Super-resolution of Multi Dimensional Diffusion MRI data: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

Use the title to convey the essential information on the challenge mission.

Super-resolution of Multi Dimensional Diffusion MRI data

Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

SuperMUDI2020

Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

Magnetic Resonance Imaging (MRI) is a fundamental asset for clinical assessment and diagnosis in modern healthcare. A key component to its success is the possibility of non-invasively performing quantitative measurements of physical properties of the living tissue (e.g. the brain) such as the diffusion coefficient and the relaxation times T1 and T2. However, this typically requires lengthy acquisitions to collect high resolution images at different contrasts: e.g. different echo-times (TE) to estimate T2, different inversion-times (TI) to estimate T1, or different b-values to estimate diffusivity. At the same time, in order to guarantee a sufficient signal-to-noise ratio (SNR) the image resolution is often sacrificed. A straightforward way of achieving high resolution and high SNR consists of performing repeated acquisitions and averaging them. However, this would lead to a too long acquisition time which, for MUDI data, is already prohibitive for clinical applications.

In order to mitigate the need of trading off between SNR and resolution, as well as keep the acquisition time unaltered, one possibility is to acquire MRI images at low resolution, e.g. thick slices (2.5x2.5x5 mm3). In this setup, each image, e.g. an axial slice, will share information with that immediately above/below. Such information will then be disentangled with a super-resolution algorithm to obtain the desired high resolution dataset. Similar strategies are currently being used in fetal and neonatal imaging, as they show the additional advantages of being more robust to motion.

The Super-MUDI challenge addresses the critical problem of guaranteeing high SNR and resolution, while keeping the scan time from increasing, by super-resolving MUDI images.

MUDI data: the dataset consists of 1344 volumes comprising 106 unique diffusion gradient directions uniformly spread over four b-shells (500, 1000, 2000, 3000 s/mm2), three TEs, and 28 TIs. These were acquired from five healthy human volunteers (3 f, 2 m, age = 19-46 years), after informed consent was obtained (REC 12/LO/1247), on a clinical 3T Philips Achieva scanner (Best, Netherlands) with a 32-channel adult head coil. Single-shot PGSE EPI with the modifications proposed recently to include relaxometry was employed. Other parameters are TR=7.5s, Resolution=2.5mm isotropic, FOV=220x230x140mm3, SENSE=1.9, halfscan=0.7, Multiband factor 2, TA=52min (including preparation time).

The challenge is comprised of two tasks (see individual task description for further details). Only submissions that attempt all the tasks will be considered for the evaluation. The winner of the challenge will be the submission with the overall best score (lower value) across the two tasks, computed as the sum of the scores in each task (see individual task description for further details about the metrics used for evaluation). Each task explores a different MRI acquisition strategy, leading to two different types of images to super-resolve. Task 1 aims at super-resolving data having high in-plane resolution but thick (axial) slices, whereas Task 2 data having isotropically lower resolution.

Therefore, the outcome of the challenge will be two-fold: 1) evaluating how reliable and stable is a super-resolution method; 2) which combination of sub-sampling strategy and super-resolution method is the best alternative.

The challenge's outcome will provide guidelines on how to obtain MR images having high SNR and resolution, with no additional acquisition time prolongation, by adopting a subsampling strategy in combination with super-resolution methods.

Challenge keywords

List the primary keywords that characterize the challenge.

super resolution, upsampling, diffusion MRI, relaxation

Year

The challenge will take place in ...

2020

FURTHER INFORMATION FOR MICCAI ORGANIZERS

Workshop

If the challenge is part of a workshop, please indicate the workshop.

Computational Diffusion MRI CDMRI

Duration

How long does the challenge take?

Half day.

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

20

Publication and future plans

Please indicate if you plan to coordinate a publication of the challenge results.

We have prepared both a MEDIA submission presenting last year's challenge and are preparing with all participants another journal paper to outline the results.
We aim to do the same for this year's challenge.

Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

The challenge is planned as part of CDMRI - we do not need anything in addition.

