Improving Robustness of Deep Learning Based Knee MRI Segmentation: Mixup and Adversarial Domain Adaptation

Egor Panfilov[1] Aleksei Tiulpin[1,2] Stefan Klein[3] Miika T. Nieminen[1,2]
Simo Saarakkala[1,2]

[1]University of Oulu, Oulu, Finland [2]Oulu University Hospital, Oulu, Finland [3]Erasmus MC, Rotterdam, The Netherlands

egor.panfilov@oulu.fi

Abstract

Degeneration of articular cartilage (AC) is actively studied in knee osteoarthritis (OA) research via magnetic resonance imaging (MRI). Segmentation of AC tissues from MRI data is an essential step in quantification of their damage. Deep learning (DL) based methods have shown potential in this realm and are the current state-of-the-art, however, their robustness to heterogeneity of MRI acquisition settings remains an open problem. In this study, we investigated two modern regularization techniques – mixup and adversarial unsupervised domain adaptation (UDA) – to improve the robustness of DL-based knee cartilage segmentation to new MRI acquisition settings. Our validation setup included two datasets produced by different MRI scanners and using distinct data acquisition protocols. We assessed the robustness of automatic segmentation by comparing mixup and UDA approaches to a strong baseline method at different OA severity stages and, additionally, in relation to anatomical locations. Our results showed that for moderate changes in knee MRI data acquisition settings both approaches may provide notable improvements in the robustness, which are consistent for all stages of the disease and affect the clinically important areas of the knee joint. However, mixup may be considered as a recommended approach, since it is more computationally efficient and does not require additional data from the target acquisition setup.

1. Introduction

Knee osteoarthritis (OA) is the most common musculoskeletal disease in the world. OA is poorly understood and no disease-modifying treatment currently exists for it [43]. Magnetic resonance imaging (MRI) methods are commonly used to clinically study the structural changes within the knee joint and, specifically, in articular cartilage [27]. A variety of MRI acquisition protocols has been introduced, each tailored to visualize specific tissues of interest or measure particular tissue properties [1, 2]. Moreover, there is a large number of MR scanner models available on the market, with major differences in hardware and reconstruction software. As a result, MR images qualitatively vary from institution to institution, from study to study, and from dataset to dataset.

Since OA is a long-term and complex disease, large longitudinal studies have been carried out to investigate onset and progression of OA. Currently, one of the major areas of interest is assessment of compositional and morphological changes in articular cartilage tissues [15]. In order to perform these analyses from MRI, the tissues need to be segmented. However, manual delineation of cartilage tissues is time-consuming, prone to high intra- and inter-rater variability [34], and challenging due to the large size of datasets and the aforementioned issues related to data heterogeneity. Consequently, there is a clear need for automatic methods for knee cartilage segmentation, which are accurate and robust to variations in data acquisition setting.

Recently, in OA and other fields of medicine, deep learning (DL) methods have become the new state-of-the-art in computer-aided diagnosis [39, 38, 45, 40, 16, 17]. Latest advances in automatic segmentation methods, in particular DL-based, have demonstrated promising results in knee tissue segmentation [28, 35, 3, 12, 36, 6]. Such methods produce accurate and consistent results, but they often lack evaluation on independent datasets and, therefore, are potentially prone to large variations in the input data characteristics. The issue originates from the fact that supervised DL-based algorithms, when trained on medical imaging datasets that are often limited in size and diversity, tend to incorporate dataset bias and fail to generalize to new domains [20].

In this paper, we focus on regularization of DL-based methods for knee cartilage segmentation from MRI, and investigate two state-of-the-art approaches to improve the generalization to new data. The contributions of this study are the following:

- We introduce an efficient and accurate DL-based baseline method for knee cartilage segmentation that performs comparably or improves on the previous state-of-the-art.
- We investigate the use of an end-to-end unsupervised domain adaptation (UDA) approach for knee MRI segmentation, and show how both labelled and unlabelled data can be leveraged within the same segmentation framework.
- We explore the use of data augmentation via mixup in the considered semantic segmentation problem, and report its effectiveness in multiple setups.
- We validate the baseline method and its modifications with mixup and UDA on an independent test set and demonstrate the improvements in model robustness. We also provide a detailed analysis of the results and examine the performance of the methods in relation to the anatomical locations that are the most clinically relevant (e.g. weight bearing areas of the knee joint).
- Finally, we make our source code and the pre-trained models publicly available.

2. Related Work

Due to the wide adoption of MR imaging methods, semi-automatic and automatic knee cartilage segmentation from MRI has been studied already for several decades, with more focus recently on purely automatic methods [33, 24]. However, despite the availability of large imaging cohorts, such as Osteoarthritis Initiative (OAI) [30], large-scale analysis of such data in scope of OA research remains extremely challenging due to the lack of annotations. Same applies to the datasets from numerous hospitals, which are typically less standardized and the annotations are even more sparse and of lower quality.

Several recent studies [28, 26, 3, 12] have shown that specifically DL-based approaches for knee cartilage and meniscus segmentation can achieve accuracy close to the human level and superior to the conventional atlas-based methods [10]. However, no validation of those DL-based methods on independent datasets acquired in various hospitals has yet been conducted. Therefore, the general applicability of all the previously published DL-based cartilage segmentation methods remains unclear.

To tackle the robustness-related issues in modern deep neural networks, a wide range of techniques of different complexity has been proposed [23]. Their effectiveness in the specific tasks and domains, however, is still to be practically investigated.

One of the recent effective techniques to improve model generalization and reduce memorization of the training data was mixup [48]. The idea of mixup was to use a convex combination of the inputs and the targets to augment the training data with such interpolated examples. Mixup has been applied in several image classification problems and has shown to notably reduce the overfitting and stabilize the convergence of models [47, 19]. Nonetheless, the applicability of the technique and its performance in semantic segmentation problems remains unclarified, and very few studies investigated this topic [14, 5]. Our goal was to adapt the technique to knee cartilage segmentation problem and evaluate its performance in different settings.

