Received: 27 March 2019 | Revised: 17 August 2019 | Accepted: 6 September 2019

DOI: 10.1002/mrm.28022

FULL PAPER

Magnetic Resonance in Medicine

FatSegNet: A fully automated deep learning pipeline for adipose tissue segmentation on abdominal dixon MRI

Santiago Estrada[1,2] | **Ran Lu**[2] | **Sailesh Conjeti**[1] | **Ximena Orozco-Ruiz**[2] | **Joana Panos-Willuhn**[2] | **Monique M. B. Breteler**[2,3] | **Martin Reuter**[1,4,5]

[1]Image Analysis, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

[2]Population Health Sciences, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany

[3]Faculty of Medicine, Institute for Medical Biometry, Informatics and Epidemiology (IMBIE), University of Bonn, Bonn, Germany

[4]A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, Massachusetts

[5]Department of Radiology, Harvard Medical School, Boston, Massachusetts, USA

Correspondence
Santiago Estrada, DZNE, Venusberg-Campus 1 BLDG 99, 53127 Bonn, Germany.
Email: santiago.estrada@dzne.de

Funding information
Diet-BB Competence Cluster in Nutrition Research, Federal Ministry of Education and Research (BMBF), Grant/Award Number: 01EA1410C and FKZ: 01EA1809C; JPI HDHL on Biomarkers for Nutrition and Health, BMBF, Grant/Award Number: 01EA1705B; National Institutes of Health (NIH), Grant/Award Number: R01NS083534 and R01LM012719

Purpose: Introduce and validate a novel, fast, and fully automated deep learning pipeline (FatSegNet) to accurately identify, segment, and quantify visceral and subcutaneous adipose tissue (VAT and SAT) within a consistent, anatomically defined abdominal region on Dixon MRI scans.

Methods: FatSegNet is composed of three stages: (a) Consistent localization of the abdominal region using two 2D-Competitive Dense Fully Convolutional Networks (CDFNet), (b) Segmentation of adipose tissue on three views by independent CDFNets, and (c) View aggregation. FatSegNet is validated by: (1) comparison of segmentation accuracy (sixfold cross-validation), (2) test–retest reliability, (3) generalizability to randomly selected manually re-edited cases, and (4) replication of age and sex effects in the Rhineland Study—a large prospective population cohort.

Results: The CDFNet demonstrates increased accuracy and robustness compared to traditional deep learning networks. FatSegNet Dice score outperforms manual raters on VAT (0.850 vs. 0.788) and produces comparable results on SAT (0.975 vs. 0.982). The pipeline has excellent agreement for both test–retest (ICC VAT 0.998 and SAT 0.996) and manual re-editing (ICC VAT 0.999 and SAT 0.999).

Conclusions: FatSegNet generalizes well to different body shapes, sensitively replicates known VAT and SAT volume effects in a large cohort study and permits localized analysis of fat compartments. Furthermore, it can reliably analyze a 3D Dixon MRI in ∼1 minute, providing an efficient and validated pipeline for abdominal adipose tissue analysis in the Rhineland Study.

KEYWORDS
deep learning, dixon MRI, neural networks, semantic segmentation, subcutaneous adipose tissue, visceral adipose tissue

1 | INTRODUCTION

The excess of body fat depots is an increasing major public health issue worldwide and an important risk factor for the development of metabolic disorders and reduced quality of life.[1,2] While the body mass index (BMI) is a widely used indicator of adipose tissue accumulation in the body, it does not provide information on fat distribution[3] neither with respect to different fat tissue types nor with respect to deposit location. Different compartments of adipose tissue are associated with different physiopathological effects.[4,5] Abdominal adipose tissue (AAT), composed of subcutaneous and visceral adipose tissue (SAT and VAT), has long been associated with an increased risk of chronic cardiovascular diseases, glucose impairment, and dyslipidemia.[6,7] Recently, several studies have indicated a stronger relation between the accumulation of VAT with an adverse metabolic and inflammatory profile compared to SAT.[8,9] Therefore, an accurate and independent measurement of VAT and SAT volumes (VAT-V and SAT-V) is of significant clinical and research interest.

Currently, the gold standard for measuring VAT-V and SAT-V is the manual segmentation of abdominal fat images from Dixon magnetic resonance (MR) scans—a very expensive and time-consuming process. Thus, especially for large studies, automatic segmentation methods are required. However, achieving good accuracy is challenging due to complex AAT structures, a wide variety of VAT shapes, large anatomical differences across subjects, and the inherent properties of the Dixon images: low intensity contrast between adipose tissue classes, inhomogeneous signals, and potential organ motion. So far, those limitations impeded the widespread implementation of automatic and semi-automatic techniques based on intensity and shape features, such as fuzzy-clustering,[10] *k*-means clustering,[11] graph cut[12,13] active contour methods,[14] and statistical shape models.[15]

Recently, fully convolutional neural networks (F-CNNs)[16,17] have been widely adopted in the computer vision community for pixel/voxel-wise image segmentation in an end-to-end fashion to overcome above-mentioned challenges. With these methods there is no need to extract manual features, divide images into patches, or implement sliding window techniques. F-CNNs can automatically extract intrinsic features and integrate global context to resolve local ambiguities thereby improving the results of the predicted models.[17] Langer et al[18] proposed a three-channel UNet for AAT segmentation, which is a conventional architecture for 2D medical image segmentation.[19] While this method showed promising results, we demonstrate that our network architecture outperforms the traditional UNet for segmenting AAT on our images with a wide range of anatomical variation. More recent architectures such as the SD-Net[20] and Dense-UNet, a densely connected network,[21] have the potential to improve generalizability and robustness by encouraging feature re-usability and strengthening information propagation across the network.[21] In prior work, we introduced a competitive dense fully convolutional network (CDFNet)[22] as a new 2D F-CNN architecture that promotes feature selectivity within a network by introducing maximum attention through a maxout activation unit.[23] The maxout boosts performance by allowing the creation of specialized sub-networks that target a specific structure during training.[24] Therefore, this approach facilitates the learning of more complex structures[22,24] with the added benefit of reducing the number of training parameters relative to the aforementioned networks.

In this paper, we propose FatSegNet, a novel fully automated deep learning pipeline based on our CDFNet architecture to localize and segment VAT and SAT on abdominal Dixon MR images from the Rhineland Study, an ongoing large population-based cohort study.[25,26] To constrain AAT segmentations to a consistent anatomically defined region, the proposed pipeline consists of three stages:

1. **Localization** of the abdominal region using a semantic segmentation approach by implementing CDFNet models on sagittal and coronal planes; we use the lumbar vertebrae positions as reference points for selecting the region of interest.
2. **Segmentation** of VAT and SAT within the abdominal region through 2D CDFNet models on three different planes (axial, sagittal, and coronal).
3. **A view aggregation** stage where the previous generated label maps are combined to generate a final 3D segmentation.

We initially evaluate and compare the individual stages of the pipeline with other deep learning approaches in a sixfold cross-validation. We show that the proposed network architecture (CDFNet) improves segmentation performance and simultaneously reduces the number of required training parameters in step 1 and 2. After asserting segmentation accuracy, we evaluate the whole pipeline (FatSegNet) with respect to robustness and reliability against two independent test sets: a manually edited and a test–retest set. Finally, we present a case study on unseen data comparing the VAT-V and SAT-V calculated from the FatSegNet segmentations against BMI to replicate age and sex effects on these volumes in a large cohort.

