Large Scale Vertebrae Segmentation Challenge: Structured description of the challenge design

Remark: This challenge have been slightly modified. All changes are highlighted in red.

CHALLENGE ORGANIZATION

Title

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

Large Scale Vertebrae Segmentation Challenge

Challenge acronym

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

VerSe'20

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.

A primary step in automated quantification of spinal morphology and pathology is vertebral labelling and segmentation. Aimed at these tasks, the first iteration of the 'Large Scale Vertebrae Segmentation Challenge' (VerSe'19) was held at MICCAI 2019 and received considerable participation from the community (>250 registrations and data downloads, 20 participating teams). With its first iteration, VerSe addressed a severe shortage of publicly-available, large, accurately annotated CT spine data in the community by releasing 160 CT scans and their voxel-level annotations comprised of a large variety in fields of view, spatial resolutions, spinal and vertebral pathologies, collected over several scanners from two major vendors.

Building on the data, experience, and learning from VerSe'19, we propose to organise a second iteration for of the Large Scale Vertebrae Segmentation Challenge (VerSe'20) at MICCAI 2020. With VerSe'20, we aim to work with 300 CT scans (~100% increase over its previous iteration). While retaining the richness of its predecessor, the data will now be multi-centre with five different institutions and and all four major scanner manufacturers. Additionally, challenging the learning algorithms, focus is given to include atypical anatomies such as transitional vertebrae and additional vertebrae such as L6. With clinical-translation being the primary objective of data preparation, we believe the algorithms using this data would be more robust and generalisable.

Challenge keywords

List the primary keywords that characterize the challenge.

spine, vertebrae, labeling, segmentation, computed tomography

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.

None

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.

40 teams.

We base this on the participation count of VerSe'19 and our learning from it. VerSe'19 received 20 submissions with the participants having ~2.5 months to work on the data [1]. For this iteration, we expect this duration to be longer for two reasons: 1) challenge acceptance decisions are announced earlier, 2) all the paraphernalia for annotating the data is already in place, allowing us to release the data faster. Therefore, we expect an increased participation.

Publication and future plans

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

Yes. We intend to compile a journal article based on our findings in VerSe'20. This iteration of the challenge focusses on generalisability across centres and robustness to anatomical outliers. Thus the manuscript will involve: A description of the methods that generated the dataset annotations, description of the participants' algorithms, and a detailed analysis of their cross-modality transfer as well as their anomaly-detection and robustness characteristics.

Timeline: Tentatively, we plan to have a detailed evaluation ready in November '20, submit a first draft to arXiv in December '20 and open to suggestions and corrections from all teams, and finally submit the final journal article in February '21.

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.

Online platform: grand-challenge.org

TASK: Vertebrae Labelling

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.

Localise and label vertebrae in CT scans.

Keywords

List the primary keywords that characterize the task.

spine, vertebrae, labelling, landmark detection, CT

ORGANIZATION

Organizers

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

1. Anjany Sekuboyina
(Informatics & Klinikum rechts der Isar, Technical University of Munich.)

2. Bjoern Menze,
(Informatics, Technical University of Munich.)

3. Jan Kirschke,
(Klinikum rechts der Isar, Technical University of Munich.)

b) Provide information on the primary contact person.

Anjany Sekuboyina
(email: anjany.sekuboyina@tum.de)

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).

verse2020.grand-challenge.org/

Participation policies

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

Fully automatic, Semi automatic.

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.

Publicly available 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.

Prospective sponsorship from NVIDIA (similar to VerSe'19).

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 5 performing submissions are announced at the challenge. Detailed analysis of performances is available upon request.

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).

1. At least two authors from every submission (first and last) listed in the challenge paper will be included in the authors list. Additional authors will be included upon request with justification according to ICMJE authorship guidelines.

2. Participating teams can submit their results separately without any embargo. The challenge paper will focus on the evaluation and will only include an overview of the individual submissions.

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.

Evaluation will be done in two phases with different data for both phases:

Phase 1 (image data public and annotations private): Participants submit predictions on test set images.

