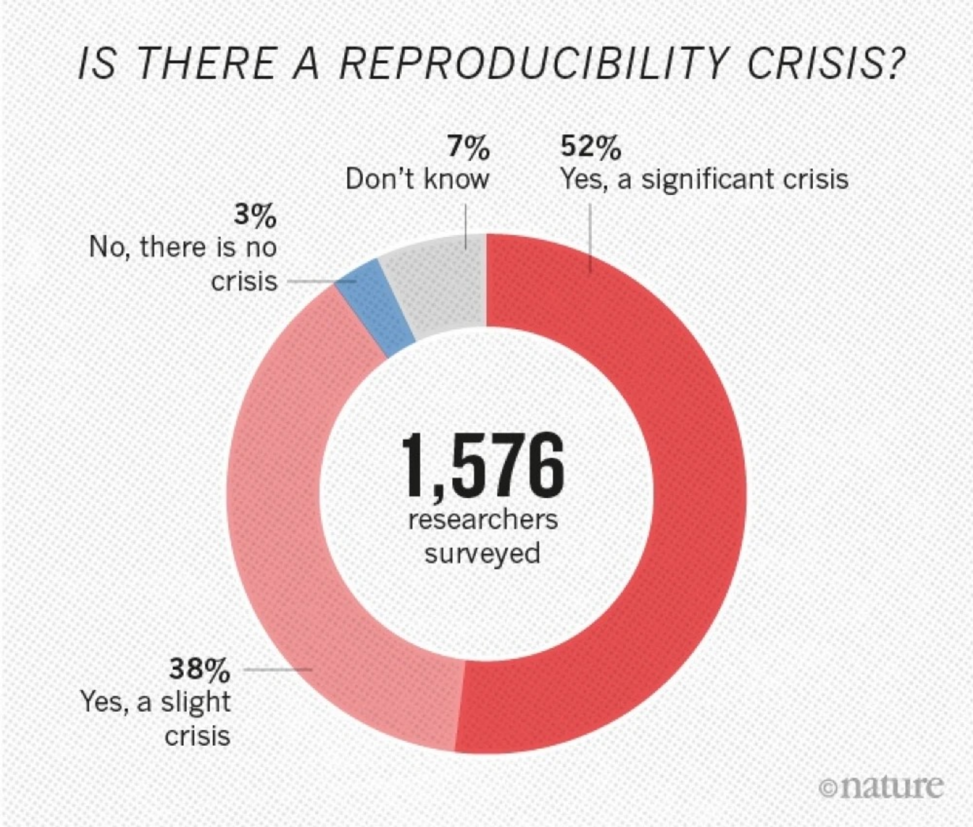
Reproducibility
Tutorial

MICCAI reproducibility checklist

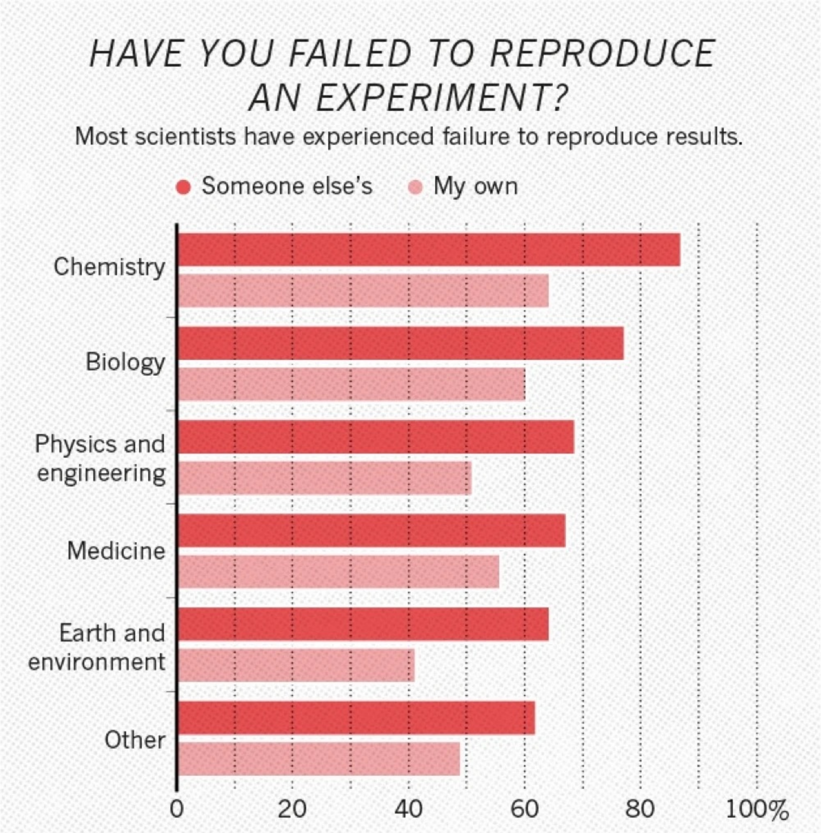
Why it matters & Practical tips
to achieve reproducibility