TASK: Slice thickness super-resolution for DW-MRI

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

In this challenge, the participants will work with MUDI data. Briefly, the dataset consists of 1344 volumes sampled with 28 different inversion times (TI), 3 echo times (TE) and 112 directions on 3 shells.

The original resolution of the MUDI images is 2.5x2.5x2.5 mm3. For this specific task, we will use a downsampled version to 2.5x2.5x5 mm3, obtained from the original 2.5x2.5x2.5 mm3 images by averaging each two consecutive slices (e.g. slice 1 and 2; then slice 2 and 3; and so on...).

Keywords

List the primary keywords that characterize the task.

MRI, super resolution, MUDI data, DW-MRI

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

Marco Pizzolato, EPFL
Marco Palombo, UCL
Jana Hutter, KCL
Vishwesh Nash, Vanterbilt University
Fan Zhang, Harvard Medical School
Noemi Gyori, UCL

b) Provide information on the primary contact person.

Jana Hutter, KCL

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event open call.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

http://cmic.cs.ucl.ac.uk/cdmri/challenge.html

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully interactive.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Private data is allowed.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate but not eligible for awards and not listed in leaderboard.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

Winner certificate (potential cash prize and GPU, as we provided in the challenge from last year).

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

Top 3 performing methods will be announced publicly.
The participants will be given the option to present their algorithms openly at the MICCAI CDMRI workshop.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

All participating teams qualify as authors. Teams can publish independently 6 months after the CDMRI workshop.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Previous versions are welcome, but only the final version will be used.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Release training data: 1/5/2020
Release testing data: 1/8/2020
Submission data: 15/9/2020

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

Ethical approval was obtained: https://www.developingbrain.co.uk/data/

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC.

Additional comments: An open access agreement needs to be signed (part of our ethics).

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

After the evaluation this will be made available.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

Only if the participating teams agree.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

Only the organisers and members of their immediate team have access.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Research, Longitudinal study, Training, Data reduction, Decision support.

Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Reconstruction.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Healthy adults.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Healthy adults.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

MRI, combined Diffusion and Relaxometry.

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

Tissue segmentation (gray matter, white matter, CSF, etc.) computed from the corresponding anatomical T1w image.

b) ... to the patient in general (e.g. sex, medical history).

Gender, age.

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Brain shown in magnetic resonance imaging data.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Whole brain.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Robustness, Reliability, Precision, Accuracy.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Acquired using a clinical 3T Philips Achieva scanner (Best, Netherlands) with a 32-channel adult head coil.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

MUDI data: the dataset consists of 1344 volumes comprising 106 unique diffusion gradient directions uniformly spread over four b-shells (500, 1000, 2000, 3000 s/mm2), three TEs, and 28 TIs. These were acquired from 12 healthy human volunteers, after informed consent was obtained (REC 12/LO/1247), on a clinical 3T Philips Achieva scanner (Best, Netherlands) with a 32-channel adult head coil. Single-shot PGSE EPI with the modifications proposed recently to include relaxometry was employed. Other parameters are TR=7.5s, Resolution=2.5mm isotropic, FOV=220x230x140mm3, SENSE=1.9, halfscan=0.7, Multiband factor 2, TA=52min (including preparation time).

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

Data from the MUDI 2019 challenge will be used. Acquired at King's College London, London, UK.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

None.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Training and test cases both represent a sequence of images(1344 volumes) of the human brain.

b) State the total number of training, validation and test cases.

10 cases in total.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

8 cases for training, 2 for testing (different from Task 2). Each training/testing case includes 1344 volumes. Based on our successful experience in the last MICCAI challenge, we are confident that such large amounts of data and the distribution of training and testing data are sufficient for evaluating the algorithms from the participants.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Our training and testing data includes 1344 volumes per case.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