Phase 2 (image data private and annotations private): Participants submit docker containers which are run by organisers in-house on hidden test set images.

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.

Participants will have access to phase-1 test set. Predictions on this data can be submitted to grand-challenge.org to get them instantly evaluated. However, in order to prevent overfitting, the number of these submissions is limited to one per day.

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 of Training data: May 15th, 2020
- Release of public Test data: July 20th, 2020
- Opening of submission system: July 20th, 2020
- Closing of submission system for public test data: July 24th, 2020
- Closing of submission system for dockers: Aug 7th, 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).

Ethics approval by the local ethics committee of the University Hospital of TUM granted 18.3.2019, No. 27/19/S-SR.

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 SA.

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.

The evaluation is already available online on grand-challenge.org's evaluation portal. We will release the python code for all systems on grand-challenge.org at the time of training data release.

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

Participants own the rights on distribution of their code. We will not re-distribute their code.

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.

Funding Agency: European Research Council.
Funding of the prize: NVIDIA.
Only the main organizers and their local annotation team will have access to all test labels and the private test datasets.

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

Intervention planning, Decision support, Treatment planning, Diagnosis, Screening, Assistance, Longitudinal study.

Task category(ies)

State the task category(ies).

Examples:

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

Localization.

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.

All patients that get a CT scan of the trunk or any part of the spine.

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

All patients that get a CT scan of the trunk or any part of the spine with a particular focus on anatomical anomalies, such as L6, sacralised L5, C7 with cervical ribs.

Imaging modality(ies)

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

Computed tomography with and without contrast from different scanners, institutions, and acquisition settings.

Context information

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

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

No additional context information will be given.

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

No additional context information will be given.

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.

Computed tomography (CT) data of the trunk and neck, including parts of the spine or the entire spine.

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.

All vertebra of the spine.

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)).

Hardware requirements, Specificity, Accuracy, Sensitivity, Precision, Robustness.

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).

All data was acquired on state of the art Multislice CT scanners from 4 different vendors: Siemens, Philips, Toshiba and GE.

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

All clinically-used routine image protocols have been used, including scans with and without i.v. and oral contrast, different kVp settings, and adaptive tube load.

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 was acquired at the University hospital TUM (60%) and multiple different smaller institutions (30%). Public data from a previous CSI 2014 labelling challenge (10%) will also be included.

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).

All patients above 18 years of age are included in the data acquisition process.

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 CT image stack of a human body. Training cases have vertebra-specific centroid-annotations while the test cases include the images only.

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

Training data: 100 CT scans (images & centroids are public)
Public test data: 100 CT scans (images public & centroids are hidden)
Private test data: 100 CT scans (images & centroids are hidden)

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

From our point of view, this was the best compromise between:
- presenting a balanced dataset dealing with multi-centre acquisitions and different anomalies.
- a dataset of a size appropriate enough to avoid overfitting while illustrating the differences in algorithmic novelty.
- acceptable effort in manual annotation.

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.

In this iteration of VerSe, we focused on two points:
- 50% of the data is multi-centre data
- 50% of the data includes anomalies

Based on the experience with VerSe'19, current state of the art algorithms perform poorly in spinal anomalies. As anatomical anomalies are severely underrepresented in non-selective cross sectional data, it can be expected that learning based approaches under-perform on these. Also, the algorithms' generalisability to multi-centre data is unknown. We believe the composition of our dataset and its size will help address these issues.

CT data included are consecutive patients identified in retrospective searches in the PACS limited to the last five years. We include all scans, except CT-myelographies with contrast material present in the spinal canal.

Note 1: VerSe'20 builds on the VerSe'19 dataset. To be able to present 50% multi-centre data, 50% anatomical anomalies, we were able to include 62 scans from VerSe'19 among the 300 scans for VerSe'20 (without data-

leakage, i.e. a test scan remains test scan and a training scan remains a train one). As the training data of VerSe'19 is publicly available, users can always use the scans not included in VerSe'20 (44 scans).

Note 2: As scans with anomalies are rare, we will have to take those from our local database (University hospital TUM), with 7 different scanners from two vendors, overlapping with VerSe'19.