THE REPRODUCIBILITY CRISIS



Baker, Nature, 2016

THE REPRODUCIBILITY CRISIS



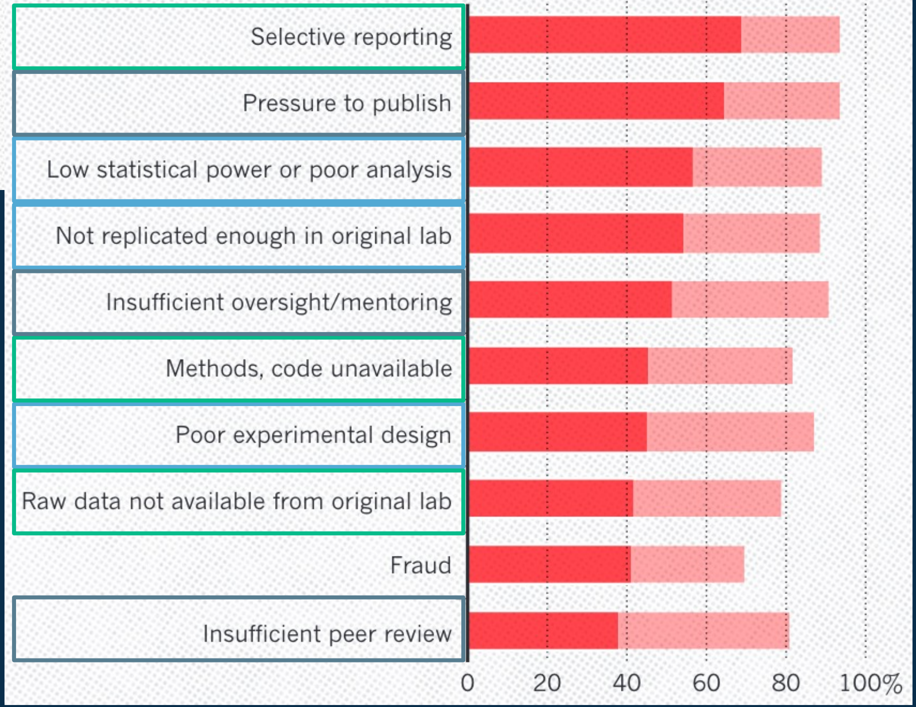
Baker, Nature, 2016

THE REPRODUCIBILITY CRISIS

WHAT FACTORS CONTRIBUTE TO IRREPRODUCIBLE RESEARCH?

Many top-rated factors relate to intense competition and time pressure.