NA

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

NA

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

NA

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

NA

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Both the training and testing data sets will be pre-processed (distortion correction) after the MRI acquisition and volume sub-sampling succesively. In particular, a PCA filtering in the complex domain will be performed to reconstruct the images. This step is then followed by image registration to prevent errors due to motion of the volunteers during the acquisition. Such image registration is based on an affine volume registration specifically designed for the type of dataset. Collinear magnitude diffusion-weighted images (DWIs) acquired with different pairs (TI,TE), i.e. inversion time and echo time, are first co-registered together using the highest TI and lowest TE volume as reference: this will produce 106 groups of registered volumes, one for each unique diffusion-gradient direction. The 106 reference volumes are then registered together based on a mutual information metric using the open software Dipy, and the registrations are then propagated across the corresponding collinear DWIs. As for the training data, both the pre-processed and the pre-processed downsampled datasets will be released. The subsampling will be performed by averaging each two consecutive slices (e.g. slice 1 and 2; then slice 2 and 3; and so on...). As for the test data, only the pre-processed downsampled data will be released. Note that we will provide the FreeSurfer segmentation and the T1w data during the training phase only, not the validation data. We will provide the segmentation and T1w at the original resolution and also the downsampled resolution.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

The pre-processed images will be used as ground-truth to compute error metrics for the evaluation of the challenge. The sub-sampled data, used for performing the task, is obtained through a deterministic process (averaging) on the pre-processed images. Therefore, no specific source of errors are expected.

b) In an analogous manner, describe and quantify other relevant sources of error.

No other relevant sources of error are expected.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Mean squared error (MSE)

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

The MSE allows for an objective evaluation of the performance of the task as the ground-truth is known (by the organizers). Given the specifics of the proposed challenge, we believe that MSE is the ideal metric for evaluating both accuracy and precision of the reconstruction. While we will possibly evaluate additional metrics for the characterization of the results (for instance correlation metrics and structural deformation), for the quantification of the results and for the ranking we intend to use the MSE. The use of the MSE will make the evaluation clear, unbiased, and easy to reproduce for the comparison with eventual future studies.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

The MSE will be calculated between the ground-truth, originally pre-processed, high resolution images (unknown to participants) and the super-sampled images submitted by the participants (for the same dataset). The lower the MSE, averaged on the testing datasets, the better the ranking of the submission.

b) Describe the method(s) used to manage submissions with missing results on test cases.

In the challenge we will accept only submissions were the task is performed on all the test datasets.

c) Justify why the described ranking scheme(s) was/were used.

The ranking proposed is justified by the fact that the super-resolution method must perform well on different subjects/datasets, and this justify the average of the MSE values across datasets. Moreover, no need of defining specific regions of interests is present, since the super-resolved images need to be correct regardless of the underlying tissue type (i.e. brain region) or contrast (echo time, inversion time, b-value, etc.)
The mean squared error (MSE) will be calculated between the ground-truth, originally pre-processed, high resolution images (unknown to participants) and the super-sampled images submitted by the participants (for the same dataset). The lower the MSE, averaged on the testing datasets (test cases), the better the ranking of the submission. The proposed ranking is justified by the fact that the super-resolution method must perform well on different subjects/datasets, and this justifies the average of the MSE values across datasets.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

The assessment of the variability in the ranking will be performed by calculating the variability of the local, i.e. voxel-wise, error across brain regions and tissue types. Such variability accounted by computing averages of the MSE per tissue type based on a segmentation of the brain in white matter, gray matter, and cerebrospinal fluid regions. Such segmentation is obtained with software like FreeSurfer (https://surfer.nmr.mgh.harvard.edu/). Similarly, an analysis will be run based on the contrast type, looking if the method proposed by the participants perform well regardless of the contrast (TE, TI, b-value). An analysis will be performed with in house software to identify the contribution of each contrast to the above mentioned tissue type based averages of MSE.

b) Justify why the described statistical method(s) was/were used.

Although the submitted super-resolved images need to perform well regardless of the underlying tissue type and MRI contrast, it is possible that the methods proposed by the participants are sensitive to such things. The overall ranking will still be based on the average MSE (see point 27a), as this will describe the overall performance.

However, this will be accompanied by the statistical analysis described above.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

TASK: Complete 3D super-resolution for DW-MRI

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

In Task 2, the participants will work with the same data as in Task 1.

