Foot Ulcer Segmentation Challenge 2021: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Foot Ulcer Segmentation Challenge 2021

Challenge acronym

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

FU-Seg

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.

Acute and chronic nonhealing wounds represent a heavy burden to healthcare systems, affecting millions of patients around the world [1]. Accurate measurement of wound areas is an important part of the diagnosis and care protocol since it is crucial to provide quantitative parameters to monitor the wound healing trajectory and to determine future interventions. Unfortunately, Manual measurement is time-consuming and often inaccurate which can cause a negative impact on patients. Lack of expertise can also lead to improper diagnosis of wound etiology and inaccurate wound measurement and documentation. Wound segmentation from images is a popular solution to these problems that not only automates the measurement of the wound area but also allows efficient data entry into the electronic medical record to enhance patient care.
With the collaboration between the University of Wisconsin-Milwaukee and Advancing the Zenith of Healthcare Wound and Vascular Center, we build a dataset containing over 1000 foot ulcer images professionally labeled with binary masks. We provide this dataset to the challenge and aim at encouraging and supporting the development of new solutions for the automated and accurate segmentation of foot ulcers from natural images taken in common clinical settings.

Challenge keywords

List the primary keywords that characterize the challenge.

Foot Ulcer, Semantic Segmentation

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.

This is a challenge proposed for the first time. We have never published the data nor received requests for the dataset yet.

Publication and future plans

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

We plan to publish a paper on the summary of the challenge results. We plan to continue collecting and labeling new images in the future.

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.

Our challenge will be hosted on grand-challenge.org.

TASK: Foot Ulcer Segmentation

SUMMARY

Keywords

List the primary keywords that characterize the task.

Foot Ulcer, Semantic Segmentation

ORGANIZATION

Organizers

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

Chuanbo Wang, University of Wisconsin-Milwaukee
Behrouz Rostami, University of Wisconsin-Milwaukee
Jeffrey Niezgoda, Advancing the Zenith of Healthcare Wound and Vascular Center
Sandeep Gopalakrishnan, University of Wisconsin-Milwaukee
Zeyun Yu, University of Wisconsin-Milwaukee

b) Provide information on the primary contact person.

Zeyun Yu
University of Wisconsin-Milwaukee
yuz@uwm.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.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Open call challenge.

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).

https://fusc.grand-challenge.org/

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.

There is no cash prize for now.

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.

All results will be made available through the planned publication and the top 3 teams will be invited to present their work.

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).

The first author and the corresponding author of each participating team will qualify as authors. The participating teams may publish their own results separately and there are no restrictions on the timing of publishing their own results.

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.

Participating algorithms need to be submitted as Docker containers. We will evaluate the performance of participating teams by applying the Docker containers to our testing set. The participating teams need to provide us a Docker container via email or a link to the project repository containing a Docker container. For example, a GitHub repository URL is welcomed.

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.

A validation set will be provided for sanity checks and evaluating algorithms before submitting final results.

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 set: Late February 2021
Release of validation set: Late February 2021
Final Docker container submission: August 1, 2021
Results announcement: Late August 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).

Our data is completely de-identified by removing personal identifiers defined by HIPAA. The IRB review and approval are waived by the Institutional Review Board of University of Wisconsin-Milwaukee.

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.

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 code will be made openly available on GitHub and grand-challenge.org

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

We encourage participating teams to publish codes and files on GitHub. Foot Ulcer Segmentation Challenge 2021 and MICCAI 2021 should be mentioned in the repository.

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.

We claim no conflict of interest. The work under this challenge is supported by the University of Wisconsin-Milwaukee and the AZH Wound Center. Anyone participating in the challenge or not will have access to the test case labels as soon as the proposal is accepted.

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

Prognosis, Decision support, Research, Treatment planning.

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.

1. Patients with foot ulcers, mainly chronic ulcers caused by diabetes, venous insufficiency, surgeries, or long-term pressure.
2. Care providers like physicians and nurses.

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

Any patients visited the AZH clinic in a two-year period with foot ulcers, venous insufficiency, surgeries, or long-term pressure. There are no further inclusion or exclusion criteria and we included patients of all ages, genders, etc. In total, there are more than 800 patients included in the dataset.

Imaging modality(ies)

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

Photography.

Context information

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

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

Foot ulcer images.

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

Patients with foot ulcers, mainly chronic ulcers caused by diabetes, venous insufficiency, surgeries, or long-term pressure.

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.

Foot showed in the photography of foot ulcer patients.

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.

Ulcers on feet.

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)).

Specificity, Sensitivity, 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).

A Canon SX 620 HS camera and an iPad Pro were used for taking pictures of foot ulcers.

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

The images are taken as close-ups with the camera facing the wound surface. To ensure adequate illumination, flashes are used whenever possible.

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.

Foot ulcer images were collected from the Advancing the Zenith of Healthcare Wound and Vascular Center.