● Always/often contribute ● Sometimes contribute



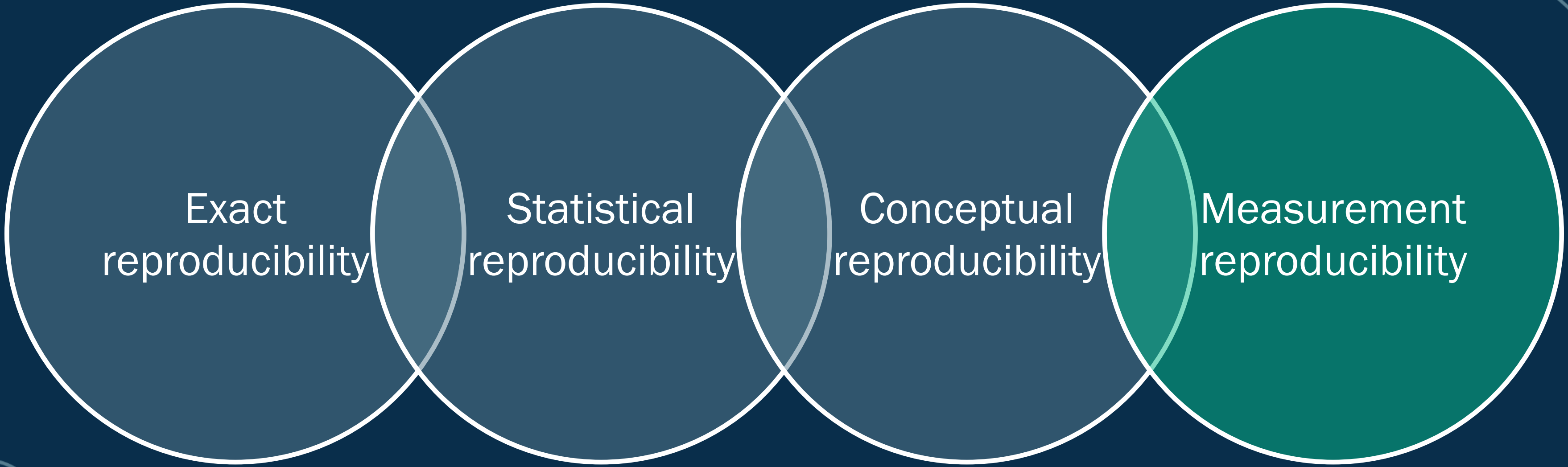
Methodological failure

Insufficient reporting

Failures of the scientific community

Baker, Nature, 2016

DIFFERENT TYPES OF REPRODUCIBILITY



Exact reproducibility

Statistical reproducibility

Conceptual reproducibility

Measurement reproducibility

Colliot et al., MLBD, 2023

DIFFERENT TYPES OF REPRODUCIBILITY

Exact reproducibility

Reproduction of strictly identical results as those of a previously published paper

Example: reproducing classification accuracies using the exact same code, data and random seeds

Colliot et al., MLBD, 2023

DIFFERENT TYPES OF REPRODUCIBILITY

Statistical reproducibility

Reproduction of the results of a study under statistically equivalent conditions. Results should be statistically compatible but not identical.

Example: reproducing a study using another sample of patients drawn from the same population or from a population with the same characteristics

Colliot et al., MLBD, 2023

DIFFERENT TYPES OF REPRODUCIBILITY

Conceptual reproducibility

Reproduction of the results of a study under conceptually equivalent conditions.
This includes generalizability studies.

Example: reproducing a study using a different sample of patients, affected by the same disorder, but with different socio-demographic characteristics and from different hospitals

Colliot et al., MLBD, 2023

DIFFERENT TYPES OF REPRODUCIBILITY

Measurement reproducibility

Variability of a measurement (computed at the patient level) under variations of the input data.

Example: variability of a volumetric measurement (coming for an ML-based segmentation method) when using different scans of the same patient (often called test-retest reproducibility)

Colliot et al., MLBD, 2023

WHAT IS REQUIRED TO ACHIEVE REPRODUCIBILITY

Exact reproducibility

- Access to all components that led to the results, including
  - data
  - code
  - trained models

Colliot et al., MLBD, 2023

WHAT IS REQUIRED TO ACHIEVE REPRODUCIBILITY

Statistical reproducibility

- Detailed description of the data
- Access to code
- Reporting of error margins

Colliot et al., MLBD, 2023

WHAT IS REQUIRED TO ACHIEVE REPRODUCIBILITY

Conceptual reproducibility

- Detailed description of the
  - datasets
  - methods
  - experimental procedure

