PENGWIN: Pelvic Bone Fragments with Injuries Segmentation Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

PENGWIN: Pelvic Bone Fragments with Injuries Segmentation Challenge

Challenge acronym

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

PENGWIN

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.

Pelvic fractures, typically resulting from high-energy traumas, are among the most severe injuries, characterized by a disability rate over 50% and a mortality rate above 13%, ranking them as the deadliest of all compound fractures. The complexity of pelvic anatomy, along with surrounding soft tissues, makes surgical interventions especially challenging. Recent years have seen a shift towards the use of robotic-assisted closed fracture reduction surgeries, which have shown improved surgical outcomes [1]. Accurate segmentation of pelvic fractures is essential, serving as a critical step in trauma diagnosis and aiding in image-guided surgery [2]. In 3D CT scans, fracture segmentation is crucial for fracture classification, pre-operative planning for fracture reduction, and screw fixation planning [3]. For 2D X-ray images, segmentation plays a vital role in transferring the surgical plan to the operating room via registration, a key step for precise surgical navigation [4].

The task of segmenting fractured pelvic fragments is technically challenging due to the diverse shapes and positions of bone fragments and the complex fracture surfaces caused by bone collisions. Despite the success of deep learning in many domains, its application in pelvic fracture segmentation is still limited. This is primarily attributed to the lack of extensive image datasets and annotations specific to fractured pelvises.

The PENGWIN segmentation challenge is designed to advance the development of automated pelvic fracture segmentation techniques in both 3D CT scans (Task 1) and 2D X-ray images (Task 2), aiming to enhance their accuarcy and robustness. Our dataset comprises CT scans from 150 patients scheduled for pelvic reduction surgery, collected from multiple institutions using a variety of scanning equipment. This dataset represents a diverse range of patient cohorts and fracture types. Ground-truth segmentations for sacrum and hipbone fragments have been semi-automatically annotated and subsequently validated by medical experts. Furthermore, we have generated high-quality, realistic X-ray images and corresponding 2D labels from the CT data using the

DeepDRR method, incorporating a range of virtual C-arm camera positions and surgical tools [5]. The primary objective of the PENGWIN challenge is to foster innovative research in image segmentation while establishing a foundational resource for future studies and cross-disciplinary collaborations in pelvis-related research.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Pelvic fracture, Image Segmentation, Computer-assisted orthopedic surgery

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

N/A

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.

The anticipated number of participants for our challenge is approximately 40. This expectation is based on the interest shown in similar challenges within the medical imaging community. For instance, the RibFrac challenge in MICCAI 2020 have attracted 356 registered participants, 243 submissions, and 25,000 downloads from dataset host, demonstrating the community's interests in fracture-related tasks [7]. Also in the field of orthopedics and comparable in dataset size to ours, the VerSe challenges on vertebrae segmentation in MICCAI 2019 and 2020 have attracted 20 submissions in 2019 and about 40 in 2020 [8].

Publication and future plans

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

The challenge organizers will publish at least one challenge journal paper and potentially more. Up to two team members in each of the top three teams for each tasks qualify as authors for joint publication. This manuscript will comprehensively cover various aspects of the challenge, including a detailed description of the methodologies employed to generate the dataset annotations. It will also provide in-depth descriptions of the algorithms developed by the participants, along with a thorough analysis of their performances and distinctive characteristics.

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 will use grand-challenge.org as the online platform.

TASK 1: Pelvic Fracture Segmentation in CT

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.

Pelvic fractures, often resultant of high-energy traumas, are critically serious injuries with a disability rate exceeding 50% and a fatality rate exceeding 13%, making them the most lethal among all compound fractures. The intricate nature of the pelvic anatomy and the encompassing soft tissues make surgical interventions particularly demanding. In recent years, there has been a promising shift towards employing robotic-assisted closed fracture reduction surgery, a development that has demonstrated improvements in surgical outcomes [1]. The precise segmentation of pelvic fracture in pre-operative CT is a fundamental process that underpins many important tasks in trauma diagnosis and image-guided surgery, including fracture typing, reduction planning, screw fixation planning, registration, and surgical navigation [2, 3].

