2021 Kidney and Kidney Tumor Segmentation Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

2021 Kidney and Kidney Tumor Segmentation Challenge

Challenge acronym

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

KiTS21

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.

There is currently a great interest in quantitatively studying the morphology of kidney tumors in order to better characterize surgical complexity and inform treatment planning. Semantic segmentation is a powerful tool for this, but it requires expert reading and considerable manual effort. The KiTS19 challenge introduced the first large-scale public dataset of kidney and kidney tumor semantic segmentations, representing a considerable step towards reliable automatic segmentation of these structures. Unfortunately, it was limited in both the scope of the dataset and the structures that were annotated. The goal of the KiTS21 challenge is to address these limitations by incorporating data from disparate geographical locations and acquisition times, and by providing segmentation labels for more extensive anatomical structures such as the ureters and renal vessels. This will enhance both the clinical utility of the resulting methods, as well as the technical challenge for participants.

Challenge keywords

List the primary keywords that characterize the challenge.

Semantic Segmentation; Kidney Tumors; Computed Tomography

Year

The challenge will take place in ...

2021

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.

In 2019, we had more than 100 teams who participated in the challenge. We expect the 2021 participation to be similar.

Publication and future plans

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

We plan to coordinate a publication of challenge results with the top five participating teams, as we did in 2019.

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.

We do not have any on-site requirements.

TASK: Segmentation of Kidney and Associated Structures

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.

The purpose of this task is 3D semantic segmentation of kidneys, renal cysts, kidney tumors, ureters, and renal vessels in computed tomography imaging.

Keywords

List the primary keywords that characterize the task.

Semantic Segmentation, 3D Segmentation, Kidney Tumors

ORGANIZATION

Organizers

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

Nicholas Heller (University of Minnesota)
Nikolaos Papanikolopoulos (University of Minnesota)
Christopher Weight (University of Minnesota)

b) Provide information on the primary contact person.

Nicholas Heller (helle246@umn.edu)

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).

None yet.

Participation policies

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

Fully 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 not participate.

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

As in 2019, an industry partner will be providing a cash prize for the highest-scoring team. The details of this will be announced at a later date.

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.

Scores and rankings will be announced publicly for all participating teams.

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).

Participating teams are required to submit a manuscript to accompany their methods, and these manuscripts will be published as a standalone collection, similar to a satellite event proceedings. The top five scoring teams will, in addition, be included on a central journal submission describing the findings of the challenge, to be submitted prior to the end of the 2021 calendar year. There will be no embargo period.

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.

Submissions will be made via Docker container on Sage Bionetworks' Synapse Platform.

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.

Within 24 hours of submitting, participants will receive an email prompting whether or not they would like to hear their "approximate" score. Approximate scores will provided only twice to each team, but they may keep submitting after receiving two scores. The most recent submission prior to the deadline will be the one used for the competition.

Scores reported by this mechanism are "approximate" in that they are calculated using only on a randomly selected 20 of the test cases. This random sampling is repeated with each submission.

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

Training data will be released with annotations as they become available, but the entire training set will be made public no later than March 1, 2021.
Registration for the challenge will open on grand-challenge.org on January 1, 2021
All submissions and evaluation will happen via Docker container, therefore no test cases will be released.
Submissions will be allowed for the three weeks prior to challenge deadline of August 1, 2021

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).

This study was reviewed an approved by the Institutional Review Board of the University of Minnesota under Study 1611M00821 until October 31, 2020. We expect that it will be renewed next fall as it has been for the past three years.

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 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.

Evaluation code will be made available with the training data.

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

Participating teams are strongly encouraged but not required to make their code publicly available.

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.

The challenge organizers only will have access to the test case labels. The challenge is funded by the National Cancer Institute of the National Institutes of Health under award R01CA225435. The organizers have no conflicts of interest.

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

CAD, Segmentation.

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.

Patients found to have a solid renal tumor.

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

Patients at our institution as well as at least one partner institution who were found to have a solid renal tumor between a well-defined time period which varies from institution to institution.

Imaging modality(ies)

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

Contrast-enhanced Computed Tomography (CT)

Context information

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

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

The image data will be released in as compressed Nifti1Image objects stored in .nii.gz files. These files will include an "affine matrix" defining the voxel size of each image.

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

No additional information regarding each patient will be released.

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.

CT including abdomen. Occasionally the chest and/or pelvis as well.

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.

