Diminished Reality for Emerging Applications in Medicine through Inpainting: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

Diminished Reality for Emerging Applications in Medicine through Inpainting

Challenge acronym

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

DREAMING

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.

While Augmented Reality (AR) is extensively studied in medicine, it represents just one possibility for modifying the real environment. Other forms of Mediated Reality (MR) remain largely unexplored in the medical domain. Diminished Reality (DR) is such a modality. DR refers to the removal of real objects from the environment by virtually replacing them with their background [1]. Combined with AR, powerful MR environments can be created. Although of interest within the broader computer vision and graphics community, DR is not yet widely adopted in medicine [2]. However, DR holds huge potential in medical applications. For example, where constraints on space and intra-operative visibility exist, and the surgeons' view of the patient is further obstructed by disruptive medical instruments or personnel [3], DR methods can provide the surgeon with an unobstructed view of the operation site. Recently, advancements in deep learning have paved the way for real-time DR applications, offering impressive imaging quality without the need for prior knowledge about the current scene [4]. Specifically, deep inpainting methods stand out as the most promising direction for DR [5,6,7].
The DREAM challenge focuses on implementing inpainting-based DR methods in oral and maxillofacial surgery. Algorithms shall fill regions of interest concealed by disruptive objects with a plausible background, such as the patient's face and its surroundings. The facial region is particularly interesting for medical DR, due to its complex anatomy and variety through age, gender and ethnicity. Therefore, we will provide a dataset consisting of synthetic, but photorealistic, surgery scenes focusing on patient faces, with obstructions from medical instruments and hands holding them. These scenes are generated by rendering highly realistic humans together with 3D-scanned medical instruments in a simulated operating room (OR) setting.
This challenge represents an initial frontier in the realm of medical DR, offering a simplified setting to pave the way for MR in medicine. In the future, the potential for more sophisticated applications is expected to unfold.

References:

[1] Mori, S., Ikeda, S., & Saito, H. (2017). A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects. IPSJ Transactions on Computer Vision and Applications, 9(1), 1-14.
[2] Ienaga, N., Bork, F., Meerits, S., Mori, S., Fallavollita, P., Navab, N., & Saito, H. (2016, September). First deployment of diminished reality for anatomy education. In ISMAR-Adjunct (pp. 294-296). IEEE.
[3] Egger, J., & Chen, X. (Eds.). (2021). Computer-Aided Oral and Maxillofacial Surgery: Developments, Applications, and Future Perspectives. Academic Press.
[4] Gsaxner, C., Mori, S., Schmalstieg, D., Egger, J., Paar, G., Bailer, W. & Kalkofen, D. (2023). DeepDR: Deep Structure-Aware RGB-D Inpainting for Diminished Reality. arXiv preprint arXiv: 2312.00532.
[5] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In CVPR (pp. 2536-2544).
[5] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2019). Free-form image inpainting with gated convolution. In ICCV (pp. 4471-4480).
[7] Kim, D., Woo, S., Lee, J. Y., & Kweon, I. S. (2019). Deep video inpainting. In CVPR (pp. 5792-5801).

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Diminished Reality, Face, Surgery, Computer Vision, Deep Learning, Inpainting, Medical Instruments

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.

/

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.

N/A

TASK 1: DREAMING

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.

While Augmented Reality (AR) is extensively studied in medicine, it represents just one possibility for modifying the real environment. Other forms of Mediated Reality (MR) remain largely unexplored in the medical domain. Diminished Reality (DR) is such a modality. DR refers to the removal of real objects from the environment by virtually replacing them with their background [1]. Combined with AR, powerful MR environments can be created. Although of interest within the broader computer vision and graphics community, DR is not yet widely adopted in medicine [2]. However, DR holds huge potential in medical applications. For example, where constraints on space and intra-operative visibility exist, and the surgeons' view of the patient is further obstructed by disruptive medical instruments or personnel [3], DR methods can provide the surgeon with an unobstructed view of the operation site. Recently, advancements in deep learning have paved the way for real-time DR applications, offering impressive imaging quality without the need for prior knowledge about the current scene [4]. Specifically, deep inpainting methods stand out as the most promising direction for DR [5,6,7].
The DREAM challenge focuses on implementing inpainting-based DR methods in oral and maxillofacial surgery. Algorithms shall fill regions of interest concealed by disruptive objects with a plausible background, such as the patient's face and its surroundings. The facial region is particularly interesting for medical DR, due to its complex anatomy and variety through age, gender and ethnicity. Therefore, we will provide a dataset consisting of synthetic, but photorealistic, surgery scenes focusing on patient faces, with obstructions from medical instruments and hands holding them. These scenes are generated by rendering highly realistic humans together with 3D-scanned medical instruments in a simulated operating room (OR) setting.
This challenge represents an initial frontier in the realm of medical DR, offering a simplified setting to pave the way for MR in medicine. In the future, the potential for more sophisticated applications is expected to unfold.