Another approach that has attracted great interest during the recent years is domain adaptation (DA) [11, 9]. A great number of the popular DA techniques is based on the following hypothesis: in order to have a good generalization for any machine learning method, the representations of data samples (including the samples from different domains or datasets) have to share a large common subspace or be somewhat aligned. Ganin *et al.* [18] and Tzeng *et al.* [42] were among the first to discover how this alignment can be applied to DL-based models, and moreover, how to perform it in a semi-supervised way, such that both labelled and unlabelled data from different domains can be incorporated into the training process. The framework was called Unsupervised Domain Adaptation (UDA) and its potential for the development of robust models has been shown in various applications.

In medical imaging, UDA has been studied in several fields for which multiple diverse datasets are publicly available: brain MRI [20, 4], chest X-ray [7], cardiac MRI-CT [13] and others. However, very few studies investigated the use of UDA techniques in knee MRI domain and, more specifically, knee cartilage segmentation. Joint multitask learning of deep segmentation and registration networks was suggested by Xu *et al.* [46]. Liu [25] proposed to train joint segmentation and cycle-consistent image-to-image translation between the labelled and unlabelled domains. Both of the approaches, however, are rather complicated and depend on the performance of an auxiliary task – registration or image-to-image translation respectively.

In this work, we explored UDA via cross-domain alignment of deep representation spaces. The chosen method has only moderate computational costs, can be easily scaled to larger number of domains and extended with other regularization techniques.

3. Materials and Methods

3.1. Problem Statement

Our goal was to assess how different regularization techniques, namely, mixup and UDA, perform in combination with a strong baseline method yielding state-of-the-art results in knee cartilage segmentation task. To assess the gen-

eralization, we validated all our methods on independent data. For UDA, we assumed that the unlabelled data from a dataset similar to the test dataset is available during training. In total, our setup included three datasets: two datasets from different MRI scanners and data acquisition protocols (Datasets A and B), and a third dataset (Dataset C) acquired similarly to Dataset B, but in an independent clinical study (see Figure 1). Train subset of Dataset A and whole Dataset B were used for development of the approaches. Test subset of Dataset A and whole Dataset C were used for evaluation purposes.

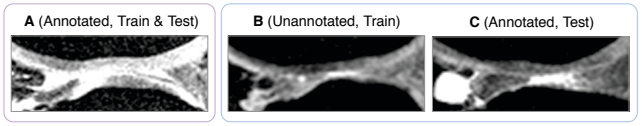


Figure 1. Examples of MRI images from Datasets A, B, and C. Here, we show only the tibiofemoral areas, which enclose femoral cartilage, tibial cartilage, and menisci.

Let $\mathbf{X}^a$ be a mini-batch of image samples from the annotated dataset, $\mathbf{Y}^a$ - corresponding reference annotations, $\mathbf{X}^b$ - a mini-batch of image samples from the unannotated dataset. S is a model that takes as input a mini-batch of images $\mathbf{X}$ (either $\mathbf{X}^a$ or $\mathbf{X}^b$) and produces the segmentation masks $\hat{\mathbf{Y}}$. In all the experiments the models were trained to perform the segmentation slice-wise (i.e. in 2D).

3.2. Baseline Method

Our baseline approach (Figure 2a) was based on U-Net [32]. Similarly to [37] we used 24 filters in the first convolutional block and doubled the number of filters at each depth level. The total model depth was set to 6. In the expanding path we used bilinear upsampling instead of 2×2 up-convolutions. As a results of an extensive experimental search we found that such model parameters yielded the best performance in the considered task. The network was trained to produce 5 mutually exclusive segmentation masks: no cartilage, femoral cartilage (FC), tibial cartilage (TC), patellar cartilage (PC), and menisci (M). For training we used multi-class cross-entropy (MCE) loss, which was calculated between the randomly sampled masks $\mathbf{Y}^a$ and the model predictions $\hat{\mathbf{Y}}$ produced from the corresponding $\mathbf{X}^a$.

3.3. Regularization Techniques

Mixup. We followed the original implementation[1] and adapted mixup to our problem. Here, the samples from the mini-batch and its permuted version were paired. Then, the virtual inputs were constructed from the pairs and passed through the network S:

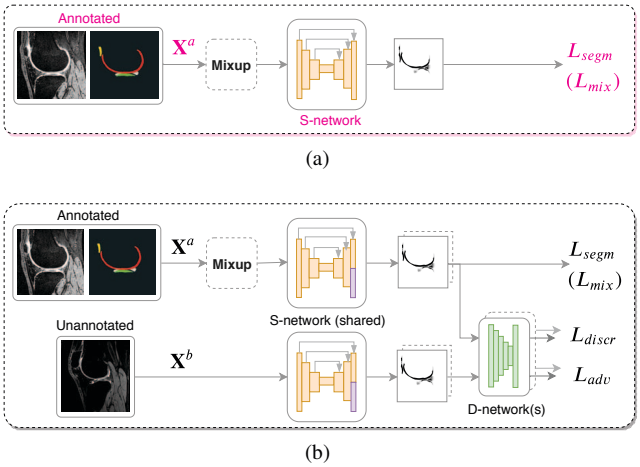


Figure 2. Schematic view of our approaches – without (a) and with (b) UDA. Mixup, if used, is applied only during the training. In UDA setting (b), the S- and D- networks are trained in an adversarial manner. During the testing, only the S-network is utilized.

$$\lambda \sim \mathrm{Beta}(\alpha, \alpha) \tag{1}$$

$$\mathbf{X}_{perm}, \mathbf{Y}_{perm} = \mathrm{permute}(\mathbf{X}, \mathbf{Y}) \tag{2}$$

$$\mathbf{X}_{mix} = \lambda \mathbf{X} + (1 - \lambda)\mathbf{X}_{perm} \tag{3}$$

$$\hat{\mathbf{Y}} = S(\mathbf{X}_{mix}), \tag{4}$$

where parameter α configured the augmentation strength. The segmentation loss for the settings with mixup was defined using the model predictions and the corresponding pairs of reference annotations:

$$L_{mix} = \lambda L_{segm}(\hat{\mathbf{Y}}, \mathbf{Y}) + (1 - \lambda) L_{segm}(\hat{\mathbf{Y}}, \mathbf{Y}_{perm}) \tag{5}$$