2 | METHODS

2.1 | Data

2.1.1 | MR imaging acquisition

MR image acquisition was performed at two different sites both with identical 3T Siemens MAGNETOM Prisma MR scanners (Siemens Healthcare, Erlangen, Germany).

The body coil was used for signal reception of a three-dimensional two-point Dixon sequence (acquisition time = 12 s, echo time TE1 = 1.23 ms, TE2 = 2.46 ms, repetition time TR = 4.12 ms, axial field of view = 500 mm × 437 mm, flip angle = 6°, left-right readout bandwidth = 750 Hz/pixel, partial Fourier factor 6/8 × 5/8). Based on a preceding moving-table abdominal localizer, the field-of-view was centered on the middle of the third lumbar vertebra (L, L3). Data were acquired during a single breath-hold in supine position with arms placed at the sides. The image resolution was finally interpolated from 2.0 mm × 2.7 mm × 10.0 mm to 2.0 mm × 2.0 mm × 5.0 mm (matrix size = 256 × 224 × 72).

2.1.2 | Datasets

The Rhineland Study is an ongoing population-based prospective cohort (https://www.rheinland-studie.de/) which enrolls participants aged 30 years and above at baseline from Bonn, Germany. The study is carried out in accordance with the recommendations of the International Council for Harmonisation (ICH) Good Clinical Practice (GCP) standards (ICH-GCP). Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki.

The first 641 subjects from the Rhineland Study with BMI and abdominal MR Dixon scans are included. The sample presents a mean age of 54.2 years (range 30 to 95) and 55.2% of the subjects are women. The BMI of the participants ranges from 17.2 to 47.7 kg/m^2 with a mean of 25.2 kg/m^2. Subjects were stratified into two subsets: 38 scans were manually annotated for training and testing; the remaining 603 subjects were segmented using the proposed pipeline. After visual inspection, 16 subjects were excluded due to poor image quality or extreme motion artifacts (e.g. potentially caused by breathing). Thus, 587 participants were used for the case study analysis and a subset of 50 subjects were randomly selected for manual corrections of the predicted label maps. This manually edited set and an independent test–retest set of 17 healthy young volunteers were used to assess reliability of the automated segmentation and volume estimates.

Ground truth data

38 subjects were randomly selected from sex and BMI strata to ensure a balanced population distribution. These scans were manually annotated by two trained raters without any semi-automated support such as thresholding, which can reduce accuracy in the ground truth and lead to overestimation of the performance of the proposed automated method.

Specific label schemes were created for each individual task of the pipeline. For localizing the abdominal region, raters divided the scans into three different blocks defined by the location of the vertebrae as follows: the abdominal region (from lower bound of twelfth thoracic vertebra (Th12) to the lower bound of L5), the thoracic region (all above the lower bound of Th12), and the pelvic region (everything below the lower bound of L5), as illustrated in Figure 1E). For AAT segmentation, 60 slices per subject were manually labeled into three classes: SAT, VAT, and bone with neighbouring tissues. The bone was labeled to prevent bone marrow from being misclassified as adipose tissue. In order to improve spatial context and prevent misclassification of the arms, the dataset was complemented by a synthetic class defined as "other tissue" that was composed of any soft tissue inside the abdomen cavity that is not VAT or SAT. The manual annotations are illustrated in Figure 1B,C. Furthermore, four subjects were labeled by both raters to evaluate the inter-rater variability.

Test–retest data

17 additional subjects were recruited with the exclusive purpose of measuring the acquisition protocol reliability.

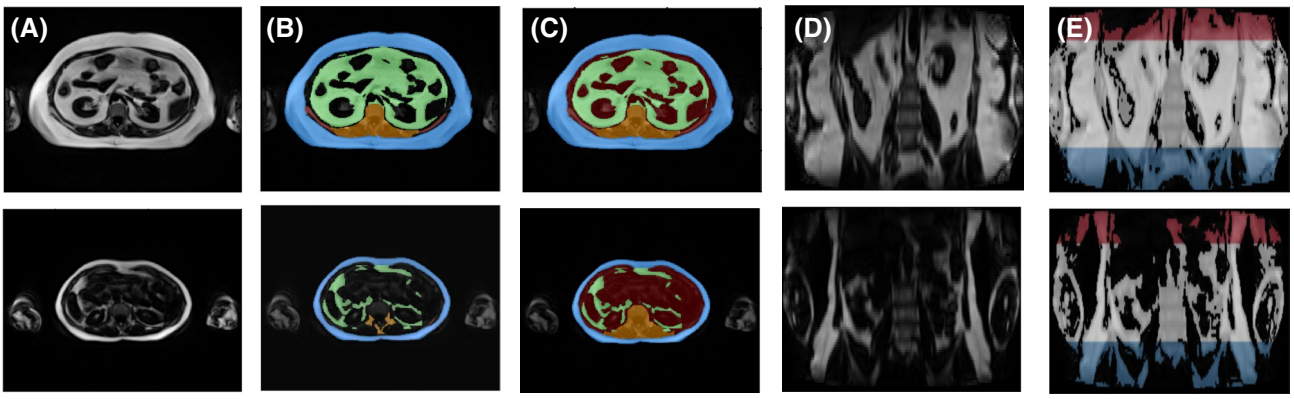

FIGURE 1 MR Dixon images and ground truth from two subjects with different BMI (obese (upper), normal (lower). A, Fat images: axial plane. B, Initial manual segmentation (blue: SAT, green: VAT, orange: bone and surrounding structures). C, Ground truth with additional synthetic class (red: other-tissue) and filled-in bone structures (orange). D, Fat images: coronal plane. E, Ground truth for localization of region of interest (red: thoracic region, white: abdominal region (region of interest), blue: pelvic region)

The group presents a mean age of 25.5 years (range: 20 to 31) and 65.0% of the participants are women; all of them have a normal BMI (BMI <25 kg/m^2). Subjects were scanned in two consecutive sessions. Before starting the second session, subjects were removed from the scanner and re-positioned.

2.2 | FatSegNet pipeline

The FatSegNet is to be deployed as a post-processing adipose analysis pipeline for the abdominal Dixon MR images acquired in the Rhineland Study. Therefore, it should meet the following requirements: (1) be fully automated, (2) segment the different adipose tissue types within the anatomically defined abdominal region, and (3) be robust to body type variations and generalizable in presence of high population heterogeneity. Following the prior conditions, we designed FatSegNet as a fully automated deep learning pipeline for adipose segmentation (Figure 2).

The proposed pipeline consists of three stages: (1) the abdominal region is localized by averaging bounding boxes from two abdominal segmentation maps generated by CDFNets on the sagittal and coronal view. For each view a bounding box is set to the full image width. The height is extracted by localizing the highest and lowest slice with at least 85% of none background voxels classified as abdominal region. Highest and lowest slice position are averaged across the views. (2) Afterward, adipose tissue is segmented within the abdominal region by three CDFNets on different views (axial, coronal, and sagittal) with standardized input sizes (zero padding). (3) Finally, a view aggregation network merges the predicted label maps from the previous stage into a final segmentation; the implemented multi-view scheme is designed to improve segmentation of structures that are not clearly visible due to poor lateral resolution. This 2.5D strategy produces a fully automated pipeline to accurately segment adipose tissue inside a consistent anatomically defined abdominal region.