Colliot et al., MLBD, 2023

THE MICCAI REPRODUCIBILITY CHECKLIST

ORIGIN: MICCAI Hackathon 2020

Two questions investigated:

- What does it need for a MICCAI paper to be reproducible?
- What could MICCAI do to encourage reproducibility?

https://2020.miccai-hackathon.com

Balsiger et al., arXiv, 2021

THE MICCAI REPRODUCIBILITY CHECKLIST

ORIGIN: MICCAI Hackathon 2020

➢ Immediate measures

- Incorporate the reproducibility checklist in the paper submission form
- Introduce a reproducibility chair at MICCAI
- Include a statement of data availability in the conference management toolkit (CMT)
- Promote reproducibility efforts of authors

  (github.com/JunMa11/MICCAI-OpenSourcePapers | github.com/yiqings/MICCAI2022_paper_with_code)
- Communicate best practices on reproducibility and code submission

Balsiger et al., arXiv, 2021

THE MICCAI REPRODUCIBILITY CHECKLIST

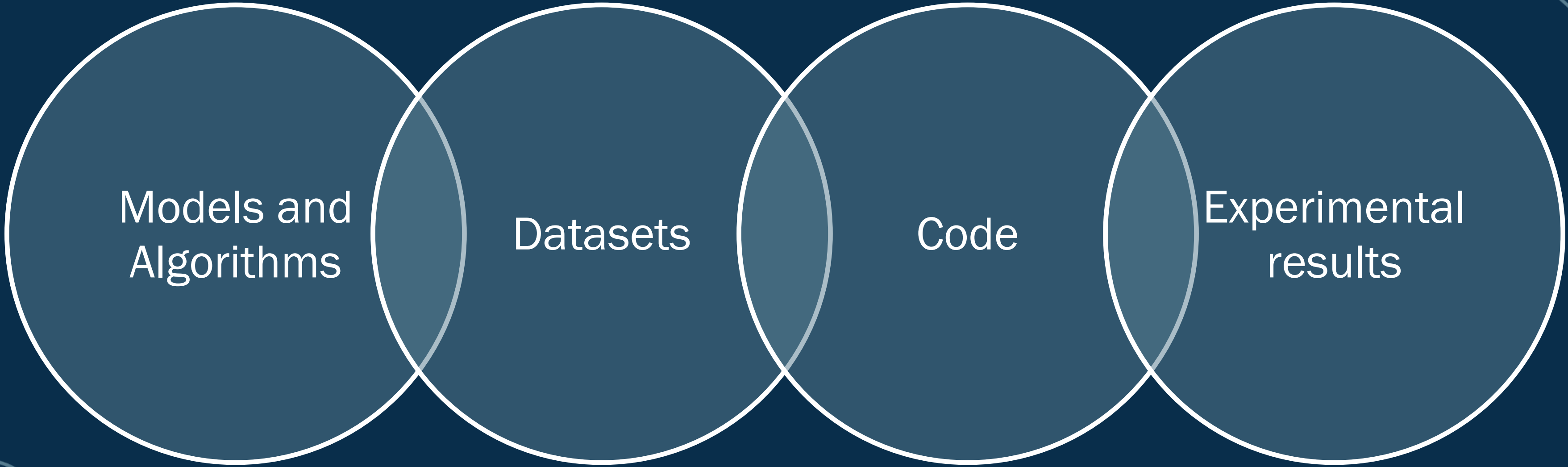
ORIGIN: MICCAI Hackathon 2020

- Long-term measures
  - Introduce a reproducibility award
  - Code submission
  - Best practices for evaluation

    (Maier-Hein et al; (2022). Metrics reloaded: Pitfalls and recommendations for image analysis validation. ArXiv.org 2206.01653)

Balsiger et al., arXiv, 2021

THE MICCAI REPRODUCIBILITY CHECKLIST



Models and Algorithms

Datasets

Code

Experimental results

THE MICCAI REPRODUCIBILITY CHECKLIST

Models and Algorithms

A clear description of the mathematical setting, algorithm, and/or model.