For this specific task, we will use a downsampled version to 5x5x5 mm3, obtained from the original 2.5x2.5x2.5 mm3 images by averaging 8 adjacent 2.5x2.5x2.5mm3 voxels organized in a cubic pattern (cubic patched) at each original voxel location, in order to obtain overlapping low resolution voxels.

Keywords

List the primary keywords that characterize the task.

MRI, super resolution, MUDI data, DW-MRI

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

Marco Pizzolato, EPFL
Marco Palombo, UCL
Jana Hutter, KCL
Vishwesh Nash, Vanterbilt University
Fan Zhang, Harvard Medical School
Noemi Gyori, UCL

b) Provide information on the primary contact person.

Jana Hutter, KCL

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.

Examples:

- One-time event with fixed submission deadline
- Open call
- Repeated event with annual fixed submission deadline

Repeated event open call.

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI.

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

grand-challenge.org

c) Provide the URL for the challenge website (if any).

http://cmic.cs.ucl.ac.uk/cdmri/challenge.html

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

Fully interactive.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or to publicly available data including (open) pre-trained nets.

Private data is allowed.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate but not eligible for awards and not listed in leaderboard.

d) Define the award policy. In particular, provide details with respect to challenge prizes.

Winner certificate (potential cash prize and GPU, as we provided in the challenge from last year).

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

Top 3 performing methods will be announced publicly.
The participants will be given the option to present their algorithms openly at the MICCAI CDMRI workshop.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

All participating teams qualify as authors. Teams can publish independently 6 months after the CDMRI workshop.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Previous versions are welcome, but only the final version will be used.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Release training data: 1/5/2020
Release testing data: 1/8/2020
Submission data: 15/9/2020

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

Ethical approval was obtained: https://www.developingbrain.co.uk/data/

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

CC BY NC.

Additional comments: An open access agreement needs to be signed (part of our ethics).

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

After the evaluation this will be made available.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

Only if the participating teams agree.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

Only the organisers and members of their immediate team have access.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Research, Training, Data reduction, Decision support.

Task category(ies)

State the task category(ies).

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Reconstruction.

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Healthy adults.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Healthy adults.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

MRI, combined Diffusion and Relaxometry.

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

Tissue segmentation (gray matter, white matter, CSF, etc.) computed from the corresponding anatomical T1w image.

b) ... to the patient in general (e.g. sex, medical history).

Gender, age.

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Brain shown in magnetic resonance imaging data.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Whole brain.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Robustness, Precision, Accuracy.

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Acquired using a clinical 3T Philips Achieva scanner (Best, Netherlands) with a 32-channel adult head coil.

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

MUDI data: the dataset consists of 1344 volumes comprising 106 unique diffusion gradient directions uniformly spread over four b-shells (500, 1000, 2000, 3000 s/mm2), three TEs, and 28 TIs. These were acquired from 12 healthy human volunteers, after informed consent was obtained (REC 12/LO/1247), on a clinical 3T Philips Achieva scanner (Best, Netherlands) with a 32-channel adult head coil. Single-shot PGSE EPI with the modifications proposed recently to include relaxometry was employed. Other parameters are TR=7.5s, Resolution=2.5mm isotropic, FOV=220x230x140mm3, SENSE=1.9, halfscan=0.7, Multiband factor 2, TA=52min (including preparation time).

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

Data from the MUDI 2019 challenge will be used. Acquired at King's College London, London, UK.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

None.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

Training and test cases both represent a sequence of images(1344 volumes) of the human brain.

b) State the total number of training, validation and test cases.