References:
[1] Mori, S., Ikeda, S., & Saito, H. (2017). A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects. IPSJ Transactions on Computer Vision and Applications, 9(1), 1-14.
[2] Ienaga, N., Bork, F., Meerits, S., Mori, S., Fallavollita, P., Navab, N., & Saito, H. (2016, September). First deployment of diminished reality for anatomy education. In ISMAR-Adjunct (pp. 294-296). IEEE.
[3] Egger, J., & Chen, X. (Eds.). (2021). Computer-Aided Oral and Maxillofacial Surgery: Developments, Applications, and Future Perspectives. Academic Press.
[4] Gsaxner, C., Mori, S., Schmalstieg, D., Egger, J., Paar, G., Bailer, W. & Kalkofen, D. (2023). DeepDR: Deep Structure-Aware RGB-D Inpainting for Diminished Reality. arXiv preprint arXiv: 2312.00532.
[5] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In CVPR (pp. 2536-2544).
[5] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2019). Free-form image inpainting with gated convolution. In ICCV (pp. 4471-4480).
[7] Kim, D., Woo, S., Lee, J. Y., & Kweon, I. S. (2019). Deep video inpainting. In CVPR (pp. 5792-5801).

Keywords

List the primary keywords that characterize the task.

Diminished Reality, Face, Surgery, Computer Vision, Deep Learning, Inpainting, Medical Instruments

ORGANIZATION

Organizers

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

Christina Gsaxner (Graz University of Technology; Medical University of Graz)
Shohei Mori (Graz University of Technology)
Gijs Luijten (Graz University of Technology; University Hospital Essen)
Viet Duc Vu (University Hospital Giessen)
Timo van Meegdenburg (University Hospital Essen)
Gabriele A. Krombach (University Hospital Giessen)
Jens Kleesiek (University Hospital Essen; University Medicine Essen; German Cancer Consortium (DKTK))
Ulrich Eck (Technical University Munich)
Nassir Navab (Technical University Munich)
Yan Guo (Shanghai Jiao Tong University)
Xiaojun Chen (Shanghai Jiao Tong University)
Frank Hölzle (University Hospital RWTH Aachen)
Behrus Puladi (University Hospital RWTH Aachen)
Jan Egger (University Hospital Essen; University Medicine Essen)

b) Provide information on the primary contact person.

gsaxner@tugraz.at (Christina Gsaxner)

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 submission deadline

Challenge venue and platform

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

ISBI 2024

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

The challenge will be hosted on a common platform (grand-challenge.org). Participants will be able to submit their methods via a docker container over the platform.

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

https://dreaming.ikim.nrw/

Participation policies

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

Algorithms should be fully automatic (no user interactions allowed).

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 and public pre-trained networks are allowed.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate but not eligible for awards and not listed in leaderboard.

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

We will award best paper & highest-ranking certificates to the winning teams. We will also look for company sponsorship for monetary awards.

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 valid submissions will be listed in the leaderboard publicly available on the challenge website.
Submitted valid papers (4 pages with introduction, data description, method description, cross-validation results on training data and results on validation data, discussion and conclusion) will be published in the ISBI proceedings. Participating teams can choose to contribute to a joint publication summarizing the challenge results which will be submitted to a journal.

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).