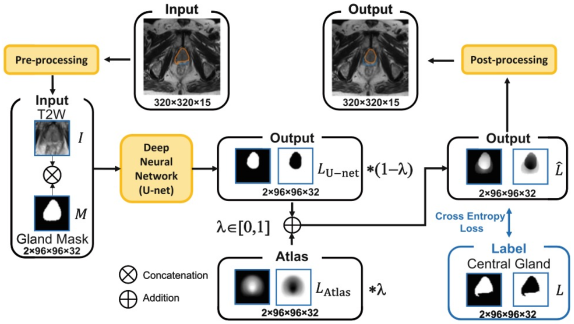
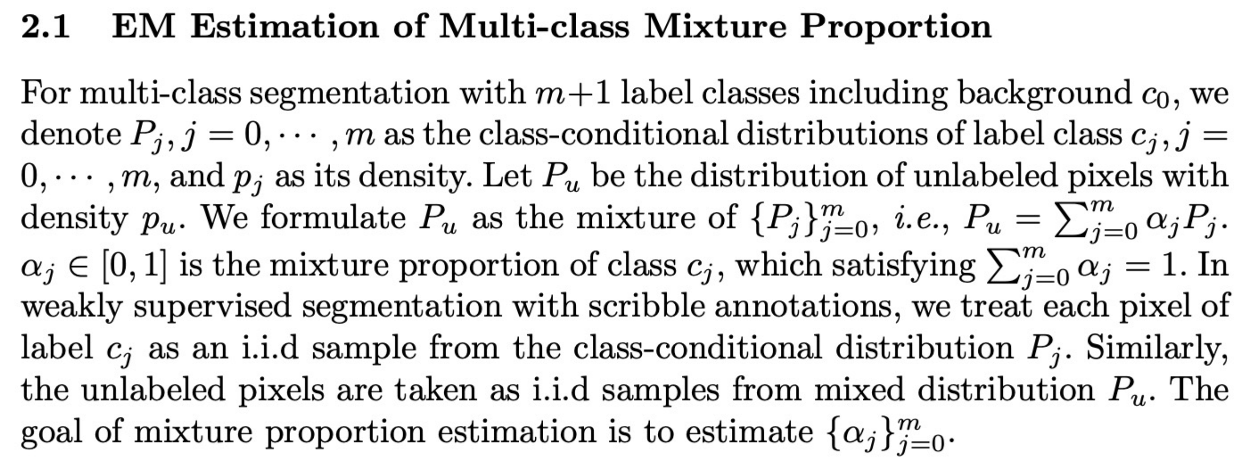


Fig. 1. Atlas-based semantic segmentation framework. Our proposed framework consists of four modules: (1) pre-processing to transform imaging data from the *native space* to the *atlas space*; (2) a deep neural network to segment the CG with the aid of a probabilistic CG shape atlas L_{Atlas}; (3) a hyperparameter λ to dynamically adjust the atlas weight; and (4) post-processing to transform segmentation results from the *atlas space* back to the *native space*.

THE MICCAI REPRODUCIBILITY CHECKLIST

Models and Algorithms

A clear explanation of any assumptions.



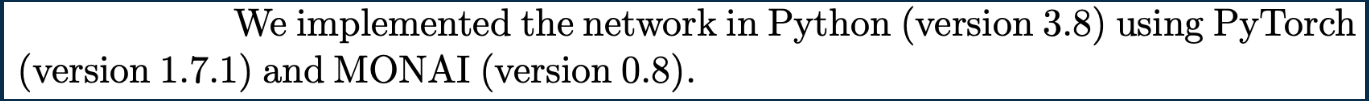
2.1 EM Estimation of Multi-class Mixture Proportion

For multi-class segmentation with $m+1$ label classes including background c_0, we denote $P_j, j = 0, \cdots, m$ as the class-conditional distributions of label class $c_j, j = 0, \cdots, m$, and p_j as its density. Let P_u be the distribution of unlabeled pixels with density p_u. We formulate P_u as the mixture of $\{P_j\}_{j=0}^{m}$, *i.e.*, $P_u = \sum_{j=0}^{m} \alpha_j P_j$. $\alpha_j \in [0, 1]$ is the mixture proportion of class c_j, which satisfying $\sum_{j=0}^{m} \alpha_j = 1$. In weakly supervised segmentation with scribble annotations, we treat each pixel of label c_j as an i.i.d sample from the class-conditional distribution P_j. Similarly, the unlabeled pixels are taken as i.i.d samples from mixed distribution P_u. The goal of mixture proportion estimation is to estimate $\{\alpha_j\}_{j=0}^{m}$.