2.2.1 | Pipeline components

Competitive dense fully convolutional network (CDFNet)

For the segmentation task, we introduce the CDFNet architecture due to its robustness and generalizability properties. The proposed network improves feature selectivity and, thus, boosts the learning of fine-grained anatomies without increasing the number of learned parameters.[22] We implemented the CDFNet by suitably adopting the Dense-UNet architecture proposed by Roy et al[27] and extending it toward competitive learning via maxout activations.[24]

The Dense-UNet proposed in[27] follows the usual dumbbell like architecture with four dense-block encoders, four dense-block decoders and one bottleneck layer. Each dense-block is based on short-range skip connections between convolutional layers as introduced for densely connected neural networks[28]; the dense connection approach stacks multiple convolutional layers in sequence and the input of a layer is iteratively concatenated with the outputs of the previous layers. This type of connectivity improves feature reusability, increases information propagation, and alleviates vanishing gradients.[28] The architecture additionally incorporates the traditional long-range skip connections between all encoder and decoder blocks of the same spatial resolution as introduced by Ronnenberger et al[19] which improves gradient flow and spatial information recovery.

Within the Dense-UNet, the information aggregation through these connections is performed by concatenation layers. Such a design increases the size of the output feature map along the feature channels, which in turn results in the need to

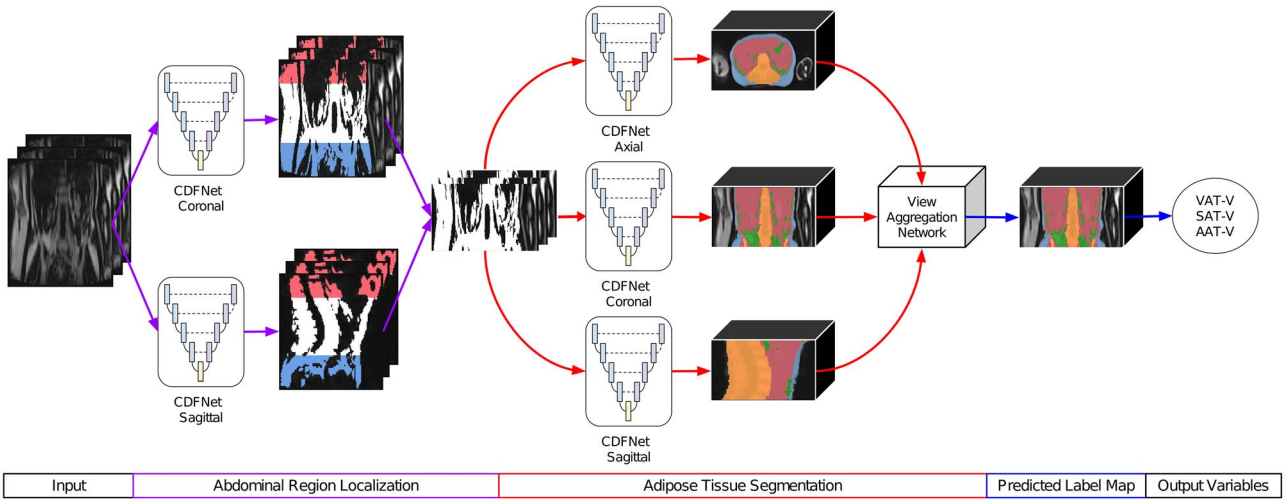

FIGURE 2 Proposed FatSegNet Pipeline for segmenting AAT. The pipeline is divided into three stages: First, localization of abdominal region. Then, tissue segmentation on the abdominal region and finally, view aggregation. Both local and global volume estimates of individual structures are calculated on the final prediction

learn filters with a higher number of parameters. Goodfellow et al introduced the idea of competitive learning through maxout activations,[23] which was adapted by Liao and Carneiro[24] for competitive pooling of multi-scale filter outputs. Both[23] and[24] proved that the use of maxout competitive units boosts performance by creating a large number of dedicated sub-networks within a network that learns to target specific sub-tasks and reduces the number of required parameters significantly, which in turn can prevent over-fitting.

The maxout is a simple feed-forward activation function that chooses the maximum value from its inputs.[23] Within a CNN, a maxout feature map is constructed by taking the maximum across multiple input feature maps for a particular spatial location. The proposed CDFNet uses competitive layers (maxout activation) instead of concatenation layers. Our preliminary results[22] demonstrate that these competitive units promote the formation of dedicated local sub-networks in each of the densely connected blocks within the encoder and the decoder paths. This encourages sub-modularity through a network-in-network design that can learn more efficiently. Toward this, we propose two novel architectural elements targeted at introducing competition within the short- and long-range connections, as follows:

1. **Local Competition—Competitive Dense Block (CDB):** By introducing maxout activations within the short-range skip connections of each of the densely connected convolutional layers (at the same resolution), we encourage local competition during learning of filters. The multiple convolution layers in each block prevent filter co-adaptation.
2. **Global Competition—Competitive Un-pooling Block (CUB):** We introduce a maxout activation between a long-range skip connection from the encoder and the features up-sampled from the prior lower resolution decoder block. This promotes competition between finer feature maps with smaller receptive fields (skip connections) and coarser feature maps from the decoder path that spans much wider receptive fields encompassing higher contextual information.

In brief, the proposed CDFNet comprises a sequence of four CDBs, constituting the encoder path (down-sampling block), and four CDBs constituting the decoder path (up-sampling block), which is joined via a bottleneck layer. The bottleneck consists of a 2D convolutional layer followed by a Batch Normalization. The skip-connections from each of the encoder blocks feed into the CUB that subsequently forward features into the corresponding decoder block of the same resolution as illustrated in Figure 3.

View aggregation network

The proposed view aggregation network is designed to regularize the prediction for a given voxel by considering spatial information from the coronal, axial, and sagittal view. The network, therefore, merges the probability maps of the three different CDFNets from the previous stage by applying a $(3 \times 3 \times 3)$ 3D-convolution (30 filters) followed by a Batch Normalization. Then a $(1 \times 1 \times 1)$ 3D-convolution is employed to reduce the feature maps to the desired number of classes ($n = 5$). The final prediction probabilities are obtained via a concluding softmax layer (as illustrated in Supporting Information Figure S1). Our approach learns to weigh each view differently on a voxel