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.

We use a human hybrid algorithm, based on our automatic labelling and segmentation algorithm. An automated centroid is provided based on the centre of mass of the vertebral body and corrected by one experienced medical student, if necessary. Final annotations are checked consecutively by one radiologist with 5 years and by one neuroradiologist with 17 years of experience in spine imaging. The final decision is made by the most experienced reader, in consensus with both other raters.

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.

The annotator was instructed to place the centroid in the centre of mass of the vertebral body. For exact enumeration, we refer to the review article [10]:
- In general, the shape of a vertebra should be responsible for its nomenclature.
- Cervical, thoracic and lumbar vertebrae are enumerated consecutively in a cranio-caudal direction. We will use additional imaging material as whole-spine MRI or radiographs to establish the gold standard, as it has been stated that only a continuous enumeration from C2 downwards can define this classification in ambiguous cases.
- Cervical spine enumeration variants are rare, ribs have been described as “cervical ribs” in C7 and rarely above.
- Thoracolumbar transition vertebrae are treated according to their shape and the presence of ribs: if transverse processes are present, the vertebra is considered an L1 (thus maybe leading to only 11 thoracic vertebrae), if ribs are present and no transverse processes, it is an thoracic vertebra (maybe T13, even if only 4 lumbar vertebrae remain).
- Lumbosacral transition vertebrae are called lumbar including a Castellvi Grad IV LSTV [11]. Additionally, in Grade III and IV, the lumbar vertebra has to have a squared shape (Grade 4, according to O’Driscoll CM, et al. [12]). They are called sacral, if both transverse processes are fully fused with the sacrum and the vertebra has a “sacral” shape, even if a small disc is still present (Grade 3, according to O’Driscoll CM). If none of these descriptions fit, the normal enumeration with 5 lumbar vertebrae is chosen.

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.

Algorithm: Uses a three-stage interactive pipeline with spine localisation, vertebrae labelling (Btrfly Net [3]), and vertebrae segmentation (U-Net based architecture). The centroids predicted by the second stage are used as initialisation for human annotation process.

Humans: One specifically trained medical student, one radiologist with 5 years of experience for direct supervision, one neuroradiologist with 17 years of experience in spine imaging for final approval. This annotation process is organised hierarchically.

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.

N/A

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.

In case of sub-mm resolution, the spatial resolution will be downsampled to 1mm in those dimensions.

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.

Possible error sources include cases where no clear anatomical definition of one specific label is available (i.e. being a lumbalized S1 or an L6). In these cases we will use additional imaging material as whole-spine MRI or radiographs to establish the gold standard. As described above in (23b), we will treat all cases consistently and all cases will be checked by one experienced neuroraidologist.

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

N/A

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)

Identification Rate (Id. rate) and Localisation Distances (d_mean) as defined in [5] Precision (P) and Recall (R) as defined in [3].

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

1. Id. rate and d_mean together capture the identification and localisation capability of the algorithm.
2. Vertebral-level Precision and Recall capture the number of false positive localisations, as all the vertebrae are not always visible in the scan.

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.

Following VerSe'19 challenge and Brain Tumor Segmentation (BraTS) challenge: We follow a 'point' based evaluation system as described in Sekuboyina et al. [2]. Added to this, we perform bootstrapping to make the ranks more robust. This results in a ranking scheme detailed below:

Step 1: Compute metrics at a scan level (Id. rate and localisation distance).
Step 2: For each metric: Compare every possible pair of teams using Wilcoxon Signed Rank Test. In each comparison, the 'statistically-better' team (p-value < 0.001) gets one point. After all comparisons, each team has a 'total point count' indicating the number of comparisons this team was better than its counterparts.
Step 3 (Leave-one-out): Drop one scan in the test set and repeat Step 2 for the remaining scans. This results in one 'total point count' per team for every dropped scan.
Step 4: Combine the 'total points counts' to determine the ranking.

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

Missing cases will be assigned 'extreme' or most penalising values according to the metric. For example: Id. rate of a missed vertebra is 0.0 while its localisation distance would be 1000 mm. These values are also employed at a scan-level for missed cases.

