# Ultrasound Image Enhancement challenge 2023: Structured description of the challenge design

## CHALLENGE ORGANIZATION

### Title

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

Ultrasound Image Enhancement challenge 2023

### Challenge acronym

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

USenhance 2023

### 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.

We propose to hold the challenge of enhancement for ultrasound images in conjunction with MICCAI 2023. We will provide various ultrasound data of multiple organs, including the thyroid, carotid artery, abdomen, and breast. The challenging task of reconstructing high-quality ultrasound images from low-quality ones will be conducted. A total of 3000 ultrasound images (1500 pairs of low- and high-quality images) from 109 patients will be provided in the challenge.

Ultrasound imaging is commonly used for aiding disease diagnosis and treatment, with advantages in noninvasive. Lately, medical ultrasound shows prospects revolving from expensive big-size machines in hospitals to economical hand-held devices in wider use. The barrier is that ultrasound examination with a handheld device has the drawback of low imaging quality due to hardware limitations. Toward this, ultrasound image enhancement provides a potential low-cost solution. Restoring high-quality images from low-quality ones using computer algorithms would exempt requirements for hardware improvements and promote ultrasound device revolutions and wider applications.

By releasing the training data for ultrasound imaging enhancement, running the online evaluation server, and holding the challenge talks, the USenhance 2023 is expected to attract much attention from the research community, and advance the research on high-quality ultrasound imaging significantly.

### Challenge keywords

List the primary keywords that characterize the challenge.

Ultrasound, image quality enhancement, multiple organs

### Year

The challenge will take place in ...

2023

## FURTHER INFORMATION FOR MICCAI ORGANIZERS

### Workshop

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

### 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.

### Publication and future plans

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

### 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.

# TASK: Ultrasound Image Enhancement

## SUMMARY

### Keywords

List the primary keywords that characterize the task.

## ORGANIZATION

### Organizers

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

Yi Guo, Fudan University, China
Shichong Zhou, Fudan University Shanghai Cancer Center, China
Jun Shi, Shanghai University, China
Yuanyuan Wang, Fudan University, China

b) Provide information on the primary contact person.

Yi Guo, Fudan University, China

### 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

Repeated event with fixed submission deadline.

### 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.

As stated in our responses to reviewers' comments, we will use grand-challenge.org for the challenge host, information release, and potential result discussions.

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

None at this moment

### 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.

No additional data 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 stated in our responses to reviewers' comments, top-10 teams will be invited to present talks and winning certificates will be provided.

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-10 performing results will be made public.

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

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

All participating team members are qualified as authors. Participating teams may publish their own results separately.

## 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.

Submission instructions will be on webpage. We will let the participants to submit their Docker submissions to our evaluation server.

The participating teams will use portion of the training data as the validation set.

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.

Submission of multiple results is allowed. Only the last run is officially counted to compute challenge results.

As stated in our responses to reviewers' comments, during the test stage, we will provide participants with a maximum of 10 submissions in phase 1 and 2 submissions in phase 2. This should allow the participants to obtain a functional Docker container after phase 1 and submit either 1 or 2 final algorithms for the final leaderboard.

## 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: July. 10th, 2023;
Submission deadline for results: Sept. 10th, 2023;
Announcement of final results: Sept. 15th, 2023.

## 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).

Approval on data from USenhance 2023 has been obtained from all participants.

## 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.

Evaluation code will be made public.

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

As stated in our responses to reviewers' comments, there are no requirements for participating teams to disclose their code. However, every team is acquired to attach a license agreement.

### 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.

None

## 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

Diagnosis, Screening.

### Task category(ies)

State the task category(ies).

Examples:

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

Reconstruction.

### Cohorts

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

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

Volunteers with potential thyroid tumors, carotid plaque or breast cancer.

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

The challenge cohort included healthy volunteers and volunteers with potential thyroid tumors, carotid plaque, or breast cancer.

### Imaging modality(ies)

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

Ultrasound devices were used to scan the organs of volunteers, which causes no radiation to them.

