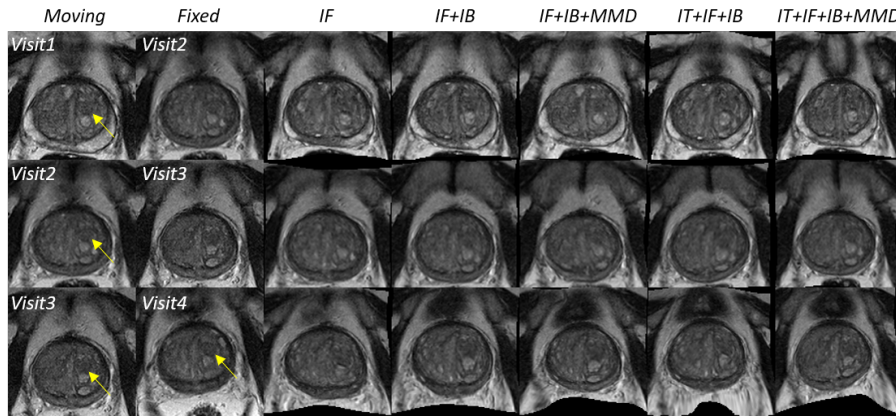
Fig. 3. Example registration results of a patient with 4 visits. The 1st and 2nd columns are the moving and fixed images. The remainder represent the network-warped images.

cancer active surveillance patients. We have also shown that adopting different training strategies significantly changes the network generalisability on holdout data.

Acknowledgment

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)
2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging **38**(8), 1788–1800 (2019)
3. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain separation networks. In: Advances in neural information processing systems. pp. 343–351 (2016)
4. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. Journal of Machine Learning Research **13**(Mar), 723–773 (2012)

5. Hu, S., Wei, L., Gao, Y., Guo, Y., Wu, G., Shen, D.: Learning-based deformable image registration for infant mr images in the first year of life. Medical physics **44**(1), 158–170 (2017)

6. Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C.M., Emberton, M., et al.: Weakly-supervised convolutional neural networks for multimodal image registration. Medical image analysis **49**, 1–13 (2018)

7. Kim, C.K., Park, B.K., Lee, H.M., Kim, S.S., Kim, E.: Mri techniques for prediction of local tumor progression after high-intensity focused ultrasonic ablation of prostate cancer. American Journal of Roentgenology **190**(5), 1180–1186 (2008)

8. Liao, S., Jia, H., Wu, G., Shen, D., Initiative, A.D.N., et al.: A novel framework for longitudinal atlas construction with groupwise registration of subject image sequences. NeuroImage **59**(2), 1275–1289 (2012)

9. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. Computer methods and programs in biomedicine **98**(3), 278–284 (2010)

10. Moore, C.M., Giganti, F., Albertsen, P., Allen, C., Bangma, C., Briganti, A., Carroll, P., Haider, M., Kasivisvanathan, V., Kirkham, A., et al.: Reporting magnetic resonance imaging in men on active surveillance for prostate cancer: the precise recommendations—a report of a european school of oncology task force. European urology **71**(4), 648–655 (2017)

11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)

12. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: application to breast mr images. IEEE transactions on medical imaging **18**(8), 712–721 (1999)

13. Schwartz, E., Jakab, A., Kasprian, G., Zöllei, L., Langs, G.: A locally linear method for enforcing temporal smoothness in serial image registration. In: International Workshop on Spatio-temporal Image Analysis for Longitudinal and Time-Series Image Data. pp. 13–24. Springer (2014)

14. Simpson, I.J., Woolrich, M., Groves, A.R., Schnabel, J.A.: Longitudinal brain mri analysis with uncertain registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 647–654. Springer (2011)

15. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014)

16. Vapnik, V.N.: An overview of statistical learning theory. IEEE transactions on neural networks **10**(5), 988–999 (1999)

17. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)