Unsupervised Domain Adaptation. Differently from mixup, UDA allows to utilize both labelled and unlabelled data. In this study we adapted the method from [41] (see Figure 2b). In particular, the segmentation model here was trained to produce the representations that do not incorporate dataset-specific biases. This was done by aligning the output and, optionally, penultimate representation spaces of the network across the datasets. For this, two networks – a segmentation network S and a discriminator network D – were trained in an adversarial manner. Network S was trained to predict the segmentation masks of the cartilage tissues and menisci by minimizing a sum of losses:

$$\begin{aligned} &\gamma_{segm} L_{segm}(\mathbf{S}, \mathbf{X}^a, \mathbf{Y}^a) + \\ &\gamma_{adv} L_{adv}(\mathbf{D}, \mathbf{S}, \mathbf{X}^b) \rightarrow \min_{\mathbf{S}}, \end{aligned} \tag{6}$$

where $\mathbf{S}$ and $\mathbf{D}$ are parameters of S and D, and L_{adv} is an adversarial loss:

$$L_{adv}(\mathbf{D}, \mathbf{S}, \mathbf{X}^b) = \mathrm{BCE}(\mathbf{0}, D(S(\mathbf{X}^b))), \tag{7}$$

[1] https://github.com/facebookresearch/mixup-cifar10

where BCE is the binary cross-entropy and $\mathbf{0}$ is a matrix of all zeros having the same shape as $\mathbf{X}^b$.

Network D, which enforced the domain-agnostic behaviour of S, acted as a domain discriminator and was trained as follows:

$$L_{discr}(\mathbf{D}, \mathbf{S}, \mathbf{X}^a, \mathbf{X}^b) = L^0_{discr} + L^1_{discr} \rightarrow \min_{\mathbf{D}}, \quad (8)$$

where

$$L^0_{discr}(\mathbf{D}, \mathbf{S}, \mathbf{X}^a) = \text{BCE}(\mathbf{0}, D(S(\mathbf{X}^a))) \quad (9)$$

$$L^1_{discr}(\mathbf{D}, \mathbf{S}, \mathbf{X}^b) = \text{BCE}(\mathbf{1}, D(S(\mathbf{X}^b))) \quad (10)$$

Here, $\mathbf{1}$ is a matrix of all ones having the same shape as $\mathbf{X}^a$ and $\mathbf{X}^b$. The discriminator was built from 5 convolutional layers (with 64, 128, 256, 512, and 1 filters) alternated by 4 LeakyReLU and followed by bilinear upsampling to the input image shape. Hereinafter we call the described approach UDA1.

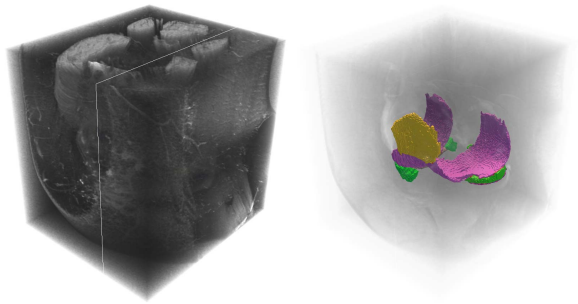
Additionally, we evaluated the extension of the method, where the adaptation is also applied to the penultimate decoder block of S. We hypothesized that adaptation at two levels can yield better alignment of the representations and compensate for the potential spatial shift between the domain in the output space. An ASPP block [8] was added on top of the last decoder block, and its output was bilinearly upscaled to the dimensions of the S output. The upscaled activations were used to compute the auxiliary segmentation loss and also as an input to the second discriminator. This discriminator had the same architecture as the first one, was trained following the same procedure, and contributed to the minimization criteria in Equation 6 with a smaller γ_{adv}. Further implementation details can be found in [41, Section 4.2]. This approach with the adaptation of two representation spaces is further referred to as UDA2.

3.4. Evaluation

To assess the segmentation results, we used both planar (slice-wise) and volumetric (scan-wise) Dice similarity coefficients (DSCs). To examine the localization of the segmentation errors, we registered the scans such that the lateral and medial sides of the knees are oriented identically, mirroring the scans where needed. Subsequently, we computed the distribution of planar DSCs for each sagittal slice index over all registered scans in the test sets. The 95% confidence intervals were estimated using bootstrapping.

3.5. Data

Overview. As mentioned previously, our setup included three different datasets of knee sagittal 3D double echo steady state (DESS) MRI (see Figure 3a): Datasets A, B, and C. Datasets B and C were collected at our hospital using the same scanner. Dataset A contained the data from a different scanner and a distinct imaging protocol (see examples in Figure 1).



(a) (b)

Figure 3. DESS MRI scan (a) and the annotations of knee cartilage and meniscal tissues (b), both rescaled to isotropic resolution. White lines in (a) indicate the orientation of sagittal slices.

All the datasets were supplemented with the Kellgren-Lawrence (KL) scores [21], derived from the radiographs associated to each subject. KL grading is the gold standard system for radiographic OA severity assessment. According to this scale, OA severity is graded from KL0 (no visible OA) to KL4 (end stage OA).