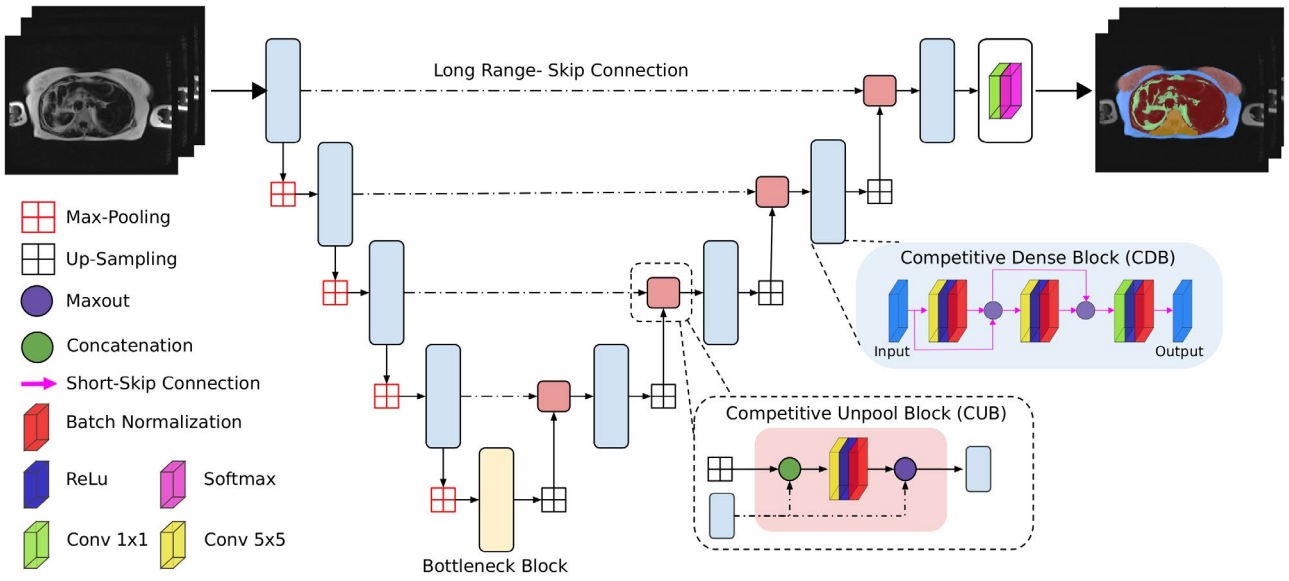

FIGURE 3 Proposed network architecture: Competitive Dense Fully Convolutional Network (CDFNet), with 4 competitive dense blocks (CDB) on each encoder and decoder path and 4 competitive unpool blocks (CUB) between them. CDB and CUB induce local and global competition within the network. Note—the output filters for all convolutional layers in CUB, CDB, and Bottleneck were standardized to 64 channels

level, compared to standard hard-coded global view aggregation schemes. Such hard-coded weighting schemes can be suboptimal when working with anisotropic voxels sizes (e.g., here 2 mm × 2 mm × 5 mm) as resolution differences impose a challenge when combining the spatial information from the finer (within-plane) and coarser (across slice) resolutions. Additionally, in the presence of high variance, abdominal body shapes across subjects segmentation benefit from data-driven approaches that can flexibly adopt weights to individual situations and even spatial locations, which are not possible if hard-coded global weights are being used.

2.3 | Experimental setup

For training and testing the pipeline, we perform a sixfold cross-validation subject-space split on the ground truth dataset. For each fold, 32 subjects are used for training and 6 held out for testing; the test sets splits are approximately balanced based on their BMI classification (underweight [BMI <18.5 kg/m2], normal [18.5 ≤ BMI <25 kg/m2], overweight [25 ≤ BMI <30 kg/m2], and obese [BMI ≥ 30 kg/m2]). This selection process ensures that all BMI categories are used for bench-marking the cross-validation models. Additionally, a final model is implemented using 33 subjects for training holding out 5 subjects spanning different BMI levels for a final performance sanity check (visual quality check and stability of Dice score). Given the limited ground truth data, for all models a validation set to assets convergence during training was created by randomly separating 15% of the slices from the corresponding training set. This allows evaluating performance and generalizability on a completely separate test set.

2.3.1 | Baselines and comparative methods

We validate the FatSegNet by comparing the performance of each stage of the pipeline against the cross-validation test sets using Dice score index (DSC) to measure similarity between the prediction and the ground truth. Let M (ground truth) and P (prediction) denote the labels binary segmentation, the Dice score index is defined as

$$DSC = \frac{2 \cdot |M \cap P|}{|M| + |P|} \tag{1}$$

where $|M|$ and $|P|$ represents the number of elements in each segmentation and $|M \cap P|$ the number of common elements. Therefore, the DSC ranges from 0 to 1 and a higher DSC represents a better agreement between segmentations.

Additionally, we benchmark the proposed CDFNet models for abdominal region localization and AAT delineation with state-of-the-art segmentation F-CNNs such as UNet,[19] SD-Net,[20] and Dense-UNet.[27] We use the probability maps generated from the aforementioned networks to train the view aggregation model and measure performance with and without view aggregation. The proposed view aggregation performance for each FCNNs is compared against two non-data-driven (hard-coded) methods: equally balanced weights for all views and axial focus weights (accounting for higher in-plane resolution, axial = 0.5, coronal = 0.25, sagittal = 0.25). Finally, to permit a fair comparison, all benchmark networks follow the same architecture of four encoder blocks, four decoders blocks, and one bottleneck layer as illustrated in Figure 3 with an input image size of 224 × 256. Note, significant differences between our proposed methods and comparative baselines are evaluated by a Wilcoxon signed-rank test[29] after multiple comparisons correction using a one-sided adaptive FDR.[30]

The aforementioned models are implemented in Keras[31] with a TensorFlow back-end using an NVIDIA Titan Xp GPU with 12 GB RAM and the following parameters: batch size of 8, momentum set to 0.9, constant weight decay of 10^{-06}, and an initial learning rate of 0.01 decreased by a order of 10 every 20 epochs. The models are trained for 60 epochs with an early-stopping criterion (no relevant changes on the validation loss after the last 8 epochs—convergence was observed around 50 epochs). A composite loss function of median frequency balanced logistic loss and Dice loss[20] is used. This loss function emphasizes the boundaries between classes and supports learning of unbalanced classes such as VAT. Finally, online data augmentation (translation, rotation and global scaling) is performed to increase training set size and improve the networks generalizability. Note, the FatSegNet implementation is available at https://github.com/reuter-lab/FatSegNet.

2.3.2 | Pipeline reliability

We assess the FatSegNet reliability by comparing the difference of VAT-V and SAT-V across sessions for each subject of the test–retest and manually edited set. Given a predicted label map and $N_i(l)$ the number of voxels classified as l (VAT or SAT) in session i (test–retest, or manual–automated), the absolute percent difference (APD(l)) of a label volume measures variability across sessions. It is defined as

$$\mathrm{APD}(l) = \frac{2 \cdot \left| N_1(l) - N_2(l) \right|}{N_1(l) + N_2(l)} \cdot 100 \tag{2}$$

Additionally, we calculate the agreement of total VAT-V and SAT-V between sessions by an intra-class correlation (ICC) using a two-way fixed, absolute agreement and single measures ICC(A,1).[32]

2.3.3 | Case study analysis on the Rhineland study

We compare the volumes of abdominal adipose tissue (AAT-V, SAT-V, and VAT-V) generated from FatSegNet with BMI on the unseen dataset. A fast quality control is

TABLE 1 Mean (and standard deviation) Dice scores (cross-validation) of the FCNN models for abdominal adipose tissue segmentation

	Subcutaneous (SAT)				Visceral (VAT)			
Models (PRM)[a]	**Axial**	**Coronal**	**Sagittal**	**V. Aggregation**	**Axial**	**Coronal**	**Sagittal**	**V. Aggregation**
UNet (~20 M)	0.965 (0.029)[b]	0.960 (0.034)[b]	0.960 (0.035)[b]	0.972 (0.019)[b]	0.810 (0.111)[b]	0.804 (0.113)[b]	0.820 (0.101)	0.837 (0.095)[b]
SD-Net (~1,5M)	0.969 (0.027)[b]	0.954 (0.040)[b]	0.956 (0.034)[b]	0.972 (0.020)[b]	0.820 (0.097)[b]	0.812 (0.099)[b]	0.822 (0.091)[b]	0.843 (0.081)[b]
Dense-UNet (~3,3M)	0.972 (0.025)[b]	0.959 (0.037)[b]	0.963 (0.029)[b]	0.975 (0.019)[b]	0.824 (0.091)[b]	0.814 (0.097)[b]	0.827 (0.090)[b]	0.847 (0.080)[b]
Proposed (~2,5M)	0.970 (0.025)	0.966 (0.029)	0.966 (0.027)	0.975 (0.018)	0.826 (0.095)	0.826 (0.085)	0.824 (0.092)	0.850 (0.076)
Inter-rater variability	0.982 (0.018)				0.788 (0.060)			

Note: We show FDR corrected significance indicators of Wilcoxon signed-rank test[29] comparing the proposed CDFNet vs. benchmark FCNNs.