From a technical perspective, the task of segmenting fractured pelvic fragments from CT images presents a myriad of challenges, stemming largely from the various shapes and positions of the bone fragments, and the intricate fracture surfaces arising from bone collisions. Despite the proven effectiveness of deep learning in various applications, fracture segmentation, especially in the pelvic region, remains a notable exception. This is largely due to the scarcity of fractured pelvis image datasets.

The CT segmentation task of the PENGWIN challenge is designed to advance the development of automated fracture segmentation methods for pelvic CT scans, with a focus on enhancing their accuracy and efficiency. We have collected CT scans from 150 patients who were to undergo pelvic reduction surgery, sourced from multiple institutions and various types of scanners. The dataset encompasses a diverse range of patient cohorts and fracture types. Ground-truth segmentations for the sacrum and hipbone fragments have been annotated semi-automatically and validated by experts. The performance in this task will be evaluated primarily using the Dice Similarity Coefficient, focusing on the accuracy of segmentation in both the pelvic anatomy and the specific bone fragments.

Keywords

List the primary keywords that characterize the task.

Computed tomography, Image Segmentation, Pelvic fracture

ORGANIZATION

Organizers

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

Yudi Sang, Rossum Robot
Gang Zhu, Rossum Robot

Yanzhen Liu, Beihang University
Sutuke Yibulayimu, Beihang University
Yu Wang, Beihang University
Mingxu Liu, Johns Hopkins University
Ping-Cheng Ku, Johns Hopkins University
Benjamin D. Killeen, Johns Hopkins University
Mehran Armand, Johns Hopkins University
Mathias Unberath, Johns Hopkins University
Xinbao Wu, Beijing Jishuitan Hospital
Chunpeng Zhao, Beijing Jishuitan Hospital
Dan Ruan, University of California, Los Angeles
S. Kevin Zhou, University of Science and Technology of China

b) Provide information on the primary contact person.

Yudi Sang, PhD
Rossum Robot
sangyudi@rossumrobot.cn

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

One-time event with fixed conference submission deadline

Challenge venue and platform

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

N/A

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).

A challenge website is planned to be launched upon acceptance of the challege proposal.

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.

Any organizations/companies affiliated with members of the organizing committee are not excluded from participation in the challenge, but must ensure that their submissions are completely independent of the members of the organizing committee.

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

The top 3 team will receive cash awards ($1000, $500, and $200, funded by Rossum Robot) and certificates. These teams will also be invited to participate in preparation of the challenge manuscript to become byline authors.

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 shown publicly on the challenge leaderboard, which will be updated daily during the open validation round and announced on the final round phases of the closed test round of the task.

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 challenge organizers will publish at least one challenge journal paper and potentially more. Up to two team members in each of the top performing teams qualify as authors for joint publication. Participants are allowed to publish their results separately after the challenge. The publication embargo period will be 12 months after the announcement of the challenge 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.

Algorithm container submission (type 2) on Grand Challenge.

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 can have multiple submissions on the validation data during the validation phase. For the testing phase, only the last run of the submitted Docker container is officially counted to compute challenge 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

Challenge websites open: 04/01/2024
Training data release: 04/05/2024
Validation data release and results submission: 05/15/2024
Final Submission deadline: 07/31/2024
Result release: 08/31/2024
Presentation at conference: 10/2024

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).

The retrospective data collection, share, and usage strictly conformed to ethical standards and was approved by the ethical committee of Beijing Jishuitan Hospital (Approval number 202009-04).

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.

The metric calculation code will be released along with the validation data.

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

The top 3 performing methods are required to share their code on a public repository. Other teams are also encouraged to share their code. Links to the availble source code will be referenced on the challenge website.

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 is sponsored by Rossum Robot.
Only the organizers (Rossum Robot, IMR Lab of Beihang University, and BIGSS Lab and ARCADE Lab of Johns Hopkins University) will have access to the test case labels during the challenge.

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, Diagnosis, Research

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.

Pelvic trauma patients who have performed CT scanning.

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

Subject of pelvic trauma who planned to undergo pelvic reduction surgery after pre-operative CT scanning.

Imaging modality(ies)

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

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).

Along with the images, ground-truth segmentation labels of the sacrum, left hipbone, right hipbone, and bone fragments will be provided.

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

No further patient information will be provided.

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.

Pelvic region shown in CT.

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.

Bone fragments of fractured pelvis.

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)).

Find pelvic fracture segmentation algorithm with high accuracy, efficiency, and robustness for CT 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).