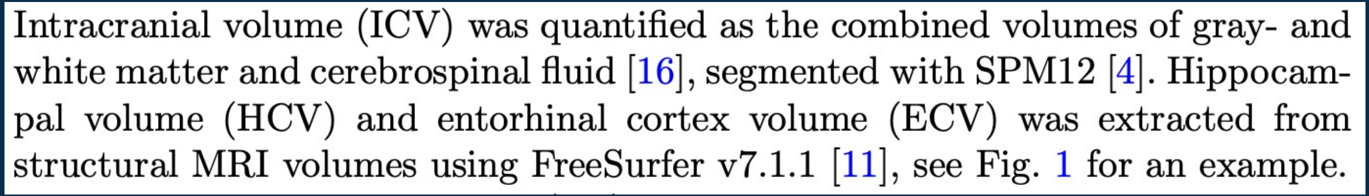
THE MICCAI REPRODUCIBILITY CHECKLIST

Models and Algorithms

A clear declaration of what software framework and version you used.



We implemented the network in Python (version 3.8) using PyTorch (version 1.7.1) and MONAI (version 0.8).



Intracranial volume (ICV) was quantified as the combined volumes of gray- and white matter and cerebrospinal fluid [16], segmented with SPM12 [4]. Hippocampal volume (HCV) and entorhinal cortex volume (ECV) was extracted from structural MRI volumes using FreeSurfer v7.1.1 [11], see Fig. 1 for an example.

THE MICCAI REPRODUCIBILITY CHECKLIST

Datasets

The relevant statistics, such as number of examples.

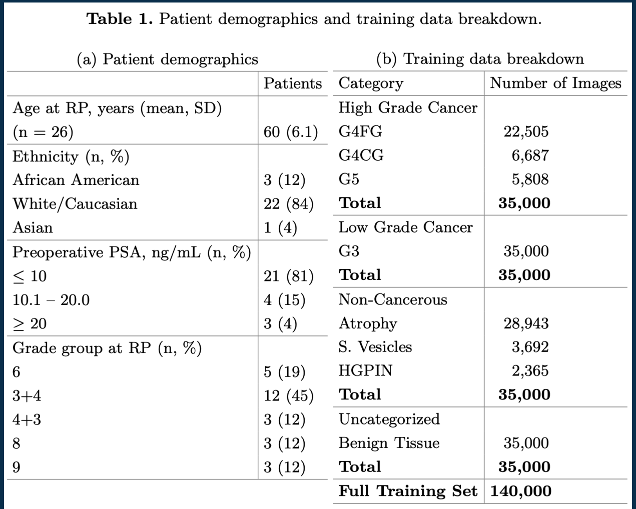


Table 1. Patient demographics and training data breakdown.

(a) Patient demographics

	Patients
Age at RP, years (mean, SD) (n = 26)	60 (6.1)
Ethnicity (n, %)	
African American	3 (12)
White/Caucasian	22 (84)
Asian	1 (4)
Preoperative PSA, ng/mL (n, %)	
≤ 10	21 (81)
10.1 – 20.0	4 (15)
≥ 20	3 (4)
Grade group at RP (n, %)	
6	5 (19)
3+4	12 (45)
4+3	3 (12)
8	3 (12)
9	3 (12)