Dataset A. The data were obtained from the Osteoarthritis Initiative (OAI, http://www.oai.ucsf.edu/) database. Dataset A included 88 subjects from the OAI baseline and 12-month follow-up examinations: 0/4/59/101/12 scans with KL0/1/2/3/4 respectively. The data were acquired with 3T Siemens MAGNETOM Trio scanners and quadrature transmit-receive knee coils (USA Instruments, Aurora, OH, USA). Sagittal DESS sequence was used (160 slices; voxel size: $0.37 \times 0.37 \times 0.7mm$, matrix: 384×384, field of view (FOV): $140mm$; repetition time (TR): $16.3ms$, echo time (TE): $4.7ms$, flip angle: $25°$). Manual annotations were available for femoral, tibial, and patellar cartilage tissues, and also menisci (Figure 3b).

Dataset B. The dataset included 108 subjects: 14/42/28/22/2 scans with KL0/1/2/3/4 respectively (ClinicalTrials.gov Identifier: NCT02937064). The knees were imaged with 3T Siemens MAGNETOM Skyra scanner using a 15-channel transmit-receive knee coil (QED, Mayfield Village, OH, USA). Sagittal DESS sequence was used (160 slices; voxel size: $0.59 \times 0.59 \times 0.6mm$, matrix: 256×256, FOV: $150mm$; TR: $14.1ms$, TE: $5ms$, flip angle: $25°$). No reference annotations for the tissues were available.

Dataset C. The dataset [31] included 44 subjects: 0/16/13/15/0 scans with KL0/1/2/3/4 respectively. The

scanner and the data acquisition protocol were the same as for Dataset B. The annotations were produced by the research group and consisted of the segmentation masks for femoral and tibial cartilage tissues.

3.6. Implementation details

Data Pre-processing. In our experiments, firstly, all the data were rescaled to the pixel size of 0.37×0.37 mm. Secondly, the intensity histograms of the images were truncated (from 10^{th} to 99^{th} percentiles). Finally, the central image regions of 300×300 pixels were cropped and used for training and evaluation.

To avoid overfitting, we used the data augmentations during the training. In particular, we used random left-right flipping, gamma correction, randomly applied downscaling followed by upscaling, and also bilateral filtering.

Training. We trained and evaluated the described regularization techniques in several settings. The baseline segmentation network was trained from scratch as-is, with mixup (with and without weight decay), with UDA1, with UDA1 and mixup (without weight decay for S), and with UDA2.

Dataset A was randomly split into the train and the test subsets using stratification by subject ID and balancing with respect to KL grading scores. Dataset B was used solely for training and validation, Dataset C – solely for testing. All the methods were trained using 5-fold cross-validation following the stratification strategy described above. For the methods with UDA, Dataset B was similarly divided into 5 folds and randomly combined with the folds of Dataset A. Therefore, for each of the experiment we train 5 models. During the testing the predictions of these models were averaged.

In all of the experiments we used Adam [22], one independent optimizer per network, depending on the setting. For all experiments with weight decay, the regularization constant was set to $5 \cdot 10^{-5}$. Parameter α in mixup, which is, typically, in the range from 0 to 1, was set to 0.7.

The baseline method with and without mixup was trained for 50 epochs starting with the learning rate (LR) of 10^{-3} (reduced to 10^{-4} at the 30^{th} epoch). All the variants with domain adaptation were trained for 30 epochs. Initial learning rates were set to 10^{-4} for S and 4×10^{-5} for D, both reduced by a factor of 10 at the 25^{th} epoch. S and D were updated alternately on each batch of images. For the experiments with mixup, Equation 5 was used as a segmentation loss instead of MCE loss. In UDA1 experiments γ_{segm} and γ_{adv} were set to 1 and 10^{-3} respectively in order to prioritize the segmentation task in the adversarial training. In UDA2 experiment, in addition to the above, auxiliary γ_{segm} and γ_{adv} were set to 10^{-1} and 2×10^{-4} respectively, following the original publication. In the experiment with combination of mixup and UDA approaches, we applied mixup only for segmentation task. Additional forward pass of S with unmixed data and without accumulation of gradients was performed to produce the input for D. Otherwise, the UDA architecture was kept the same.

Testing of the methods was performed on the test subset of Dataset A and whole Dataset C. To conduct the experiments we used NVIDIA 2080 Ti GPU and PyTorch [29].

4. Results

Baseline. We compared our baseline method to the published state-of-the-art approaches in knee cartilage and menisci segmentation (see Table 1). Our baseline method performed either more accurately or on par with others depending on the tissue. Even though it was not designed to separate lateral and medial tibial cartilage tissues and menisci as in [35, 3, 36], it provided other advantages, namely, it was faster in training and inference, more lightweight, and produced masks for all the considered tissues simultaneously.

On Dataset A the method reached $0.907\,(0.019)$ for FC and $0.897\,(0.028)$ for TC, however, on Dataset C the scores for the respective tissues were $0.791\,(0.033)$ and $0.629\,(0.054)$. Such discrepancy in the scores was, presumably, caused by the several factors: lack of model robustness, which resulted in biased and noisy segmentations (see Figure 5), and lower original resolution of images and annotations in Dataset C, which made the segmentation more challenging and increased the cost of annotation errors.

Mixup. We found that applying mixup lead to a minor underfitting on Dataset A (see Table 2), yet the generalization had increased ($0.804\,(0.031)$ for FC, $0.652\,(0.051)$ for TC on Dataset C). Such phenomena was also reported for object classification in Verma *et al.* [44]. However, since mixup itself is a strong regularizer, we hypothesized that avoiding the use of weight decay could address the undefitting. In this new setting our model largely recovered the scores on Dataset A and further improved the performance for FC on Dataset C ($0.819\,(0.025)$ for FC, $0.647\,(0.049)$ for TC).

Unsupervised Domain Adaptation. An approach with UDA1 on Dataset C yielded comparable DSCs ($0.815\,(0.025)$ for FC, $0.647\,(0.049)$ for TC) to the best mixup setting, however, the scores on Dataset A were lower (see Table 2). Performing representation alignment at the multiple layers of the network (UDA2) improved on top of UDA1 for Dataset A. However, on Dataset C the performance increased only for FC, while it became worse for TC ($0.821\,(0.025)$ for FC, $0.640\,(0.055)$ for TC). What concerns the efficiency, the computational costs for training UDA approaches were up to three times higher compared to mixup.