### Context information

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

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

None

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

None

### 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.

Thyroid (ultrasound images);
Carotid artery (ultrasound images);
Abdomen (ultrasound images);
Breast (ultrasound images).

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.

Thyroid (ultrasound image quality);
Carotid artery (ultrasound image quality);
Abdomen (ultrasound image quality);
Breast (ultrasound image quality).

### 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)).

Runtime.

Additional points: Quality of reconstructed images.

# 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).

1. Low-end / High-end ultrasound devices for thyroid imaging: mSonics MU1 / Toshiba Aplio 500;
2. Low-end / High-end ultrasound devices for carotid artery imaging: SSUN / Toshiba Aplio 500;
3. Low-end / High-end ultrasound devices for abdomen imaging: SSUN / Toshiba Aplio 500;
4. Low-end / High-end ultrasound devices for breast imaging: mSonics MU1 / Aixplorer ultrasound system (SuperSonic Imaging S.A.);

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

Ultrasound imaging

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.

Fudan University Shanghai Cancer 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).

Experienced sonogarphers (well medically-trained).

## 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 means one pair of images.

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

A total of 3000 ultrasound images (1500 pairs of low- and high-quality images) from 109 patients will be provided in the challenge. The training set contains 1050 pairs. The testing set contains 450 pairs.
As stated in our responses to reviewers' comments, details are:
(1) There are 517 thyroid ultrasound image pairs from 47 people (11 pairs/person).
Training/Testing: 363 pairs from 33 people / 154 pairs from 14 people.
Low-end portable device / High-end device: mSonics MU1 / Toshiba Aplio 500.
(2) There are 535 carotid artery ultrasound image pairs from 77 people (6~7 pairs/person).
Training/Testing: 375 pairs from 54 people / 160 pairs from 23 people.
Low-end portable device / High-end device: SSUN 100 / Toshiba Aplio 500, respectively.
(3) There are 300 abdomen ultrasound image pairs from 30 people (10 pairs/person).
Training/Testing: 210 pairs from 21 people / 90 pairs from 9 people.
Low-end portable device / High-end device: SSUN 100 / Toshiba Aplio 500, respectively.
(4) There are 148 breast ultrasound image pairs from 32 people (4~5 pairs/person).
Training/Testing: 102 pairs from 23 people / 46 pairs from 9 people.
Low-end portable device / High-end devices: mSonics MU1 / Aixplorer ultrasound system (SuperSonic Imagine S.A.).

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

A large-scale dataset for ultrasound image enhancement.
70% training data and 30% test data.

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.

As stated in our responses to reviewers' comments, the train/test was split stratified over the target anatomies.

### 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.

Paired ultrasound images were acquired by an experienced sonographer, assisted by another one.

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.

Sonographers have been well-trained to acquire data using both low- and high-end devices.

Volunteers held their breath for about 10 seconds when being scanned by different devices to reduce the deformation. In the scanning process, landmark points were recorded on volunteers' scanning positions for the landmark-based nonrigid registration to obtain paired data pairs.

Specifically, landmark points were manually annotated by experienced experts. In addition to the measuring position and scanning angle, deformation caused by respiratory motion was compensated. That is, a B-spline-based nonrigid registration method [1] and a landmark-based nonrigid registration method [2] were adopted to handle all clinical training pairs. After registration, the deviation between the low-quality training samples and the corresponding reference images was reduced.

[1] R. Shekhar et al., “Mutual information-based rigid and nonrigid registration of ultrasound volumes,” IEEE Trans. Med. Imaging, vol. 21, no. 1, pp. 9-22, Jan. 2002.
[2] S. Klein et al., “Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines,” IEEE Trans. Image Process. vol. 16, no. 12, pp. 2879-2890, Dec. 2007.

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.

Experienced sonogarphers (well medically-trained).

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.

None

### 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.

Patient-identity information is removed from data to protect patient privacy.