Members that contribute to the development of the algorithms or the writing of papers qualify as authors. Participating teams should submit their papers via CMT for peer-review. Accepted valid papers (4 pages with introduction, data description, method description, cross-validation results on training data and results on validation data, discussion and conclusion) will be included in the ISBI proceedings. Authors should refer to the ISBI author instructions to prepare their papers: https://biomedicalimaging.org/2024/authors-instructions/

Authors may publish their paper submissions as pre-prints. Submissions of substantially similar content to other venues are not allowed during the entire duration of the challenge. Papers that are not accepted can be contributed to other venues.

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.

Docker container submissions on grand-challenge.org. Submission instructions: https://grand-challenge.org/documentation/create-your-own-algorithm/

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.

The Method evaluation will be split in two phases: In a first container testing phase, participants will be allowed to submit their containers up to 5 times. Their algorithms will be tested on two hidden test cases. In the second phase, participants will be allowed to submit their containers up to 2 times. Their algorithms will be tested on ten hidden synthetic test cases and ranked based on the results. Additionally, the algorithms will be tested qualitatively on two real test cases.

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

8th January 2024: Initiation of challenge. First subset of training & validation data available.
22nd January 2024: Second subset of training & validation data available.
29th January 2024: Challenge platform available on grand-challenge.org. Full training & validation data available. Method evaluation opens.
6th April 2024: Paper submission deadline.
20th April 2024: Notifications of paper reviews.
27th April 2024: Camera-ready paper submission deadline; Method evaluation closes.
6th May 2024: Final notifications (poster vs oral).
27th May 2024: Conference - Announcement of winners.
All deadlines are 23:59 Pacific Time

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).

Informed consent will be collected from volunteers prior to any video recordings and ethics approval will be obtained, if required. However, real patient recordings will not be made publicly available.

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.

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 available on GitHub.

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

Participants are strongly encouraged to make their code and trained models publicly available. Furthermore, they are strongly encouraged to describe their method in a challenge paper, to facilitate the reproducibility of the results.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

The challenge organizers have no conflict of interest. We will look for company sponsorship for monetary awards. The organizers have access to the ground truth of the test cases.

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

Decision support, Surgery, Assistance, Research.

Task category(ies)

State the task category(ies)

Examples:

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

Inpainting, Object Removal.

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.

The target cohort are patients undergoing surgery in the head and face region (e.g., oral and maxillofacial or head and neck surgery), in the view of the surgeon, who looks at them through a video see-through display / screen. Disruptive medical instruments or hands that block the operating site should be removed in the surgeons' view of the patient (in vivo).

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

The challenge cohort are highly realistic, synthetic human characters (Unreal Engine MetaHumans) in a simulated OR environment. Realistic, dynamically moving medical instruments (3D scanned from real counterparts), as well as hands holding them, are added to the scene as targets for removal. The scenes are rendered from a perspective similar to the point of view from an operating surgeon, to result in surgical video sequences. Light sources matching OR conditions are added to the scene, to produce realistic lighting and shadows. Each scene is rendered two times: once with (masked) instruments, as input to the algorithm, and once without instruments, as ground truth.
Furthermore, for testing, a small sample of surgical video sequences will be acquired using a mobile video camera to reflect the target cohort. This sample will stay hidden from participants and will be used for a qualitative assessment. This approach allows us to also assess the generalization ability of proposed solutions to real-world data. From our own experience and recent developments in the literature [1,2], we expect that state-of-the-art inpainting approaches can generalize well to out-of-distribution data. In particular, techniques such as contextual attention or transformer architectures ensure that trained models are able to utilize features from the given input data for filling missing regions. The results of the assessment will be presented and discussed during the conference's challenge event and will be part of the joint meta-analysis publication. The generalization ability of photorealistic to real data is an interesting, open research question, which will be addressed in this challenge.
[1] Paulin, G., & Ivasic-Kos, M. (2023). Review and analysis of synthetic dataset generation methods and techniques for application in computer vision. Artificial Intelligence Review, 1-45.
[2] Nikolenko, S. I. (2021). Synthetic data for deep learning (Vol. 174). Springer Nature.

Imaging modality(ies)

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

The imaging techniques applied are conventional videography, as well as 3D scanning of medical instruments [1].
[1] Luijten, G., Gsaxner, C., Li, J., Pepe, A., Ambigapathy, N., Kim, M., ... & Egger, J. (2023). 3D surgical instrument collection for computer vision and extended reality. Scientific Data, 10(1), 796.