10 cases in total.

c) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

8 cases for training, 2 for testing (different from Task 1). Each training/testing case includes 1344 volumes. Based on our successful experience in the last MICCAI challenge, we are confident that such large amounts of data and the distribution of training and testing data are sufficient for evaluating the algorithms from the participants.

d) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Our training and testing data includes 1344 volumes per case.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

NA

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

NA

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

NA

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

NA

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Both the training and testing data sets will be pre-processed (distortion correction) after the MRI acquisition and volume sub-sampling succesively. In particular, a PCA filtering in the complex domain will be performed to reconstruct the images. This step is then followed by image registration to prevent errors due to motion of the volunteers during the acquisition. Such image registration is based on an affine volume registration specifically designed for the type of dataset. Collinear magnitude diffusion-weighted images (DWIs) acquired with different pairs (TI,TE), i.e. inversion time and echo time, are first co-registered together using the highest TI and lowest TE volume as reference: this will produce 106 groups of registered volumes, one for each unique diffusion-gradient direction. The 106 reference volumes are then registered together based on a mutual information metric using the open software Dipy, and the registrations are then propagated across the corresponding collinear DWIs. As for the training data, both the pre-processed and the pre-processed downsampled datasets will be released. The subsampling will be performed by averaging 8 adjacent 2.5x2.5x2.5mm3 voxels organized in a cubic pattern (cubic patched) at each original voxel location, in order to obtain overlapping low resolution voxels. As for the test data, only the pre-processed downsampled data will be released. Note that we will provide the FreeSurfer segmentation and the T1w data during the training phase only, not the validation data. We will provide the segmentation and T1w at the original resolution and also the downsampled resolution.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

The pre-processed images will be used as ground-truth to compute error metrics for the evaluation of the challenge. The sub-sampled data, used for performing the task, is obtained through a deterministic process (averaging) on the pre-processed images. Therefore, no specific source of errors are expected.

b) In an analogous manner, describe and quantify other relevant sources of error.

No other relevant sources of error are expected.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Mean squared error (MSE)

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

The MSE allows for an objective evaluation of the performance of the task as the ground-truth is known (by the organizers). Given the specifics of the proposed challenge, we believe that MSE is the ideal metric for evaluating both accuracy and precision of the reconstruction. While we will possibly evaluate additional metrics for the characterization of the results (for instance correlation metrics and structural deformation), for the quantification of the results and for the ranking we intend to use the MSE. The use of the MSE will make the evaluation clear, unbiased, and easy to reproduce for the comparison with eventual future studies.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking.

The MSE will be calculated between the ground-truth, originally pre-processed, high resolution images (unknown to participants) and the super-sampled images submitted by the participants (for the same dataset). The lower the MSE, averaged on the testing datasets, the better the ranking of the submission.

b) Describe the method(s) used to manage submissions with missing results on test cases.

In the challenge we will accept only submissions were the task is performed on all the test datasets.

c) Justify why the described ranking scheme(s) was/were used.

The ranking proposed is justified by the fact that the super-resolution method must perform well on different subjects/datasets, and this justify the average of the MSE values across datasets. Moreover, no need of defining specific regions of interests is present, since the super-resolved images need to be correct regardless of the underlying tissue type (i.e. brain region) or contrast (echo time, inversion time, b-value, etc.)
The mean squared error (MSE) will be calculated between the ground-truth, originally pre-processed, high resolution images (unknown to participants) and the super-sampled images submitted by the participants (for the same dataset). The lower the MSE, averaged on the testing datasets (test cases), the better the ranking of the submission. The proposed ranking is justified by the fact that the super-resolution method must perform well on different subjects/datasets, and this justifies the average of the MSE values across datasets.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

The assessment of the variability in the ranking will be performed by calculating the variability of the local, i.e. voxel-wise, error across brain regions and tissue types. Such variability accounted by computing averages of the MSE per tissue type based on a segmentation of the brain in white matter, gray matter, and cerebrospinal fluid regions. Such segmentation is obtained with software like FreeSurfer (https://surfer.nmr.mgh.harvard.edu/). Similarly, an analysis will be run based on the contrast type, looking if the method proposed by the participants perform well regardless of the contrast (TE, TI, b-value). An analysis will be performed with in house software to identify the contribution of each contrast to the above mentioned tissue type based averages of MSE.

b) Justify why the described statistical method(s) was/were used.

Although the submitted super-resolved images need to perform well regardless of the underlying tissue type and MRI contrast, it is possible that the methods proposed by the participants are sensitive to such things. The overall ranking will still be based on the average MSE (see point 27a), as this will describe the overall performance. However, this will be accompanied by the statistical analysis described above.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.