[a]The approximately number of learn parameters reported is for the models without the View Aggregation Network.

[b]Statistical difference using a one-sided adaptive FDR multiple comparison correction[30] at a level of 0.05.

performed to identify drastic failure cases. The differences among BMI groups are evaluated with a one-way analysis of variance (ANOVA) with subsequent Tukey's honest significant difference (HSD) post hoc comparisons. The associations of volumes of abdominal adipose tissue and BMI are assessed using partial correlation and linear regression after accounting for age, sex, and height of the abdominal region. Separate linear regression analyses are performed to explore the effect of age on SAT-V and VAT-V in men and women. All the statistical analyses are performed in R.[33]

3 | RESULTS

3.1 | Method validation

3.1.1 | Localization of abdominal region

For assessing the performance of abdominal region detection after creation of an average bounding box from the coronal and sagittal views the average Dice overlap (sixfold cross-validation) was calculated, as illustrated on the Supporting Information Figure S2. We observe that all models perform extremely well on the relatively easy task of localizing the desired abdominal region (DSC >0.96). There is no significant difference between the models; however, we use our CDFNet because it requires substantially less parameters (see Table 1) compared to the UNet and Dense-UNet.

3.1.2 | Segmentation of AAT

In Table 1, we present the average Dice score (sixfold cross-validation) for VAT and SAT for each individual view as well as for the view aggregation model. Here, we observe that all methods work extremely well for SAT segmentation. Nevertheless, our proposed CDFNet outperforms the UNet and SD-Net on all single-view models and, when compared with the Dense-UNet, there is significant improvement in the sagittal and coronal views. For the more challenging task of VAT recognition, which is a more fine-grained compartment with large shape variation, the proposed CDFNet outperforms the SD-Net on all single planes; when compared with Dense-UNet and U-Net, there is only significant improvement in the axial and coronal plane. Nonetheless, CDFNet achieves this performance with ~30% (Dense-UNet) and ~80% (UNet) less parameters, demonstrating that the proposed architecture improves feature selectivity and simplifies network learning. Furthermore, fewer parameters can help decrease overfitting error, especially when training with limited annotated data, and thus improve generalizability.

Note, that Dice scores increase and difference of pairwise comparisons is slightly reduced after the view aggregation (Table 1), showing that this steps helps all individual networks to reach a better performance by introducing spatial

TABLE 2 Mean (and standard deviation) Dice scores (cross-validation) of hard-coded balanced weights, hard-coded axial focus weights, and the proposed view aggregation for abdominal adipose tissue segmentation

	Subcutaneous (SAT)			Visceral (VAT)		
Single-view model	**Balanced**	**Axial focus**	**Proposed**	**Balanced**	**Axial focus**	**Proposed**
UNet	0.970 (0.026)	0.970 (0.026)	0.972 (0.019)	0.830 (0.098)[a]	0.829 (0.099)[a]	0.837 (0.095)
SD-Net	0.970 (0.026)[a]	0.972 (0.025)[a]	0.972 (0.020)	0.839 (0.084)[a]	0.838 (0.085)[a]	0.843 (0.082)
Dense-UNet	0.973 (0.025)	0.974 (0.024)[a]	0.975 (0.019)	0.841 (0.081)[a]	0.840 (0.082)[a]	0.847 (0.080)
CDFNet	0.972 (0.025)[a]	0.973 (0.024)	0.975 (0.018)	0.844 (0.077)[a]	0.841 (0.080)[a]	0.850 (0.076)

Note: We show FDR corrected significance indicators of Wilcoxon signed-rank test[29] comparing the proposed data-driven aggregation scheme vs. each hard-coded method.

[a]Statistical difference using a one-sided adaptive FDR multiple comparison correction[30] at a level of 0.05.

TABLE 3 Mean absolute percent difference (APD) and interclass correlation agreement (ICC(A,1)) for the volumes estimates of VAT and SAT across sessions of the manually edited and test–retest set

	Manually edited set		Test–retest set	
Metric	**SAT-V**	**VAT-V**	**SAT-V**	**VAT-V**
ICC [95% CI]	0.999 [0.999-1.000]	0.999 [0.994-0.999]	0.996 [0.986-0.999]	0.998 [0.995-0.999]
APD (SD)	0.149% (0.424)	1.398% (0.963)	3.254% (2.524)	2.957% (2.600)

information from multiple views and regularizing the prediction maps. The proposed data-driven aggregation scheme outperforms (DSC) the hard-coded models for SAT and with statistically significance for VAT as shown in Table 2. Furthermore, learned weights are spatially varying and can adjust to subject-specific anatomy, which in turn can improve generalizability. We empirically observe that the aggregation model smoothes the label maps slightly, resulting in visually more appealing boundaries. It also significantly reduces the arms from being misclassified as adipose tissue which can otherwise be observed in different views, especially on overweight and obese subjects, where arms are located closer to the abdominal cavity, as seen Supporting Information Figure S3.

Finally it should be highlighted, that all single-view and the view aggregation models achieve similarly excellent results on the SAT segmentation compared to inter-rater variability and outperform the manual raters for the more challenging VAT segmentation by a margin.

3.1.3 | FatSegNet reliability

Table 3 presents the reliability metrics evaluated on the test–retest and the manually edited test set. The proposed pipeline presents only a small absolute percent volume difference (APD) for VAT and SAT, and excellent agreement between the predicted and corrected segmentation maps. It must be noted, that APD is larger for both tissue types in the test–retest setting as it also includes variance from acquisition noise (e.g. motion artefacts, non-linearities based on different positioning) in addition to potential variances of the processing pipelines. Nevertheless, we observe excellent agreement (ICC) between sessions for the test–retest dataset for both adipose tissue types.

3.2 | Case study: Analysis of Rhineland study data

3.2.1 | The characteristics of the study population

After visual quality inspection, 16 scans were flagged due to image artefacts, such as motion or low contrast (see Figure 4C,D for two examples). The characteristics of the remaining 587 participants with valid data on BMI and volumes of abdominal adipose tissue are presented in Supporting Information Table S1. The mean (SD) age of the subjects is 54.2 (13.3) years, and 54.7% are women. 311 (53.0%) subjects are normal weight, 209 (35.6%) overweight, and 67 (11.4%) obese. We observed a BMI increase with age ($\beta = 0.03$, $P = .007$) and a borderline significance of age difference among BMI groups ($P = .052$, ANOVA). Obvious differences are observed in AAT-V, VAT-V, and SAT-V across BMI groups ($P < .001$, ANOVA). VAT-V to SAT-V ratio is higher in overweight and obese participants compared to those with normal weight ($P < .001$), but there is no difference between overweight and obese ($P = .505$).