CT scans were acquired from a diverse array of machines, including a Toshiba Aquilion scanner, an United Imaging uCT 550 scanner, an United Imaging uCT 780 scanner, and a Philips Brilliance scanner.

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

Clinically-used routine CT protocols have been used in each CT scanner. During the acquisition process, all patients were positioned in supine pose. The average voxel spacing is 0.82 × 0.82 × 0.94 mm^3. The typical image dimensions average at 480 × 397 × 310.

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 CT data were collected from the Department of Orthopaedics and Traumatology, Beijing Jishuitan Hospital (75%) and multiple smaller institutions (25%).

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).

CT acquisition was performed by professional radiologists.

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 scan of a human pelvis, and both have an expert-validated annotation.

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

The 150 cases are split into 100 for training, 20 for validation, and 30 for testing.

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

The test cases are sufficiently broad to cover each CT scanner type and each fracture type. Considering the large image size for the CT scans, the number of test cases was set to 30 to limit the total runtime and computational cost for performing the evaluation.

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.

To ensure a balanced representation in the test set, we employed stratified random sampling based on the five primary fracture types so that the training and test sets have the same fracture type distribution.

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.

The dataset is processed by two experienced annotators and a senior expert. The data annotation was structured into a four-step workflow:
(1) Initial automatic segmentation: We employ a pre-trained segmentation network based on the nnUnet framework to produce preliminary segmentations for the sacrum and hipbones. This network was trained on roughly 1,000 pelvic CT images, with the majority of them not presenting any fractures.
(2) Manual refinement of anatomical labels: The initial anatomical labels undergo a meticulous refinement process by annotators using the 3D Slicer platform.
(3) Identification of fractured fragments: Leveraging the refined anatomical labels, the annotators identify and

label fractured bone fragments. This operation is also carried out on the 3D Slicer platform.
(4) Expert validation: As a final checkpoint, a senior expert rigorously reviews, modifies, and authenticates the annotated fracture labels, ensuring their precision and consistency.

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 annotators are asked to annotate the three anatomical labels (sacrum, left hipbone, and right hipbone) and the bone fragment within each bone, using the 3D Slicer platform. When a bone is considered intact without presenting any fracture, the only fragment is its anatomical mask itself. Fragments with volume less than 500 mm3 are omitted.

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 dataset is processed by two experienced annotators and a senior expert. All of them has more than five year experience on pelvis-related study or surgery.

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.

The CT scans and the corresponding labels are manually cropped to the pelvis region when the original field of view is too large (greater than 0.5 meter on z-axis).

Sources of error

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

The source of error may come from annotations of the half-cracked and connected patterns in some (about 10%) fracture cases, where determining the fracture line involves higher inter-observer variability.

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

Another source of error comes from the comminuted fracture regions with small bone shards, where the annotators have to determine whether it belongs to a fragment label.

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)

1. IoU on the anatomical labels
2. IoU on the fracture labels
3. HD95 on the anatomical labels
4. HD95 on the fracture labels
5 ASSD on the anatomical labels
6. ASSD on the fracture labels

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

Intersection over Union (IoU) evaluats the overlap between predicted and ground truth segmentations relative to their union, but is particularly their sensitivity to the segmentation of small structures. To address this, we have incorporated distance-based metrics into our evaluation scheme. The 95th percentile Hausdorff Distance (HD95) metric is particularly valuable for its focus on the largest discrepancies in segmentation without undue influence from outliers, thereby offering a stringent measure of segmentation accuracy in terms of boundary delineation. The Average Symmetric Surface Distance (ASSD) metric, which aligns with the Average Hausdorff Distance (AHD), provides an average measure of the distance between the surfaces of predicted and ground truth segmentations. This inclusion offers a balanced evaluation of segmentation performance across various structure sizes, ensuring a comprehensive assessment of model accuracy and robustness. The adoption of these metrics aligns with the recommendations from the metrics reloader website and has been validated in the study by Faghani et al.

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.

In our evaluation framework, the performance in fragment identification/classification is implicitly evaluated through the results of anatomical segmentation. This approach ensures that the final ranking effectively balances and reflects the overall proficiency in three key areas: anatomical segmentation, fracture indentification, and fracture segmentation. By averaging the results on a case-by-case basis, we ensure that each case, regardless of its inherent difficulty, contributes equally to the determination of the final rankings.