Method	FC	TC		PC	M	
		medial	*lateral*		*medial*	*lateral*
Norman et al. [28]	0.867(0.032)	0.777(0.029)	0.799(0.036)	0.767(0.091)	0.731(0.054)	0.812(0.030)
Tack et al. [35]	-	-	-	-	0.838(0.061)	0.889(0.024)
Ambellan et al. [3]	0.894(0.024)	0.861(0.053)	0.904(0.024)	-	-	-
Desai et al. [12]	0.89 (0.02)	-	-	-	-	-
Tack et al. [36]	-	0.880(0.046)	0.913(0.023)	-	-	-
Chaudhari et al. [6]	0.902(0.017)	-	-	-	-	-
Ours (baseline method)	0.907(0.019)	0.897(0.028)		0.871(0.046)	0.863(0.034)	

Table 1. Comparison of our baseline to the previously published methods on Dataset A. Numbers are the means and standard deviations of volumetric DSCs. The scores are given for reference and should be compared carefully. [28, 12, 6] used slightly different train/validation/test splits. [6] performed segmentation in 3D. [35, 3, 36] used multi-stage pipelines (2D segmentation, statistical shape modelling, 3D refinement), 2-fold cross validation, and reported the results for 2 examinations (we present only the highest scores).

Method	Dataset A				Dataset C	
	FC	TC	PC	M	FC	TC
Baseline	**0.907(0.019)**	**0.897(0.028)**	**0.871(0.046)**	**0.863(0.034)**	0.791(0.033)	0.629(0.054)
+ mixup	0.903(0.019)	0.892(0.031)	0.865(0.054)	0.852(0.035)	0.804(0.031)	**0.652(0.051)**
+ mixup - WD	**0.907(0.019)**	0.896(0.028)	0.864(0.054)	0.861(0.033)	0.819(0.025)	0.647(0.049)
+ UDA1	0.896(0.023)	0.887(0.031)	0.852(0.064)	0.851(0.035)	0.815(0.025)	0.647(0.049)
+ UDA2	0.901(0.021)	0.892(0.031)	0.861(0.060)	0.856(0.035)	**0.821(0.025)**	0.640(0.055)
+ mixup - WD + UDA1	0.895(0.023)	0.886(0.027)	0.846(0.066)	0.849(0.034)	0.810(0.026)	0.635(0.052)

Table 2. Regularization approaches evaluated tissue-wise on two datasets. Numbers are the means and standard deviations of volumetric DSCs. The best score for each tissue is highlighted in bold, the second best - is underlined. "- WD" indicates the experiments without weight decay.

Combined Approach. We experimented with combining mixup and UDA approaches. Here, we took UDA1 setting and applied mixup to the supervised segmentation task. Otherwise, the architecture was kept the same, including the input to the generator. The approach with both mixup and UDA1 showed the worst performance among others, yet still showed marginally better DSCs on Dataset C compared to the baseline.

Detailed Analysis. As previously said, cartilage tissues degenerate over the progression of OA. Lesions start to appear, cartilage is getting worn out and, therefore, it becomes challenging to segment. To evaluate the performance of the methods in relation to OA severity (from doubtful OA to end-stage OA), we computed the volumetric DSCs over the test sets KL-grade-wise. Here, for the sake of brevity, we compared only the baseline approach, the best of mixup, and the best of UDA. The results are presented in Table 3.

The detailed analysis showed that both modifications (mixup and UDA2) yielded similar significant improvements on Dataset C for most of the cases, while the approach with mixup better maintained the performance on Dataset A. The results were further analyzed with respect to anatomical location by following the approach described in Section 3.4. Illustrated in Figure 4, the results indicated that both mixup and UDA2 improved the segmentation accuracy for FC, with higher increase in the weight bearing areas of femoral condyles. For TC, the improvements were concentrated near the tibial plateaus, mainly located on the medial side.

Several challenging examples of MR images, with the corresponding annotations and the segmentation masks produced by the methods, are presented in Figure 5. In these cases, the baseline and UDA2 approaches tended to over-segment the tissues, while the baseline also produced shifted segmentations on Dataset C. The images also highlight inaccuracies of the reference segmentations and common limitations of the approaches.

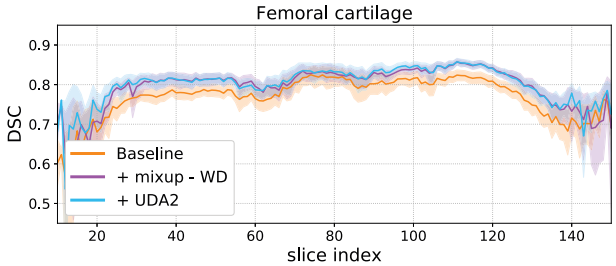
5. Conclusions

In this study, we investigated the use of mixup and adversarial domain adaptation for DL-based knee tissue segmentation from MRI. We showed that the segmentation model trained from scratch with limited data lacked generalization and performed worse on the dataset that had different resolution and contrast. Strong regularization techniques, namely, supervised mixup and UDA, helped to partially alleviate the issue and make the model more robust. We analyzed the baseline and the best performing approaches in relation to anatomical locations and different stages of OA, and showed that the improvements over the baseline are consistent and clinically relevant.