Context information

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

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

Not applicable.

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

Patients undergoing surgery in the head and face area (e.g., oral and maxillofacial surgery or head and neck surgery), with different demographics regarding gender, age and ethnicity.

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.

The data will originate from an OR shown in video data, acquired from the point of view of the operating surgeon. Since we focus on surgery in the head and face area, the video data will mostly show the patients face, but also its surroundings, such as the operating bed or attending medical staff. While the challenge cohort contains exclusively synthetic, but photorealistic data, the target cohort will be live videos from the operating room (in vivo).

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.

The algorithm target is to remove medical instruments and hands holding them from the operating site, specifically, to fill regions covering patient faces and its close surroundings with plausible content. Thus, for an input sequence in which a region of interest is masked out, the algorithm should output a filled sequence which looks realistic.

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)).

Accuracy, Consistency / Plausiblity, Runtime, Applicability, User satisfaction.

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).

Synthetic OR scenes are created and rendered in Unreal Engine using MetaHumans to simulate patients and attending staff.
Real medical instruments from facial surgery are scanned using ARTEC Leo and Autoscan Inspec. Real OR scenes will be captured using standard mobile cameras (e.g., GoPro).

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

Camera parameters (intrinsic, extrinsic) used during rendering and filming will be made available. All renderings will be done in HD resolution (1,280 x 720 pixels or similar), at 30 frames per second. Each sequence will contain around 700 -1000 frames, resulting in sequences approximately 30-40 seconds long. The medical instruments are scanned using Artec Leo HD 3D and Autoscan Inspec. The collection is publicly available [1].
[1] Luijten, G., Gsaxner, C., Li, J., Pepe, A., Ambigapathy, N., Kim, M., ... & Egger, J. (2023). 3D surgical instrument collection for computer vision and extended reality. Scientific Data, 10(1), 796.

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.

Renderings are created at the Institute of Computer Graphics and Vision of the Graz University of Technology in Austria and the Institute for AI in Medicine of the University Hospital Essen in Germany.
The medical instruments have been provided by the Department of Oral and Maxillofacial Surgery of the University Hospital RWTH Aachen in Aachen, Germany, and were acquired at the Institute for AI in Medicine (IKIM), University Hospital Essen, Germany. Real scenes will be captured at the University Hospital RWTH Aachen in Aachen, Germany.

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).

Dr. Dr. Behrus Puladi, a specialist in Oral and Maxillofacial Surgery and Medical Informatics at the University Hospital RWTH Aachen, is supervising the data acquisition process to ensure clinical resemblance and relevance.

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 refers to a rendered video sequence of one patient with at least one medical instrument and / or hand to be removed. For our synthetic dataset, we prepare two identical sequences per case, one with (masked) objects of interest as input to the algorithm and one without the objects of interest as ground truth for training and testing.

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

We will provide 100 synthetic cases for training, and 10 for testing. Furthermore, 2 sequences from real scenes will be acquired for qualitative testing. Due to the lack of ground truth, real scenes cannot be considered for metric computation. We do not release an explicit validation data set. Participants can split the training data accordingly during the training phase.

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

We create 110 photorealistic MetaHumans as patients (100 for training and validation, 10 for testing), and 3D scan over 100 medical instruments and hands to simulate holding the instruments. A selection of instruments will be dynamically added and moved over the faces to simulate surgical movements.

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 MetaHumans will be created in a way that covers a variety of patients with different demographics (e.g., gender, age, ethnicity). The training, validation and test cases will be split by scene, meaning that MetaHumans or instruments in the training set will not appear in the validation or test set.

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.

No human annotation is involved in the generation of synthetic data. In the real test sequences, two annotators will delineate the target region for inpainting (i.e., a medical instrument or hand).

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.

-

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.

-

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.

-

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.

-

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.