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

For any missing results on the test cases, the assigned metrics will default to an IoU of 0, a HD95 equal to the diameter of the instance's circumscribing circle, and an ASSD equal to the radius of the same circumscribing circle.

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

(1) Metric calculation: On an individual case level, compute the metrics for segmentation. For the fragment instance segmentation, match each ground truth instance with the predicted instance that has the highest IoU.
(2) Independent Ranking: Rank the average results on each metirc separately for each submission.
(3) Final Ranking: Determine the final rank for each submission by averaging its ranks on different metrics.
(4) Tie-Breaking: Use the total execution time of the submissions to break any ties in the final ranking.

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.

No statistical analyses methods have been used in the scope of this challenge.

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

N/A

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.

A baseline algorithm based on nnUNet will be provided for benchmarking. We will also calculate and report the inter-algorithm variability in the challenge publication.

TASK 2: Pelvic Fracture Segmentation in Simulated X-Ray

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 X-ray segmentation task is focused on enhancing intra-operative guidance for pelvic fracture fixation surgeries. Pelvic fracture fixation, a sophisticated surgical procedure, often relies on fluoroscopic imaging to strategically place Kirschner wires and orthopedic screws within specific bony corridors, adhering to a detailed pre-operative plan [4]. This operation, however, faces numerous challenges: accurately identifying pelvic anatomies, clearly visualizing fracture lines, and precisely aligning fracture fragments are all critical yet complex tasks. These challenges highlight the urgent need for improved fluoroscopic image segmentation techniques. Enhanced segmentation capabilities could significantly aid in aligning pre-operative CT data with intra-operative X-rays, enabling more effective correspondence between fracture fragments across these imaging modalities. This alignment is crucial for successful registration, which in turn facilitates the transfer of surgical plans from the pre-operative phase to the actual surgery. Improved registration and segmentation accuracy are key to enhancing the precision and effectiveness of surgical interventions.

From a technical perspective, manually annotating fluoroscopic images is essential but overly time-intensive and impractical during surgery. While deep learning has shown efficacy in segmenting bones in X-ray images, the specific task of segmenting pelvic fractures in X-ray fluoroscopic images poses significant challenges. These primarily stem from a scarcity of annotated intra-operative fluoroscopic data. To overcome this hurdle, the PENGWIN challenge incorporates SyntheX [6], a simulation-based framework designed to advance generalizable medical AI in X-ray image analysis. This challenge aims to expedite the development of effective image segmentation models specifically tailored for pelvic fracture surgeries with 2D X-ray images.

In this task, participants will have access to a dataset comprising 60,000 high-quality, realistic X-ray images. These images are generated from 120 training CT volumes using the DeepDRR framework [5], each yielding 500 X-rays. This dataset is diverse, encompassing a range of virtual C-arm X-ray camera angles and potential surgical tool interferences. The provided data set includes ground-truth 2D segmentations of the sacrum, hipbones, and bone fragments that are consistency with the corresponding 3D CT annotations. Additionally, for evaluation purposes, we have created 15,000 X-ray images derived from the 30 testing CT volumes. We anticipate that methods developed in response to this challenge to be capable of instantaneously segmenting X-ray images without using information from pre-operative CT scans.

Keywords

List the primary keywords that characterize the task.

X-ray, Image Segmentation, Pelvic fracture

ORGANIZATION

Organizers

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

Yudi Sang, Rossum Robot
Gang Zhu, Rossum Robot
Mingxu Liu, Johns Hopkins University
Ping-Cheng Ku, Johns Hopkins University
Benjamin D. Killeen, Johns Hopkins University
Mehran Armand, Johns Hopkins University
Mathias Unberath, Johns Hopkins University
Yanzhen Liu, Beihang University
Sutuke Yibulayimu, Beihang University
Yu Wang, Beihang University
Xinbao Wu, Beijing Jishuitan Hospital
Chunpeng Zhao, Beijing Jishuitan Hospital
Dan Ruan, University of California, Los Angeles
S. Kevin Zhou, University of Science and Technology of China

b) Provide information on the primary contact person.

Yudi Sang, PhD
Rossum Robot
sangyudi@rossumrobot.cn

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

One-time event with fixed conference submission deadline

Challenge venue and platform

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

N/A

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).