3.2.2 | The association between abdominal adipose tissue volumes and BMI

BMI shows a strong positive correlation with AAT-V and SAT-V (AAT-V: $r = .88$, $P < .001$; SAT-V: $r = .85$, $P < .001$),

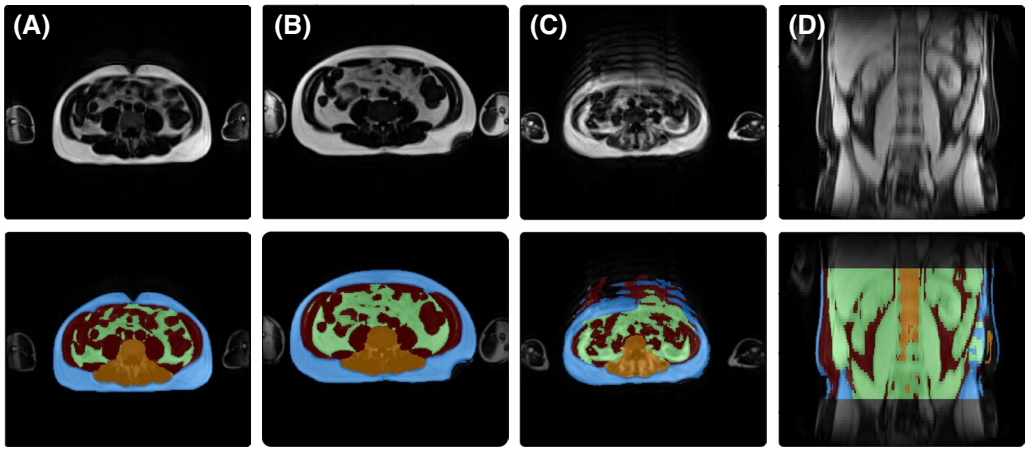

FIGURE 4 Examples of FatSegNet predictions and excluded cases on the Rhineland Study. (A, B) Subjects with different body shapes and accurate segmentations. (C, D) Excluded subjects from the case study due to extreme motion noise (C), or low image contrast quality (D).

but only a moderate correlation with VAT-V ($r = 0.65$, $P < .001$) after adjusting for age, sex, and abdominal region height. As illustrated in Figure 5, both SAT-V and VAT-V are positively associated with BMI after accounting for age, sex, and abdominal region height ($P < .001$). The accumulation of SAT-V is higher than VAT-V as BMI increases.

3.2.3 | Influence of age and sex on VAT-V and SAT-V

The influence of age and sex on VAT-V and SAT-V follows different patterns (as illustrated in Figure 6). Men tend to have lower SAT and higher VAT compared to women ($P < .001$). VAT-V significantly increase with age in both men and women. Conversely, SAT-V is weakly associated with age in women ($\beta = 0.02$, $P = .012$), but not in men ($\beta = -0.01$, $P = .337$).

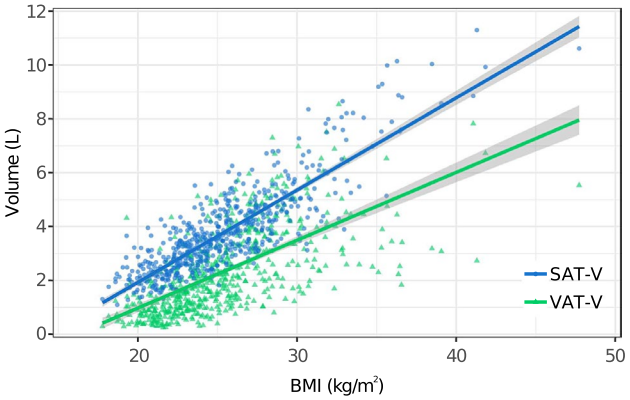

FIGURE 5 Association of BMI with SAT-Volume and VAT-Volume

4 | DISCUSSION

In our study, we established, validated, and implemented a novel deep learning pipeline to segment and quantify the components of abdominal adipose tissue, namely, VAT-V, SAT-V, and AAT-V on a fast acquisition abdominal Dixon MR protocol for subjects from the Rhineland Study, a large population-based cohort. The proposed pipeline is fully automated and requires approximately 1 minute for analyzing a subject's whole volume. Moreover, since the pipeline is based on deep learning models, it can be easily updated and retrained as the study progresses and new manual data are generated—which can further improve overall pipeline robustness and generalizability, providing a pragmatic solution for a population-based study.

The proposed pipeline, termed FatSegNet implements a three-stage design with the CDFNet architecture at the core for localizing the abdominal region and segmenting the AAT. The introduction of our CDFNet inside the pipeline boosts the competition among filters to improve feature selectivity within the networks. CDFNet introduces competition at a local scale by substituting concatenation layers with maxout activations that prevent filter co-adaptation and reduce the overall network complexity. It also induces competition at a global scale through competitive unpooling. This network design, in turn, can learn more efficiently.

For the first stage of the pipeline, i.e. localization of the abdominal region, all FCNNs can successfully determine the upper and lower limit of the abdominal region from a segmentation prediction map. However, our CDFNet requires significantly fewer parameters compared to the traditional UNet and Dense-UNet. Furthermore, the localization block is able to identify the abdominal region correctly even in cases with scoliosis (curved spine) as illustrated in Figure 7F. For the more challenging task of segmenting AAT, we demonstrate that CDFNet recovers VAT significantly better than traditional deep learning variants that rely on

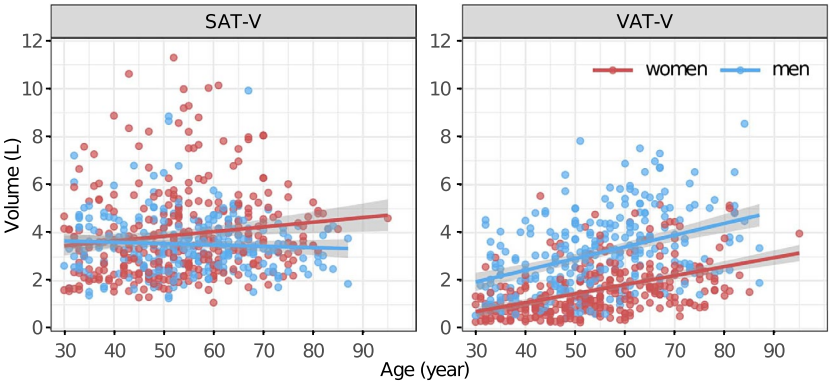

FIGURE 6 The association between age and SAT-Volume and VAT-Volume in men and women