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

Our ranking method from VerSe'19 has received positive feedback from the participants (due to its stability to outlying performances). It is similar to the one used in BraTS [7] and Medical Segmentation Decathlon [8]. These tests are inspired from Maier-Hein et al. [9].

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.

Note that our ranking scheme in inherently based in a statistically significant difference between team's performance measures (cf. Section 27a). We employ Wilcoxon signed-rank test (with 'greater' or 'lesser' hypothesis as appropriate to the metric), along with leave-one-out sampling, to determine this difference.

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

Usual metrics such a mean or median do not consider the entire distribution of the performance metric value of an algorithm on all the cases. In our case, considering case-level performance as sample of a distribution, we compare distributions and not just their statistics. We believe such an evaluation to be robust and stable.

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.

The ensuing journal article about the challenge will have a detailed analysis on inter-algorithm variability, algorithm-human variability, and an evaluation of the ensemble of algorithms.

TASK: Vertebrae Segmentation

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.

Segment the vertebrae in CT scans with corresponding labels.

Keywords

List the primary keywords that characterize the task.

spine, vertebrae, segmentation, CT

ORGANIZATION

Organizers

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

1. Anjany Sekuboyina
(Informatics & Klinikum rechts der Isar, Technical University of Munich.)

2. Bjoern Menze,
(Informatics, Technical University of Munich.)

3. Jan Kirschke,
(Klinikum rechts der Isar, Technical University of Munich.)

b) Provide information on the primary contact person.

Anjany Sekuboyina
(email: anjany.sekuboyina@tum.de)

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).

verse2020.grand-challenge.org/

Participation policies

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

Fully automatic, Semi automatic.

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.

Publicly available 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.

Prospective sponsorship from NVIDIA (similar to VerSe'19).

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 5 performing submissions are announced at the challenge. Detailed analysis of performances is available upon request.

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).

1. At least two authors from every submission (first and last) listed in the challenge paper will be included in the authors list. Additional authors will be included upon request with justification according to ICMJE authorship guidelines.

2. Participating teams can submit their results separately without any embargo. The challenge paper will focus on the evaluation and will only include an overview of the individual submissions.

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.

Evaluation will be done in two phases with different data for both phases:

Phase 1 (image data public and annotations private): Participants submit predictions on test set images.

Phase 2 (image data private and annotations private): Participants submit docker containers which are run by organisers in-house on hidden test set images.

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.

Participants will have access to phase-1 test set. Predictions on this data can be submitted to grand-challenge.org to get them instantly evaluated. However, in order to prevent overfitting, the number of these submissions is limited to one per day.

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 of Training data: May 15th, 2020
- Release of public Test data: July 20th, 2020
- Opening of submission system: July 20th, 2020
- Closing of submission system for public test data: July 24th, 2020
- Closing of submission system for dockers: Aug 7th, 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).

Ethics approval by the local ethics committee of the University Hospital of TUM granted 18.3.2019, No. 27/19/S-SR.

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 SA.

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.

The evaluation is already available online on grand-challenge.org's evaluation portal. We will release the python code for all systems on grand-challenge.org at the time of training data release.

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

Participants own the rights on distribution of their code. We will not re-distribute their code.

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.

Funding Agency: European Research Council.
Funding of the prize: NVIDIA.
Only the main organizers and their local annotation team will have access to all test labels and the private test datasets.

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

Intervention planning, Decision support, Treatment planning, Diagnosis, Screening, Assistance, Longitudinal study.

Task category(ies)

State the task category(ies).

Examples:

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

Segmentation.

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.

All patients that get a CT scan of the trunk or any part of the spine.

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

All patients that get a CT scan of the trunk or any part of the spine with a particular focus on anatomical anomalies, such as L6, sacralised L5, C7 with cervical ribs.

Imaging modality(ies)

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

Computed tomography with and without contrast from different scanners, institutions, and acquisition settings.

Context information

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

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

No additional context information will be given.

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

No additional context information will be given.

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.

Computed tomography (CT) data of the trunk and neck, including parts of the spine or the entire spine.