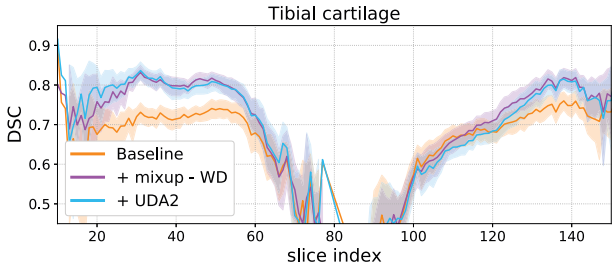
This paper is the first to address the challenge of robustness in knee MRI segmentation in an end-to-end manner

Method	KL	Dataset A #	FC	TC	PC	M	Dataset C #	FC	TC
Baseline	1	-	-	-	-	-	16	0.785(0.041)	0.667(0.038)
	2	11	0.920(0.015)	0.921(0.010)	0.860(0.061)	0.873(0.047)	13	0.794(0.031)	0.602(0.057)
	3	21	0.904(0.019)	0.891(0.026)	0.875(0.040)	0.860(0.025)	15	0.794(0.024)	0.612(0.043)
	4	4	0.892(0.003)	0.861(0.019)	0.882(0.015)	0.854(0.029)	-	-	-
	all	36	0.907(0.019)	0.897(0.028)	0.871(0.046)	0.863(0.034)	44	0.791(0.033)	0.629(0.054)
+ mixup - WD	1	-	-	-	-	-	16	**0.826(0.024)**	0.674(0.038)
	2	11	0.921(0.013)	0.922(0.007)	0.860(0.060)	0.872(0.043)	13	**0.821(0.023)**	**0.636(0.048)**
	3	21	0.903(0.019)	0.890(0.026)	**0.863(0.055)**	0.857(0.026)	15	**0.811(0.026)**	0.627(0.047)
	4	4	0.889(0.002)	0.861(0.016)	0.877(0.015)	0.856(0.027)	-	-	-
	all	36	0.907(0.019)	0.896(0.028)	**0.864(0.054)**	0.861(0.033)	44	**0.819(0.025)**	**0.647(0.049)**
+ UDA2	1	-	-	-	-	-	16	**0.827(0.024)**	0.669(0.050)
	2	11	**0.915(0.015)**	0.918(0.007)	0.851(0.067)	**0.867(0.045)**	13	**0.822(0.028)**	**0.632(0.059)**
	3	21	**0.898(0.021)**	**0.885(0.030)**	**0.863(0.061)**	**0.852(0.027)**	15	**0.815(0.022)**	0.617(0.043)
	4	4	0.883(0.008)	0.856(0.017)	0.875(0.019)	0.840(0.032)	-	-	-
	all	36	**0.901(0.021)**	**0.892(0.031)**	**0.861(0.060)**	**0.856(0.035)**	44	**0.821(0.025)**	**0.640(0.055)**

Table 3. Comparison between the baseline and the best performing approaches. Here, means and standard deviations of volumetric DSCs are presented for the subject groups of specific KL-grades (1-4) and for the full test sets. # shows the number of scans in the specific group. Statistically significant differences to the baseline method ($p < 0.05$ with two-sided Wilcoxon signed-rank test) are highlighted in bold. For Dataset A all the differences are either negative or insignificant, for Dataset C - either positive or insignificant.



(a)



(b)

Figure 4. Distributions of the planar DSCs computed slice-wise (from 0th to 159th slice, medial to lateral, respectively). Solid lines indicate the distribution means, bright bands – the 95% confidence intervals. Slices approx. 20-60 and 100-140 correspond to the locations of the medial and lateral femoral condyles (i.e. weight-bearing areas of the joint). Slices approx. 60-100 enclose the intercondylar notch and, therefore, are of less clinical interest.

using DL. On the test set derived from OAI, our model yielded state-of-the-art segmentation results for patellar cartilage and similar to other works results for other tissues, and allowed to segment all the cartilage and meniscal tissues simultaneously.

Despite the state-of-the-art results, our study has still some limitations. In particular, we considered only 2D segmentation approach. Due to several factors, such as complex cartilage geometry, partial volume effect, lack of contextual information in 2D, and imperfect and inconsistent annotations, most of the segmentation errors produced by our methods were located on the tissue surfaces or in the slices tangential to the surfaces. Volumetric methods could potentially alleviate some of those issues and provide more accuracy and shape consistency. However, the comparison done to the previous studies [36, 3, 35, 6] showed similar performance in terms of DSCs.

Another limitation of this study is that more complex applications of mixup in UDA were not investigated. We believe that more experimental work in that direction can potentially lead to higher results. Future studies on knee MRI segmentation or medical image segmentation could further explore the potential of this idea. Besides the mentioned limitations, we acknowledge that a more comprehensive framework for assessment of regularization methods should be considered. However, in knee MRI field there is a lack of public datasets available for experiments. Our future studies will consider consolidation of various datasets in order to perform more thorough investigations, including different scanner manufacturers and diverse MRI sequences.

From the methodological point of view, our study demonstrated that for, at least, a moderate range of image variations, mixup and UDA may similarly improve the robustness of medical image segmentation. However, UDA approach is computationally heavy and difficult to train due to the need of careful hyper-parameter tuning. Besides