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).

Images were collected by wound specialists and nurses during clinic visits to keep track of the healing 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.

A case in our dataset represent an foot ulcer image. Each case is fully annotated with a binary mask of the same

size as the image where non-wound or background pixels are marked with zero values and wound pixels are marked with 255.

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

Training: 610 images.
Validation: 200 images.
Testing: 200 images.

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

Currently, we have 1010 images fully annotated since labeling segmentation masks are time-consuming. In our previous research [1], we have tested popular deep segmentation neural networks with approximately 500 images and 1000 images, and the accuracy is similar. These neural networks include U-Net [2], Mask-RCNN [3], and MobileNetV2 [4]. Hence, we conclude that 1000 images are sufficient for the foot ulcer segmentation problem. We follow the Pareto principle that randomly splits with an 80/20 ratio for training and testing. The training set is further randomly split with a 60/20 ratio for training and validation so that the validation set has the same size as the testing set.

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.

The data were collected in a common clinical setting over the past few years and the characteristics follow real-world distribution. Each case in the dataset contains at least one wound. In some rare cases, the wound just healed and got photographed during the last few clinic visits and there will be no wound labeled in the annotation for these cases.

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.

All the images were manually annotated with segmentation masks and further reviewed and verified by wound care specialists and nurses from the Advancing the Zenith of Healthcare Wound and Vascular Center.

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.

To solve this two-class segmentation problem, the algorithms only need to classify each pixel into two categories, wound or non-wound. Thus, the only instruction we provided to the wound care specialists and nurses is to label each pixel as wound pixel or non-wound pixel. We closely worked with the wound care clinic to annotate the dataset. The special or edge cases were mainly done by the specialist from the clinic described in detail in item 23c).

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 annotators from the collaborating clinic are the wound professionals who took/are taking care of the patients that appear in our dataset. The first round of rough annotation was proposed by grad students in our lab. The rough annotations were then reviewed and refined by nurses and wound specialists from the clinic who have 2-5 years of professional experience. In special or edge cases, the nurses and specialists consult with either a) the doctors who took care of the specific patients during their visits, or b) Dr. Jeffrey Niezgoda, MD who supervises the annotation team and has over 30 years of professional experience. The wound images are annotated using Adobe Photoshop where the annotations are marked on a different layer on top of the image layer. Finally, the refined annotations were pre-processed by our lab to form the dataset.

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.

There is only one annotation for each image in our dataset. The images are annotated based on mutual agreement between all annotators and all images are annotated following the same protocol.

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.

The images are zero-padded to 512 by 512. The raw training date will not be pre-processed further and the participating teams can choose any pre-processing methods of their preference.

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.

Fibering, eschar, epithelial, and granulation tissues are considered part of the wound in the annotations of our dataset. The most common error is caused by the unclear boundary of wounds/ulcers due to epithelial tissue travels from the outward wound edges.

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

Another source of error that confuses the annotators appears when exposed bone structures and fibering exist in the wound simultaneously. They appear in very similar colors and could be annotated incorrectly. In the case of disagreement, annotators generate the final annotation together based on mutual agreement after consultation with the doctors.

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) / Intersection over Union Rate(IOU)

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

DSC shows the similarity between the segmentation and the ground truth. In our dataset, the areas of wound/ulcer are imbalanced comparing to the background. We choose DSC because it is computed by the harmonic mean of Precision and Recall and finds a good balance between evaluating false positives and false negatives. For non-wound/healed cases, all pixels predicted to be wound will be processed as false positives. For example, A perfect prediction of such cases is an image with every pixel labeled as non-wound/background with zero intensity.

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.

To compute the final ranking, submitted algorithms will be ranked based on the average DSC of all testing images. In the case of ties, pixel-wise accuracy will be used to rank the participants.

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

For submissions with missing results, all missing cases will be given the DSC of 0 since no true positives are predicted.

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

Images in our dataset are labeled with binary masks and there is only one class to be segmented. Thus, segmentation accuracy is the only measurement of the participating teams' performance. We will rank the participating algorithms base on average DSC scores for simplicity.

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.

For computing the final ranking, we will use bootstrap to take account of the ranking variability.

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

Bootstrapping works well with small sample sizes.

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.

[1] C. Wang, D.M. Anisuzzaman, V. Williamson, M.K. Dhar, B. Rostami, J. Niezgoda, S. Gopalakrishnan, and Z. Yu, "Fully automatic wound segmentation with deep convolutional neural networks", accepted for publication in Scientific Reports.

[2] R. G. Frykberg and J. Banks. "Challenges in the treatment of chronic wounds. Advances in wound care. 4, no. 9, 560-582 (2015).

[3] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention. 234-241 (2015).

[4] W. Abdulla, Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. Git code https://github.com/matterport/Mask_RCNN (2017)

[5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L-C Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4510-4520 (2018).