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.

All vertebra of the spine.

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)).

Hardware requirements, Specificity, Accuracy, Sensitivity, Precision, Robustness.

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).

All data was acquired on state of the art Multislice CT scanners from 4 different vendors: Siemens, Philips, Toshiba and GE.

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

All clinically-used routine image protocols have been used, including scans with and without i.v. and oral contrast, different kVp settings, and adaptive tube load.

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 was acquired at the University hospital TUM (60%) and multiple different smaller institutions (30%). Public data from a previous CSI 2014 labelling challenge (10%) will also be included.

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).

All patients above 18 years of age are included in the data acquisition process.

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 CT image stack of a human body. Training cases have vertebra-specific voxel-level segmentation masks while the test cases include the images only.

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

Training data: 100 CT scans (images & masks are public)
Public test data: 100 CT scans (images public & masks are hidden)
Private test data: 100 CT scans (images & masks are hidden)

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

From our point of view, this was the best compromise between:
- presenting a balanced dataset dealing with multi-centre acquisitions and different anomalies.
- a dataset of a size appropriate enough to avoid overfitting while illustrating the differences in algorithmic novelty.
- acceptable effort in manual annotation.

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.

In this iteration of VerSe, we focused on two points:
- 50% of the data is multi-centre data
- 50% of the data includes anomalies

Based on the experience with VerSe'19, current state of the art algorithms perform poorly in spinal anomalies. As anatomical anomalies are severely underrepresented in non-selective cross sectional data, it can be expected that learning based approaches under-perform on these. Also, the algorithms' generalisability to multi-centre data is unknown. We believe the composition of our dataset and its size will help address these issues.

CT data included are consecutive patients identified in retrospective searches in the PACS limited to the last five years. We include all scans, except CT-myelographies with contrast material present in the spinal canal.

Note 1: VerSe'20 builds on the VerSe'19 dataset. To be able to present 50% multi-centre data, 50% anatomical anomalies, we were able to include 62 scans from VerSe'19 among the 300 scans for VerSe'20 (without data-

leakage, i.e. a test scan remains test scan and a training scan remains a train one). As the training data of VerSe'19 is publicly available, users can always use the scans not included in VerSe'20 (44 scans).

Note 2: As scans with anomalies are rare, we will have to take those from our local database (University hospital TUM), with 7 different scanners from two vendors, overlapping with VerSe'19.

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.

We use a human hybrid algorithm, based on our automatic labelling and segmentation algorithm. The segmentation is corrected by one experienced medical student and re-checked and again corrected consecutively by one radiologist with 5 years and by one neuroradiologist with 17 years of experience in spine imaging. The final decision is made by the most experienced reader, in consensus with both other raters.

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.

The annotator was instructed to segment the vertebrae at the exact outline, including all posterior elements and osteophytes, while excluding cavities like Schmorl's nodes, or implants like cages or screws. Bone cement will be annotated when included within the vertebra, but will be excluded when outside of the vertebra. Syndesmophytes shall be included in the label of the respective vertebra up to the mid-IVD-level (IVD: Inter-vertebral disc).

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.

Algorithm: Uses a three-stage interactive pipeline with spine localisation, vertebrae labelling (Btrfly Net [3]), and vertebrae segmentation (U-Net based architecture). The masks predicted by the segmentation stage are used as initialisation for human annotation process.

Humans: One specifically trained medical student, one radiologist with 5 years of experience for direct supervision, one neuroradiologist with 17 years of experience in spine imaging for final approval. This annotation process is organised hierarchically.

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.

N/A

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.

In case of sub-mm resolution, the spatial resolution will be downsampled to 1mm in those dimensions.

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.

Possible error sources include cases with severe degeneration like syndesmophytes, where no clear differentiation is possible between two adjacent vertebrae. However, as only one primary annotator is involved and all cases are checked by one experienced neuroradiologists, variations will be minimal, according to proir experience <1% DICE.