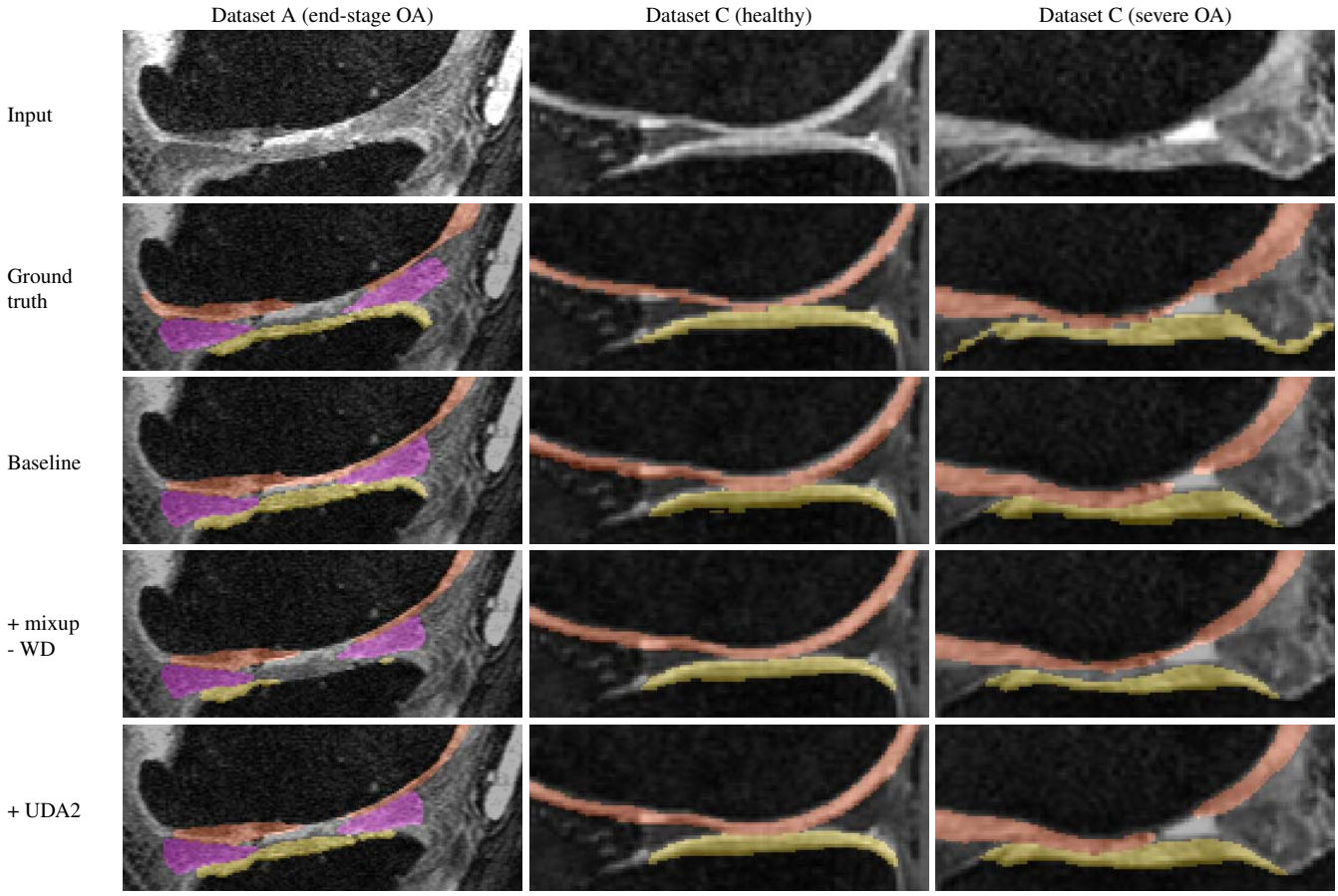


Figure 5. Example images of tibiofemoral contact zones from Datasets A and C, respective annotations, and the segmentation masks produced by the baseline and the regularized approaches. Visual differences between the datasets can be observed. Colors highlight cartilage tissues: orange – femoral, yellow – tibial, purple – menisci. Patellar cartilage was not presented in the considered Dataset A slice, for that reason patellofemoral zone is not shown.

that, our experiments showed that UDA may significantly worsen the DSCs in the source domain (Dataset A). Therefore, we think that mixup and other regularization techniques should be preferred when aiming for robust medical image segmentation using DL.

To conclude, we believe that our results will promote wider adoption of DL-based methods in OA research community and facilitate further work on development of robust segmentation methods for knee MRI. In MRI domain, such methods may become a powerful tool to leverage large and diverse imaging cohorts without available annotations and drastically speed up and improve the medical research. For instance, one important application area – disease modifying drugs development for OA – can directly benefit from reliable segmentations. To facilitate further knee MRI segmentation research, our source codes and pre-trained models are made publicly available: https://github.com/MIPT-Oulu/RobustCartilageSegmentation.

Acknowledgements

The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the National Institutes of Health, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. Private sector funding for the OAI is managed by the Foundation for the National Institutes of Health. This manuscript was prepared using an OAI public use data set and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners.

The authors would like to acknowledge the following funding sources: strategic funding of University of Oulu (Infotech Oulu), Sigrid Juselius foundation, and KAUTE foundation, Finland.

References

[1] O. M. Abdulaal et al. 3T MRI of the knee with optimised isotropic 3d sequences: Accurate delineation of intra-articular pathology without prolonged acquisition times. *European radiology*, 27(11):4563–4570, 2017. 1

[2] F. Altahawi and N. Subhas. 3d MRI in musculoskeletal imaging: Current and future applications. *Current Radiology Reports*, 6(8):27, 2018. 1

[3] F. Ambellan, A. Tack, M. Ehlke, and S. Zachow. Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the osteoarthritis initiative. *Medical Image Analysis*, 52(2):109–118, 2019. 1, 2, 5, 6, 7

[4] K. Chaitanya, N. Karani, C. Baumgartner, and E. Konukoglu. Semi-supervised and task-driven data augmentation. *arXiv e-prints*, page arXiv:1902.05396, Feb 2019. 2

[5] K. Chaitanya, N. Karani, C. F. Baumgartner, A. Becker, O. Donati, and E. Konukoglu. Semi-supervised and task-driven data augmentation. In *International Conference on Information Processing in Medical Imaging*, pages 29–41. Springer, 2019. 2

[6] A. S. Chaudhari et al. Utility of deep learning super-resolution in the context of osteoarthritis MRI biomarkers. *Journal of Magnetic Resonance Imaging*, 2019. 1, 6, 7

[7] C. Chen, Q. Dou, H. Chen, and P.-A. Heng. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 143–151. Springer, 2018. 2

[8] L.-C. Chen et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 4

[9] S. Chopra, S. Balakrishnan, and R. Gopalan. Dlid: Deep learning for domain adaptation by interpolating between domains. In *ICML workshop on challenges in representation learning*, 2013. 2

[10] E. B. Dam, M. Lillholm, J. Marques, and M. Nielsen. Automatic segmentation of high- and low-field knee MRIs using knee image quantification with data from the osteoarthritis initiative. *Journal of Medical Imaging*, 2(2):024001, apr 2015. 2