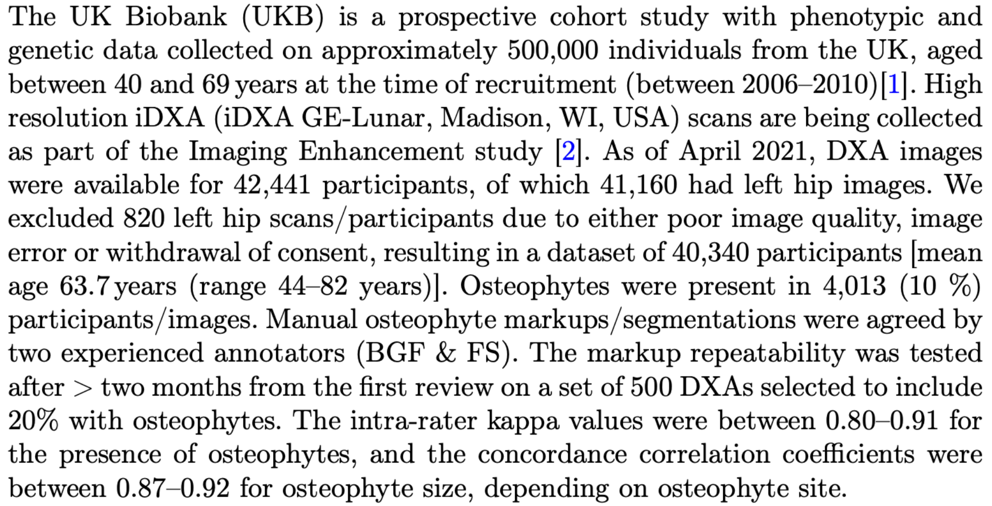
(b) Training data breakdown

Category	Number of Images
High Grade Cancer	
G4FG	22,505
G4CG	6,687
G5	5,808
Total	**35,000**
Low Grade Cancer	
G3	35,000
Total	**35,000**
Non-Cancerous	
Atrophy	28,943
S. Vesicles	3,692
HGPIN	2,365
Total	**35,000**
Uncategorized	
Benign Tissue	35,000
Total	**35,000**
Full Training Set	**140,000**

THE MICCAI REPRODUCIBILITY CHECKLIST

Datasets

Description of the study cohort.