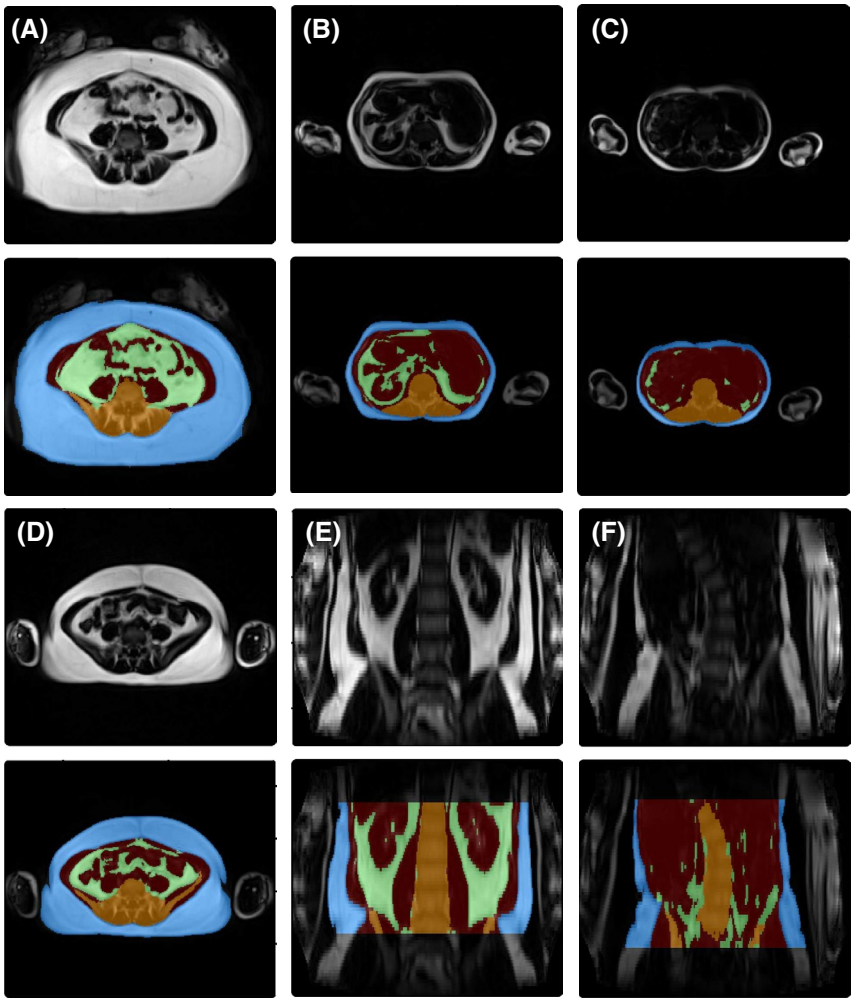

FIGURE 7 Examples of FatSegNet predictions on the Rhineland Study. (A-F) Accurate automatic segmentation of different body shapes. Extreme cases: A, arms are in front of the abdominal cavity, and F, deviated spine

concatenation layers. Additionally, each individual CDFNet view model outperforms manual raters for segmenting the complex VAT and accomplishes equivalent results on SAT. The selection of an inhomogeneous BMI testing set ensures that our method is evaluated for different body types and avoids biases, as better segmentation performance can be achieved on subjects with high content of AAT compared to lean subjects.[34,35] Moreover, images from individuals with high AAT could be accompanied by other types of issues, such as fat shadowing (Figure 7D), or arms located in close proximity to the abdominal cavity (Figure 7A,D,E). These issues are mitigated by our view aggregation model that regularizes the predicted segmentation by combining the spatial context from different views ultimately improving segmentation of tissue boundaries. Moreover, this approach automatically prevents misclassification of arms whereas previous deep learning AAT segmentation methods required manual removal of the upper extremities in a pre-processing step.[18]

Note, that we prefer the 2D over a full 3D approach in this work. A full 3D network architecture has more parameters, requiring significantly more expert annotated training data (full 3D cases) and/or artificial data augmentation, which could increase the chance of overfitting—in addition to increased GPU memory requirements.

As demonstrated on the Rhineland Study data, the proposed pipeline exhibits high robustness and generalizability across a wide range of age, BMI, and a variety of body shapes as seen in Figures 7 and 4A,B. FatSegNet successfully identifies the AAT in different abdomen morphologies, spine curvatures, adipose shadowing, arms positioning, or intensity inhomogeneities. Furthermore, the pipeline has a high test–retest reliability between the calculated volumes of VAT and SAT without the need of any image pre-processing (bias-correction, image registration, etc.) or manual selection of a slice or region. Furthermore, the manually edited test set demonstrates a high similarity of automated and manual labels and excellent agreement of volume estimates. However, as is usual with any automated method, segmentation reliability decreases when input images have low quality as illustrated in Figure 4C,D where the scans present severe motion/breathing artifacts or very low-image contrast. In order to detect these problematic images in large studies, an automated or manual quality control protocol should be implemented before passing images to automated pipelines.

In accordance with previous studies on smaller data sets,[13,36] our data showed a lower correlation of BMI with VAT-V than with AAT-V and SAT-V. We also observed a sex difference of the SAT-V and VAT-V accumulation as previously reported[37,38]: men were more likely to have higher VAT-V and lower SAT-V compared to women. Moreover, we further explored the association between age with SAT-V and VAT-V and found an obvious age effect on the accumulation of VAT-V in both men and women, and a weak age effect on SAT-V in women but not in men. This discrepancy was previously observed by Machann et al,[37] who assessed the body composition using MRI in 150 healthy volunteers aged 19 to 69 years. They reported a strong correlation between VAT-V and age both in men and women, whereas SAT-V only slightly increased with age in women. The fact that our results replicate these previous findings on a large unseen dataset corroborates stability and sensitivity of our pipeline.

In conclusion, we have developed a fully automated post-processing pipeline for adipose tissue segmentation on abdominal Dixon MRI based on deep learning methods. While reducing the number of required parameters, the pipeline outperforms other deep learning architectures and demonstrates high reliability. Furthermore, the proposed method was successfully deployed in a large population-based cohort, where it replicated well known SAT-V and VAT-V age and sex associations and demonstrated generalizability across a large range of anatomical differences, both with respect to body shape and fat distribution.

ACKNOWLEDGEMENTS

We would like to thank the Rhineland Study group for supporting the data acquisition and management, as well as Mohammad Shahid for his support on the deployment of the FatSegNet method into the Rhineland Study processing pipeline. Furthermore we acknowledge Tony Stöcker and his team for developing and implementing the Dixon MRI sequence used in this work. This work was supported by the Diet-Body-Brain Competence Cluster in Nutrition Research funded by the Federal Ministry of Education and Research (BMBF), Germany (grant numbers 01EA1410C and FKZ: 01EA1809C), by the JPI HDHL on Biomarkers for Nutrition and Health (HEALTHMARK), BMBF (grant number 01EA1705B), the NIH R01NS083534, R01LM012719, and an NVIDIA Hardware Award.