[11] H. Daume III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of artificial Intelligence research*, 26:101–126, 2006. 2

[12] A. D. Desai, G. E. Gold, B. A. Hargreaves, and A. S. Chaudhari. Technical considerations for semantic segmentation in MRI using convolutional neural networks. *arXiv preprint arXiv:1902.01977*, 2019. 1, 2, 6

[13] Q. Dou et al. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. *arXiv preprint arXiv:1804.10916*, 2018. 2

[14] Z. Eaton-Rosen, F. Bragman, S. Ourselin, and M. J. Cardoso. Improving data augmentation for medical image segmentation. MIDL 2018 submission at https://openreview.net/forum?id=rkBBChjiG, 2018. 2

[15] C. A. Emery, J. L. Whittaker, A. Mahmoudian, L. S. Lohmander, E. M. Roos, K. L. Bennell, C. M. Toomey, R. A. Reimer, D. Thompson, J. L. Ronsky, et al. Establishing outcome measures in early knee osteoarthritis. *Nature Reviews Rheumatology*, page 1, 2019. 1

[16] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017. 1

[17] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24, 2019. 1

[18] Y. Ganin et al. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 2

[19] H. Guo, Y. Mao, and R. Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3714–3722, 2019. 2

[20] N. Karani, K. Chaitanya, C. Baumgartner, and E. Konukoglu. A lifelong learning approach to brain MR segmentation across scanners and protocols. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–484. Springer, 2018. 1, 2

[21] J. Kellgren and J. Lawrence. Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases*, 16(4):494, 1957. 4

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] J. Kukačka, V. Golkov, and D. Cremers. Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*, 2017. 2

[24] D. Kumar, A. Gandhamal, S. Talbar, and A. F. M. Hani. Knee articular cartilage segmentation from MR images: A review. *ACM Computing Surveys (CSUR)*, 51(5):97, 2018. 2

[25] F. Liu. Susan: segment unannotated image structure using adversarial network. *Magnetic resonance in medicine*, 81(5):3330–3345, 2019. 2

[26] F. Liu et al. Deep convolutional neural network and 3d deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging. *Magnetic Resonance in Medicine*, 79(4):2379–2391, 2018. 2

[27] M. T. Nieminen, V. Casula, M. T. Nevalainen, and S. S. Saarakkala. Osteoarthritis year in review 2018: imaging. *Osteoarthritis and cartilage*, 2018. 1

[28] B. Norman, V. Pedoia, and S. Majumdar. Use of 2d u-net convolutional neural networks for automated cartilage and meniscus segmentation of knee mr imaging data to determine relaxometry and morphometry. *Radiology*, 288(1):177–185, 2018. 1, 2, 6

[29] A. Paszke et al. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 5

[30] C. Peterfy, E. Schneider, and M. Nevitt. The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis and cartilage*, 16(12):1433–1441, 2008. 2

[31] J. Podlipská et al. Comparison of diagnostic performance of semi-quantitative knee ultrasound and knee radiography with MRI: Oulu knee osteoarthritis study. *Scientific reports*, 6:22365, 2016. 4

[32] 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*, pages 234–241, 2015. 3

[33] M. D. Ryzhkov. Knee cartilage segmentation algorithms: a critical literature review. Master's thesis, Utrecht University, 2015. https://dspace.library.uu.nl/handle/1874/308831. 2

[34] T. Stammberger et al. Interobserver reproducibility of quantitative cartilage measurements: comparison of b-spline snakes and manual segmentation. *Magnetic resonance imaging*, 17(7):1033–1042, 1999. 1

[35] A. Tack, A. Mukhopadhyay, and S. Zachow. Knee menisci segmentation using convolutional neural networks: Data from the osteoarthritis initiative. *Osteoarthritis and Cartilage*, 26(5):680 – 688, 2018. 1, 5, 6, 7

[36] A. Tack and S. Zachow. Accurate automated volumetry of cartilage of the knee using convolutional neural networks: Data from the osteoarthritis initiative. In *IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 40 – 43, 2019. 1, 5, 6, 7

[37] A. Tiulpin et al. Deep-learning for tidemark segmentation in human osteochondral tissues imaged with micro-computed tomography. *arXiv preprint arXiv:1907.05089*, 2019. 3

[38] A. Tiulpin, S. Klein, S. Bierma-Zeinstra, J. Thevenot, E. Rahtu, J. van Meurs, E. H. Oei, and S. Saarakkala. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. *arXiv preprint arXiv:1904.06236*, 2019. 1

[39] A. Tiulpin and S. Saarakkala. Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks. *arXiv preprint arXiv:1907.08020*, 2019. 1

[40] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. *Scientific reports*, 8(1):1727, 2018. 1

[41] Y.-H. Tsai et al. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 3, 4

[42] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 2

[43] W. E. Van Spil et al. Osteoarthritis phenotypes and novel therapeutic targets. *Biochemical pharmacology*, 2019. 1

[44] V. Verma et al. Manifold mixup: Better representations by interpolating hidden states. *arXiv preprint arXiv:1806.05236*, 2018. 5

[45] J. Wang, M. Knol, A. Tiulpin, F. Dubost, M. De Bruijne, M. Vernooij, H. Adams, M. A. Ikram, W. Niessen, and G. Roshchupkin. Grey matter age prediction as a biomarker for risk of dementia: A population-based study. *BioRxiv*, page 518506, 2019. 1

[46] Z. Xu and M. Niethammer. Deepatlas: Joint semi-supervised learning of image registration and segmentation. *arXiv preprint arXiv:1904.08465*, 2019. 2

[47] Y. Yaguchi, F. Shiratani, and H. Iwaki. Mixfeat: Mix feature in latent space learns discriminative space. ICLR 2019 submission at https://openreview.net/forum?id=HygT9oRqFX, 2019. 2

[48] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. 2