### 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.

Slight deformation of images caused by slight probe movement, which has been reduced as much as possible by post registration.

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

None.

## 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 stated in our responses to reviewers' comments, metrics include:
(Peak signal-noise ratio) PSNR,
(structure similarity) SSIM,
(mutual information) MI,

(Locally-normalized-cross-correlation) LNCC,
runtime,
(the lesion segmentation Dice on the enhanced images using a pre-trained nnU-Net) Dice

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

They are widely used to measure image reconstruction quality.
As stated in our responses to reviewers' comments, these evaluation metrics adopted to access the quality of the enhanced images includes three aspects. First, PSNR and MI are used to evaluate the global similarity of enhanced images and target images. Second, SSIM is employed to measure the detailed similarity of images. Third, Dice coefficient is provided to comprehensively evaluate the clinical utility of investigated algorithms.

### 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.

As stated in our responses to reviewers' comments, for each test case, respectively calculate their 1) SSIM, 2) MI, 3) PSNR, 4) LNCC, 5) runtime and 6) Dice. Then, we will conduct statistical tests of robustness, e.g. to test for variability of ranking with Kendall's tau analysis.

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

Missing result is not allowed.

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

The ranking method has been used in several related publications.

[1] Zhou Z, Guo Y, Wang, Y. Ultrasound deep beamforming using a multiconstrained hybrid generative adversarial network. Medical Image Analysis, 2021, 71: 102086.
[2] Zhou Z, Guo Y, Wang, Y. Handheld ultrasound video high-quality reconstruction using a low-rank representation multipathway generative adversarial network. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(2): 575-588.
[3] Zhou Z, Wang Y, Guo Y, Qi Y, Yu J. Image quality improvement of hand-held ultrasound devices with a two-stage generative adversarial network. IEEE Transactions on Biomedical Engineering, 2020, 67(1): 298-311.
[4] Zhou Z, Wang Y, Guo Y, Jiang X, Qi Y. Ultrafast plane wave imaging with line-scan-quality using an ultrasound-transfer generative adversarial network. IEEE Journal of Biomedical and Health Informatics, 2020, 24(4): 943-956.
[5] Zhou Z, Wang Y, Yu J, Guo Y, Guo W, Qi Y. High spatial-temporal resolution reconstruction of plane-wave ultrasound images with a multichannel multiscale convolutional neural network. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 2018, 65(11): 1983-1996.
[6] Huang L, Zhou Z, Guo Y, Wang, Y. A stability-enhanced CycleGAN for effective domain transformation of unpaired ultrasound images. Biomedical Signal Processing and Control, 2022, 77: 103831.
[7] Liu S, Zhang B, Liu Y, Han A, Shi H, Guan T, He Y, Unpaired stain transfer using pathology-consistent constrained generative adversarial networks, IEEE Transactions on Medical Imaging, 2021, 40(8): 1977–1989.
[8] Kong L, Lian C, Huang D, Li Z, Hu Y, Zhou Q, Breaking the dilemma of medical image-to-image translation. in Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), Oct, 2021.
[9] Guo D, Shamai S, Verdu S. Mutual information and minimum mean-square error in Gaussian channels. IEEE Transactions on Information Theory, 2015, 51(4): 1261–1282.

### 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.

As stated in our responses to reviewers' comments, we will measure the stability and robustness of each algorithm over test cases. For each algorithm, we will conduct statistical tests of robustness, e.g. to test for variability of ranking with Kendall's tau analysis.

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

Algorithm overall performance: The overall performance of the algorithm can be measured by the six metrics over all cases.

Algorithm stability and roustness: As stated in our responses to reviewers' comments, statistical tests of robustness will be conducted, specifically, to test for variability of ranking with Kendall's tau analysis. Kendall’s Tau rank correlation coefficient assesses statistical associations based on the ranks of the data. Ranking data is carried out on the variables that are separately put in order and numbered. This could offer an intuitive way to gain important insights into the relative and absolute performance of algorithms.

### 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.

None

## 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.