ORCID

Santiago Estrada http://orcid.org/0000-0003-0339-8870
Monique M. B. Breteler https://orcid.org/0000-0002-0626-9305
Martin Reuter https://orcid.org/0000-0002-2665-9693

REFERENCES

1. Padwal R, Leslie WD, Lix LM, Majumdar SR. Relationship among body fat percentage, body mass index, and all-cause mortality: a cohort study. *Ann Int Med*. 2016;164:532–541.
2. Ng M, Fleming T, Robinson M, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*. 2014;384:766–781.
3. Tomiyama AJ, Hunger JM, Nguyen-Cuu J, Wells C. Misclassification of cardiometabolic health when using body mass index categories in NHANES 2005–2012. *Int J Obesity*. 2016;40:883–886.
4. Linge J, Borga M, West J, et al. Body composition profiling in the UK biobank imaging study. *Obesity*. 2018;26:1785–1795.
5. Després JP. Body fat distribution and risk of cardiovascular disease: an update. *Circulation*. 2012;126:1301–1313.
6. Kissebah AH, Vydelingum N, Murray R, Evans DJ, Kalkhoff RK, Adams PW. Relation of body fat distribution to metabolic complications of obesity. *J Clinical Endocrinol Metabolism*. 1982;54:254–260.
7. Despres JP, Moorjani S, Lupien PJ, Tremblay A, Nadeau A, Bouchard C. Regional distribution of body fat, plasma lipoproteins, and cardiovascular disease. *Arteriosclerosis Thrombosis Vascular Biol*. 1990;10:497–511.
8. Després JP, Lemieux I. Abdominal obesity and metabolic syndrome. *Nature*. 2006;444:881–887.
9. De Larochellière E, Côté J, Gilbert G, et al. Visceral/epicardial adiposity in nonobese and apparently healthy young adults:

association with the cardiometabolic profile. *Atherosclerosis*. 2014;234:23–29.
10. Zhou A, Murillo H, Peng Q. Novel segmentation method for abdominal fat quantification by MRI. *J Magn Reson Imaging*. 2011;34:852–860.
11. Thörmer G, Bertram HH, Garnov N, et al. Software for automated MRI-based quantification of abdominal fat and preliminary evaluation in morbidly obese patients. *J Magn Reson Imaging*. 2013;37:1144–1150.
12. Christensen AN, Larsen CT, Mandrup CM, et al. Automatic segmentation of abdominal fat in MRI-Scans, using graph-cuts and image derived energies. In: *Scandinavian Conference on Image Analysis*. Tromso, Norway: Springer; 2017:109–120.
13. Sadananthan SA, Prakash B, Leow MKS, et al. Automated segmentation of visceral and subcutaneous (deep and superficial) adipose tissues in normal and overweight men. *J Magn Reson Imaging*. 2015;41:924–934.
14. Mosbech TH, Pilgaard K, Vaag A, Larsen R. Automatic segmentation of abdominal adipose tissue in MRI. In: *Scandinavian Conference on Image Analysis*. Ystad, Sweden: Springer; 2011: 501–511.
15. Wald D, Teucher B, Dinkel J, et al. Automatic quantification of subcutaneous and visceral adipose tissue from whole-body magnetic resonance images suitable for large cohort studies. *J Magn Reson Imaging*. 2012;36:1421–1434.
16. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intelligence*. 2017;39: 2481–2495.
17. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, 2015. pp. 3431–3440.
18. Langner T, Hedström A, Mörwald K, et al. Fully convolutional networks for automated segmentation of abdominal adipose tissue depots in multicenter water–fat MRI. *Magn Reson Med*. 2019;81:2736–2745.
19. 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*. Munich, Germany: Springer; 2015:234–241.
20. Roy AG, Conjeti S, Sheet D, Katouzian A, Navab N, Wachinger C. Error corrective boosting for learning fully convolutional networks with limited data. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Quebec City, Canada: Springer; 2017:231–239.
21. Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: *2017 IEEE Conference on IEEE Computer Vision and Pattern Recognition Workshops (CVPRW)*. Honolulu, HI: 2017:1175–1183.
22. Estrada S, Conjeti S, Ahmad M, Navab N, Reuter M. Competition vs. concatenation in skip connections of fully convolutional networks. In: *International Workshop on Machine Learning in Medical Imaging*. Granada, Spain: Springer; 2018: 214–222.
23. Goodfellow IJ, Warde-Farley D, Mirza M, Courville A, Bengio Y. Maxout networks. In Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28 JMLR. org, Atlanta, GA, 2013. pp. III–1319.
24. Liao Z, Carneiro G. A deep convolutional neural network module that promotes competition of multiple-size filters. *Pattern Recognition*. 2017;71:94–105.
25. Breteler MM, Stöcker T, Pracht E, Brenner D, Stirnberg R. MRI in the Rhineland study: a novel protocol for population neuroimaging. *Alzheimer's Dementia*. 2014;10:P92.
26. Stöcker T. Big data: the Rhineland study. In Proceedings of the 24th Scientific Meeting of the International Society for Magnetic Resonance in Medicine (Singapore); 2016. https://index.mirasmart.com/ISMRM2016/PDFfiles/6865.html.
27. Roy AG, Conjeti S, Navab N, et al. QuickNAT: a fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage*. 2019;186:713–727.
28. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE; 2017:2261–2269.
29. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull*. 1945;1:80–83.
30. Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*. 2006;93:491–507.
31. Chollet F, et al. Keras; 2015. https://keras.io/getting-started/faq/#how-should-i-cite-keras.
32. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996;1:30–46.
33. R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria; 2013, http://www.R-project.org/
34. Kullberg J, Ahlström H, Johansson L, Frimmel H. Automated and reproducible segmentation of visceral and subcutaneous adipose tissue from abdominal MRI. *Int J Obesity*. 2007;31:1806–1817.
35. Addeman BT, Kutty S, Perkins TG, et al. Validation of volumetric and single-slice MRI adipose analysis using a novel fully automated segmentation method. *J Magn Reson Imaging*. 2015;41:233–241.
36. Sun J, Xu B, Freeland-Graves J. Automated quantification of abdominal adiposity by magnetic resonance imaging. *Am J Human Biol*. 2016;28:757–766.
37. Machann J, Thamer C, Schnoedt B, et al. Age and gender related effects on adipose tissue compartments of subjects with increased risk for type 2 diabetes: a whole body MRI/MRS study. *Magn Reson Mater Phys Biol Med*. 2005;18:128–137.
38. Kuk JL, Lee S, Heymsfield SB, Ross R. Waist circumference and abdominal adipose tissue distribution: influence of age and sex–. *Am J Clinical Nutrition*. 2005;81:1330–1334.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

FIGURE S1 View aggregation Network. The proposed network is composed of a initial 3D convolution layer with 30 channels, followed by a batch normalization and a 3D convolutional layer for reducing the feature map dimensionality into the number of classes($n = 5$)

FIGURE S2 Step 1: Abdominal region localization. Dice scores box-plot: Average Dice score (cross-validation) of the abdominal

region detection comparing the Proposed CDFNet vs. other FCNN architectures. The Dice scores are calculated on the average abdominal region generated from the average bounding boxes of the sagittal and coronal model. There is no significant difference between models, nonetheless, the proposed method achieves the same performance with ∼30% and ∼80% less parameters compared to Dense-UNet and UNet, respectively

FIGURE S3 Comparison of single view model (left) vs. view aggregation (right): AAT predictions of two unseen subjects: A, normal subject, B, obese subject. View aggregation avoids arm-misclassification (red boxes) and improves SAT (purple box)

TABLE S1 Case study analysis on the Rhineland Study data. Characteristics of the participants (n = 587) showing mean (SD) for continuous and counts (PCT) for categorical variables

How to cite this article: Estrada S, Lu R, Conjeti S, et al. FatSegNet: A fully automated deep learning pipeline for adipose tissue segmentation on abdominal dixon MRI. *Magn Reson Med*. 2020;83:1471–1483. https://doi.org/10.1002/mrm.28022