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

N/A

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)

Dice Similarity Coefficient (DSC)
Hausdorff Distance (HD)

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

1. Dice Coeff. (DSC) is a standard measure quantifying segmentation performance.
2. In cases of complete segmentation failure (as the vertebra needs to be appropriately identified too), when DSC is 0, Hausdorff Distance (HD) help in capturing the degree of failure in terms of distance from ground truth. Moreover, HD is more sensitive to noisy segmentation.

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.

Following VerSe'19 challenge and Brain Tumor Segmentation (BraTS) challenge: We follow a 'point' based evaluation system as described in Sekuboyina et al. [2]. Added to this, we perform bootstrapping to make the ranks more robust. This results in a ranking scheme detailed below:

Step 1: Compute metrics at a scan level (DSC and HD).
Step 2: For each metric: Compare every possible pair of teams using Wilcoxon Signed Rank Test. In each comparison, the 'statistically-better' team (p-value < 0.001) gets one point. After all comparisons, each team has a 'total point count' indicating the number of comparisons this team was better than its counterparts.
Step 3 (Leave-one-out): Drop one scan in the test set and repeat Step 2 for the remaining scans. This results in one 'total point count' per team for every dropped scan.
Step 4: Combine the 'total points counts' to determine the ranking.

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

Missing cases will be assigned 'extreme' or most penalising values according to the metric. For example: DSC of a missed vertebra is 0.0 while its Hausdorff distance would be 100 mm. These values are also employed at a scan-level for missed cases.

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

Our ranking method from VerSe'19 has received positive feedback from the participants (due to its stability to outlying performances). It is similar to the one used in BraTS [7] and Medical Segmentation Decathlon [8]. These tests are inspired from Maier-Hein et al. [9].

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.

Note that our ranking scheme in inherently based in a statistically significant difference between team's performance measures (cf. Section 27a). We employ Wilcoxon signed-rank test (with 'greater' or 'lesser' hypothesis as appropriate to the metric), along with leave-one-out sampling, to determine this difference.

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

Usual metrics such a mean or median do not consider the entire distribution of the performance metric value of an algorithm on all the cases. In our case, considering case-level performance as sample of a distribution, we compare distributions and not just their statistics. We believe such an evaluation to be robust and stable.

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.

The ensuing journal article about the challenge will have a detailed analysis on inter-algorithm variability, algorithm-human variability, and an evaluation of the ensemble of algorithms.

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

[1] VerSe'19 challenge page: verse2019.grand-challenge.org

[2] Sekuboyina et al., 'VerSe: A Vertebrae Labelling and Segmentation Benchmark', 2020. (arxiv.org/abs/2001.09193)

[3] Labelling algorithm: Sekuboyina et al., 'Btrfly net: Vertebrae labelling with energy-based adversarial learning of local spine prior', In: MICCAI 2018.

[4] Dataset Description for VerSe'19: Loeffler et al., 'Large Scale Vertebral Segmentation (VerSe) Dataset with Fracture Grading'. In: Radiology:Artificial Intelligence (In Press), 2020.

[5] Glocker et al., 'Vertebrae localization in pathological spine ct via dense classification from sparse annotations'. In: MICCAI 2013.

[6] VerSe'19 public data: https://osf.io/nqjyw/ (Release of the VerSe'20 data will be performed similarly)

[7] Menze et al., 'The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)'. In: IEEE Trans Med Imaging, 2014.

[8] medicaldecathlon.com

[9] Maier-Hein et al., 'Why rankings of biomedical image analysis competitions should be interpreted with care'. In: Nature Communications, 2018.

[10] Thawait, Chhabra, & Carrino, ' Spine segmentation and enumeration and normal variants', In: Radiol Clin North Am, 2012. doi:10.1016/j.rcl.2012.04.003.

[11] Castellvi, Goldstein, Chan, 'Lumbosacral transitional vertebrae and their relationship with lumbar extradural defects', In: Spine (Phila Pa 1976), 1984.

[12] O'Driscoll et al., 'Variations in morphology of the lumbosacral junction on sagittal MRI: correlation with plain radiography', In: Skeletal Radiol, 1996.

Further comments

Further comments from the organizers.

We thank the reviewers for taking the time to evaluate our proposal.