-

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

-

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Accuracy:
- Peak Signal to Noise Ratio (PSNR)
- Mean Absolute Error (MAE)
Consistency / Plausibility:
- Learned Perceptual Image Patch Similarity (LPIPS) [1]
- Frechet Inception Distance (FID) [2] and Video Frechet Inception Distance (VFID) [3]
Runtime:
- Inference time (in ms)
User satisfaction, applicability:
- Qualitative assessment

PSNR, MAE, and LPIPS will be computed for every image in which an object should be removed and averaged over all these images. FID and VFID will be computed for the entire sequence.

[2] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In CVPR (pp. 586-595).
[2] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30.
[3] Wang, T. C., Liu, M. Y., Zhu, J. Y., Liu, G., Tao, A., Kautz, J., & Catanzaro, B. (2018). Video-to-video synthesis. arXiv preprint arXiv:1808.06601.

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

Contrary to other image-to-image translation applications, object removal or DR is an ill-posed problem, where several visually plausible solutions exist. A plausible, consisten result is not necessarily an accurate result on a pixel basis. This is why, aside from metrics measuring pixel-wise concordance between algorithmic output and ground truth, i.e., PSNR and MAE, we put a special emphasis on feature-level metrics, particularly LPIPS and (V)FID. These metrics have been shown to represent human judgment to a surprisingly accurate extent when it comes to plausibility [1,2]. For computing LPIPS, image patches are passed through a CNN pre-trained on ImageNet, such as AlexNet or VGG. The activations at different layers of this CNN are then compared using l2 distance. Similarly, for (V)FID computation, images are passed through a pre-trained Inception v3 CNN. Compared to LPIPS, which works on a single-image basis, FID is computed on an entire dataset, and therefore, informs about the similarity between the distribution of features in two sets of images or videos. This is done by comparing mean and covariance of activations across real and generated data using Wasserstein-2 distance. Still, since we

have a pixel-accurate ground truth, using MAE as an accuracy metric and PSNR to quantify noise, are valid choices. Unfortunately, there are no better metrics available [3].
Runtime is an important factor for DR, since the goal is that algorithms can run in real-time, ideally on mobile hardware (although this is not a requirement for participating in the challenge). Lastly, since the goal of DR is to improve the perception of the user, results should be qualitatively pleasing.
The most convincing evaluation of inpainting quality is a human evaluation. This is why medical staff working in the clinical routine will informally evaluate the participants' results on the real test set.
However, since a conclusive and fair human evaluation / user study with a large enough number of participants may not be feasible in the short time frame of a challenge, we will rely on quantitative metrics for ranking, for full transparency. We think that our combination of metrics can provide satisfactory information about inpainting quality.
[1] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In CVPR (pp. 586-595).
[2] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30.
[3] Xiang, H., Zou, Q., Nawaz, M. A., Huang, X., Zhang, F., & Yu, H. (2023). Deep learning for image inpainting: A survey. Pattern Recognition, 134, 109046.

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.

Rendering a synthetic scene with and without instruments allows for a precise ground truth on all synthetic cases, which is important for a fair evaluation. Hence, ranking will be based on synthetic data only. The final scores for each metric will be obtained by averaging over all synthetic cases in the test set. Methods will be ranked according to feature-based plausibility / consistency metrics (LPIPS, FID and VFID) first, and pixel-based accuracy metrics (PSNR and MAE) second. A separate ranking will be provided for computational efficiency (taking runtime into account). LPIPS, FID and VFID will be normalized to values between 0 and 1 and averaged using equal weights to compute a plausibility / consistency score. PSNR and MAE will also be weighted equally to compute an accuracy score. Real cases will only be used for a qualitative assessment of the generalization ability of algorithms.

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

Will be excluded.

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

Feature-based metrics are more descriptive of the perceptual plausibility of results, which is more important in diminished reality than pixel-wise accuracy.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,

- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

Missing data handling is not applicable for our challenge. The challenge results will be evaluated statistically, with common measures, like mean, standard deviation, variance, etc.

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

Measures like mean, standard deviation, variance, etc. are common in statistics and they show the distribution and outliers of the challenge results.

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.

Facial surgeons will qualitatively assess the outcomes from a clinical standpoint. Their assessment, together with further analyses, such as common problems/biases of the submitted methods, etc. will be discussed and presented at the challenge event and will be included in a challenge summary publication.

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.

N/A

Further comments

Further comments from the organizers.

N/A