A challenge website is planned to be launched upon acceptance of the challege proposal.

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.

Any organizations/companies affiliated with members of the organizing committee are not excluded from participation in the challenge, but must ensure that their submissions are completely independent of the members of the organizing committee.

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

The top 3 team will receive cash awards ($1000, $500, and $200, funded by Rossum Robot) and certificates. These teams will also be invited to participate in preparation of the challenge manuscript to become byline authors.

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 shown publicly on the challenge leaderboard, which will be updated daily during the open validation round and announced on the final round phases of the closed test round of the task.

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 challenge organizers will publish at least one challenge journal paper and potentially more. Up to two team members in each of the top performing teams qualify as authors for joint publication. Participants are allowed to publish their results separately after the challenge. The publication embargo period will be 12 months after the announcement of the challenge 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.

Algorithm container submission (type 2) on Grand Challenge.

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 can have multiple submissions on the validation data during the validation phase. For the testing phase, only the last run of the submitted Docker container is officially counted to compute challenge 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

Challenge websites open: 04/01/2024
Training data release: 04/05/2024
Validation data release and results submission: 05/15/2024
Final Submission deadline: 07/31/2024
Result release: 08/31/2024
Presentation at conference: 10/2024

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).

Same to Task 1.

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.

The metric calculation code will be released along with the validation data. The code for a basic data loaders compatible with the PyTorch framework will also be released.

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

The top 3 performing methods are required to share their code on a public repository. Other teams are also encouraged to share their code. Links to the availble source code will be referenced on the challenge website.

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 is sponsored by Rossum Robot.
Only the organizers (Rossum Robot, IMR Lab of Beihang University, and BIGSS Lab and ARCADE Lab of Johns Hopkins University) will have access to the test case labels during the challenge.

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 assistance, Diagnosis, Research

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.

Pelvic trauma patients who have performed X-ray or who planned to undergo reduction surgery.

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

Subject of pelvic trauma who planned to undergo pelvic reduction surgery after pre-operative CT scanning.

Imaging modality(ies)

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

Simulated X-ray.

Context information

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

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

Along with the images, ground-truth 2D segmentation labels of the sacrum, left hipbone, right hipbone, and bone fragments will be provided.

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

No further patient information will be provided.

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.

Pelvic region shown in X-ray.

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.

Bone fragments of fractured pelvis.

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)).

Find pelvic fracture segmentation algorithm with high accuracy, efficiency, and robustness for X-ray 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).

The CT source is described in task 1.

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

The PENGWIN X-ray dataset consists of digitally reconstructed radiographs (DRRs) with pelvic anatomy, fragments, and surgical tools. Tool poses are simulated in random poses to ensure robustness to surgical hardwares. Utilizing the DeepDRR [5, 6] method, images are rendered based on the original 3D pose of the objects and anatomies, as well as camera intrinsic and extrinsic parameters. The virtual C-arm parameters are randomized over the training set, and each view is sampled from a uniform distribution on the solid angle up to 60 degrees from vertical (assuming a supine patient position). The image dimension is 448 x 448.

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 CT data were collected from the Department of Orthopaedics and Traumatology, Beijing Jishuitan Hospital (75%) and multiple smaller institutions (25%). The X-ray simulation was performed by the BIGSS Lab and the ARCADE Lab, Johns Hopkins University.

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).

The X-ray images was simulated automatically using the DeepDRR method [5, 6].

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 simulated X-ray image of a human pelvis and the corresponding anatomical and fracture annotations.

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

The 150 patients are split into 100 for training, 20 for validation, and 30 for testing. On each patient data, 500 X-ray images are generated, resulting in 50,000, 10,000, and 15,000 for training, validation and testing, repectively.

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

The split or proportion for this task has been set to mirror that of Task 1, ensuring consistency in the evaluation criteria across both tasks.

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.

Same to Task 1.

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.

The 3D segmentation labels described in Task 1 are projected onto the 2D simulated X-rays using the corresponding virtual C-arm parameters .

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.

No annotator is further involved.

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 3D segmentation labels described in Task 1 are projected onto the 2D simulated X-rays using the corresponding virtual C-arm parameters .

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.

A crucial component of the SyntheX pipeline is strong domain randomization (DR) to ensure sim-to-real performance [6]. This has been integrated into the data generation process. No further pre-processing on the simulated X-rays has been performed.