Kidney, kidney tumor, kidney cyst, renal vessels, ureters

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)).

Sensitivity, Precision.

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).

Various brands of CT scanners from hundreds of referring institutions

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

Acquisition occurred as part of routine clinical practice and therefore varies with referring institution and physician.

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.

The central organizing institution is the University of Minnesota, and a significant proportion of the data will come from there. Much of this data was released for the 2019 iteration of this challenge, but this cohort will be expanded by roughly 30-50%.

Data use agreements are currently being negotiated with several partner institutions. At least one yet to be announced partner institution will be providing at least as many cases as the University of Minnesota cohort.

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).

Data annotation is overseen by the challenge clinical chair, Dr. Christopher Weight. Dr. Weight is an experienced urologic cancer surgeon who specializes in treating kidney tumors.

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.

One case, in this challenge, is a single CT image associated with a semantic segmentation mask.

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

The total number of cases is yet unknown, but the estimated number is 800. A 70%-30% train-test split will be employed resulting in 560 training cases and 240 test cases.

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

The total number of cases was chosen because it is the largest number that we expect to be able to annotate between now and the data release. The 70-30 split was chosen because it is standard practice in machine learning, and in line with the previous edition of this challenge.

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.

A summary of demographic characteristics of these cases is important in order to judge the external validity of the results. These characteristics are unknown at this time but will be thoroughly reported in a data-description manuscript that will be made available prior to data release.

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.

Our procedure for annotating kidneys and kidney tumors has been reported previously (arxiv: 1904.00445). Our procedure for annotating cysts, ureters, and renal vessels is currently under development and will be thoroughly reported in the aforementioned data description manuscript.

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.

Please refer to our response to question 23a.

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.

The annotation process will be guided by Dr. Christopher Weight who has more than 15 years of medical experience.

The "legwork" will be performed by medical students (M1-M4) working in conjunction with crowdsourced workers with (presumed) no medical training.

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.

The multiple annotations on training cases will not be aggregated. All annotations for each training case will be released, and participants may or may not choose to aggregate the labels.

For the test set, voxel-wise majority voting will be used for aggregation and subsequent evaluation. The source code used for this procedure will be released with the training set.

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.

Data will not be pre-processed except for what is required for de-identification.

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.

Error in unavoidable in medical image segmentation. We plan to employ our strategy of 2019 of soliciting input from participants about possible errors that they notice in the dataset.

An analysis of inter-rater variability in our 2019 cohort found agreement (Dice) of ~0.98 for kidney and ~0.92 for tumors. We expect that of cysts to be on par with (if not slightly higher than) tumors, and we do not know what to expect for renal vessels and ureters. This, too, will be reported when known.

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

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)

As in 2019, we will be using the metric of average Dice Similarity Coefficient (DSC) across all classes across all cases.

The methods of [1] will be used to assign a score to each team with some slight modifications:

a. The methods will be used for each segmentation region, and scores will then be averaged -- this is necessary since we have more than two segmentation classes
b. Rather than setting epsilon_i bar to the average single difference between two raters, it will be set to the average difference between all pairs of annotations -- this is necessary since we have more than two annotations per case.

[1] Heimann, Tobias, et al. "Comparison and evaluation of methods for liver segmentation from CT datasets." IEEE transactions on medical imaging 28.8 (2009): 1251-1265.

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

As stated in [1] each of the five metrics measures a difference aspect of the quality of the prediction, and thus using them in combination will provide a robust measure of quality.

[1] Heimann, Tobias, et al. "Comparison and evaluation of methods for liver segmentation from CT datasets." IEEE transactions on medical imaging 28.8 (2009): 1251-1265.

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.

Simple ranking by score will be used (see 26a for score).

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

Missing results will be assigned a score of zero during mean aggregation.

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

In general, in order to earn a high average score, one needs to perform well. We understand that other alternatives have been proposed to a simple ranking of averages, but we again believe that simplicity and transparency outweigh the potential benefits to these methods.

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.

Missing data handling has been addressed previously. Ranking variability will be characterized using the bootstrap. This analysis will be performed in Python 3.7.

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

The Bootstrap is a simple nonparametric method that relies on minimal assumptions.

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.

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.

Heller, Nicholas, et al. "The KiTS19 Challenge Data: 300 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes." arXiv preprint arXiv:1904.00445 (2019).

Heller, Nicholas, et al. "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 Challenge." arXiv preprint arXiv:1912.01054 (2019).