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.

Same as task 1.

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

Same as task 1.

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)

1. IoU on the anatomical labels
2. IoU on the fracture labels
3. HD95 on the anatomical labels
4. HD95 on the fracture labels
5 ASSD on the anatomical labels
6. ASSD on the fracture labels
7. Precision on fragment indentification
8. Recall on fragment indentification

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

IoU, HD95, and ASSD are used for the same reason in task 1. Due to the overlapping bone fragments on 2D projections, task 2 has a multi-label nature. Precision and recall are used to assess the fragment identification result, apart from the segmentation performance.

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.

In this multi-label segmentation tasks, fragment identification performance has to be assessed directly and independently using precision and recall. The approach ensures that the final ranking effectively balances and reflects the overall proficiency in three key areas: anatomical segmentation, fracture indentification, and fracture segmentation. By averaging the results on a case-by-case basis, we ensure that each case, regardless of its inherent difficulty, contributes equally to the determination of the final rankings.

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

For any missing results on the test cases, the assigned metrics will default to an IoU of 0, a HD95 equal to the diameter of the instance's circumscribing circle, and an ASSD equal to the radius of the same circumscribing circle.

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

(1) Metric calculation: On an individual case level, compute the metrics for segmentation. For the fragment instance segmentation, match each ground truth instance with the predicted instance that has the highest IoU. Compute the precision and recall of bone fragment detection as defined by > 50% IoU segmentation results.
(2) Independent Ranking: Rank the average results on each metirc separately for each submission.
(3) Final Ranking: Determine the final rank for each submission by averaging its ranks on different metrics.
(4) Tie-Breaking: Use the total execution time of the submissions to break any ties in the final ranking.

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.

No statistical analyses methods have been used in the scope of this challenge.

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

N/A

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.

N/A

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] Ge, Yufeng, et al. "Robot-assisted autonomous reduction of a displaced pelvic fracture: a case report and brief literature review." Journal of Clinical Medicine 11.6 (2022): 1598.
[2] Moolenaar, Jet Zoë, Nazli Tümer, and Sara Checa. "Computer-assisted preoperative planning of bone fracture fixation surgery: A state-of-the-art review." Frontiers in Bioengineering and Biotechnology 10 (2022): 1037048.
[3] Liu, Yanzhen, et al. "Pelvic Fracture Segmentation Using a Multi-scale Distance-Weighted Neural Network." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2023.
[4] Simonian, P.T., Routt Jr, M.L.C., Harrington, R.M., Tencer, A.F. (1994). "Internal fixation of the unstable anterior pelvic ring: a biomechanical comparison of standard plating techniques and the retrograde medullary superior pubic ramus screw." Journal of Orthopaedic Trauma, 8(6), 476.
[5]. Unberath, M., Zaech, J. N., Lee, S. C., Bier, B., Fotouhi, J., Armand, M., & Navab, N. (2018). DeepDRR - a catalyst for machine learning in fluoroscopy-guided procedures. In Medical Image Computing and Computer Assisted Intervention - MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11 (pp. 98-106). Springer International Publishing.
[6]. Gao, C., Killeen, B. D., Hu, Y., Grupp, R. B., Taylor, R. H., Armand, M., & Unberath, M. (2023). Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. Nature Machine Intelligence, 5(3), 294-308.
[7] Jin, Liang, et al. "Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet." EBioMedicine 62 (2020).
[8] Sekuboyina, Anjany, et al. "VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images." Medical image analysis 73 (2021): 102166.
[9] Faghani, Shahriar, et al. "Mitigating bias in radiology machine learning: 3. Performance metrics." Radiology: Artificial Intelligence 4.5 (2022): e220061.

Further comments

Further comments from the organizers.

The organizing team has a distinguished background, with key members being leaders in their respective fields. Prof. Mehran Armand and Prof. Yu Wang are renowned for their contributions to medical robotics research. Prof. Mathias Unberath, Prof. Dan Ruan, and Prof. S. Kevin Zhou bring a wealth of expertise in the realms of deep learning and medical imaging. Dr. Xinbao Wu and Dr. Chunpeng Zhao are leading experts in orthopedic surgery. Beijing Jishuitan Hospital is recognized as the leading institution in orthopedics and traumatology in China. This diverse expertise ensures a comprehensive and informed approach to the challenge.