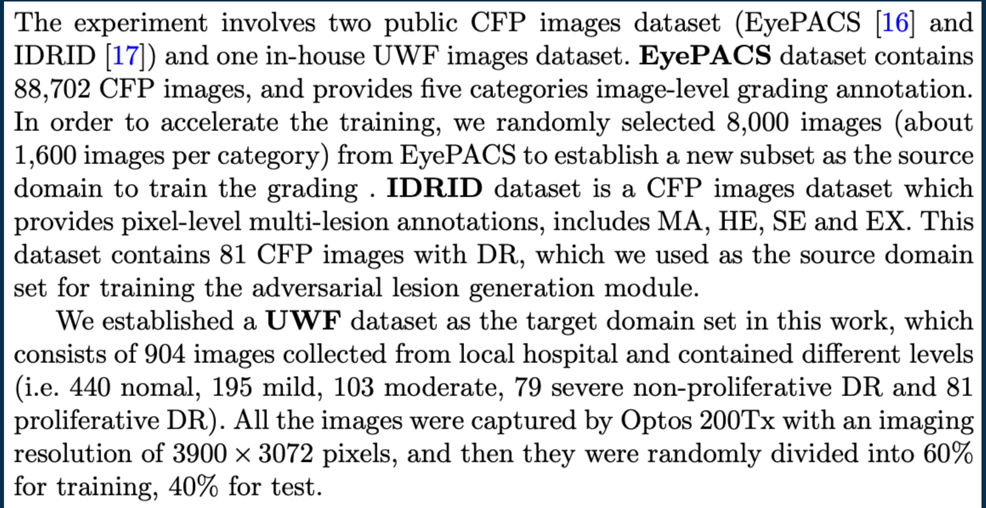


The UK Biobank (UKB) is a prospective cohort study with phenotypic and genetic data collected on approximately 500,000 individuals from the UK, aged between 40 and 69 years at the time of recruitment (between 2006–2010)[1]. High resolution iDXA (iDXA GE-Lunar, Madison, WI, USA) scans are being collected as part of the Imaging Enhancement study [2]. As of April 2021, DXA images were available for 42,441 participants, of which 41,160 had left hip images. We excluded 820 left hip scans/participants due to either poor image quality, image error or withdrawal of consent, resulting in a dataset of 40,340 participants [mean age 63.7 years (range 44–82 years)]. Osteophytes were present in 4,013 (10 %) participants/images. Manual osteophyte markups/segmentations were agreed by two experienced annotators (BGF & FS). The markup repeatability was tested after > two months from the first review on a set of 500 DXAs selected to include 20% with osteophytes. The intra-rater kappa values were between 0.80–0.91 for the presence of osteophytes, and the concordance correlation coefficients were between 0.87–0.92 for osteophyte size, depending on osteophyte site.

THE MICCAI REPRODUCIBILITY CHECKLIST

Datasets

For existing datasets, citations as well as descriptions if they are not publicly available.



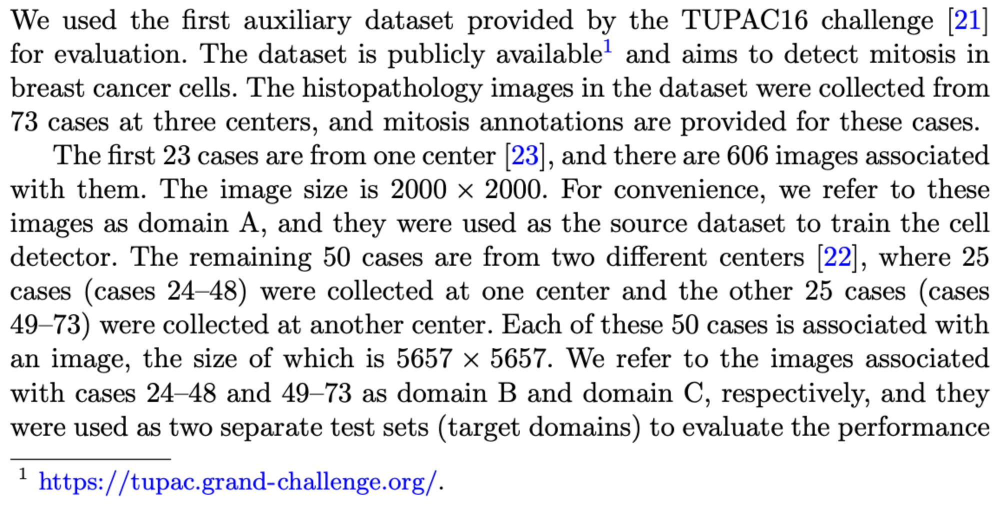
The experiment involves two public CFP images dataset (EyePACS [16] and IDRID [17]) and one in-house UWF images dataset. **EyePACS** dataset contains 88,702 CFP images, and provides five categories image-level grading annotation. In order to accelerate the training, we randomly selected 8,000 images (about 1,600 images per category) from EyePACS to establish a new subset as the source domain to train the grading . **IDRID** dataset is a CFP images dataset which provides pixel-level multi-lesion annotations, includes MA, HE, SE and EX. This dataset contains 81 CFP images with DR, which we used as the source domain set for training the adversarial lesion generation module.

We established a **UWF** dataset as the target domain set in this work, which consists of 904 images collected from local hospital and contained different levels (i.e. 440 nomal, 195 mild, 103 moderate, 79 severe non-proliferative DR and 81 proliferative DR). All the images were captured by Optos 200Tx with an imaging resolution of 3900×3072 pixels, and then they were randomly divided into 60% for training, 40% for test.

THE MICCAI REPRODUCIBILITY CHECKLIST

Datasets

A link to a downloadable version of the dataset (if public).



We used the first auxiliary dataset provided by the TUPAC16 challenge [21] for evaluation. The dataset is publicly available[1] and aims to detect mitosis in breast cancer cells. The histopathology images in the dataset were collected from 73 cases at three centers, and mitosis annotations are provided for these cases.

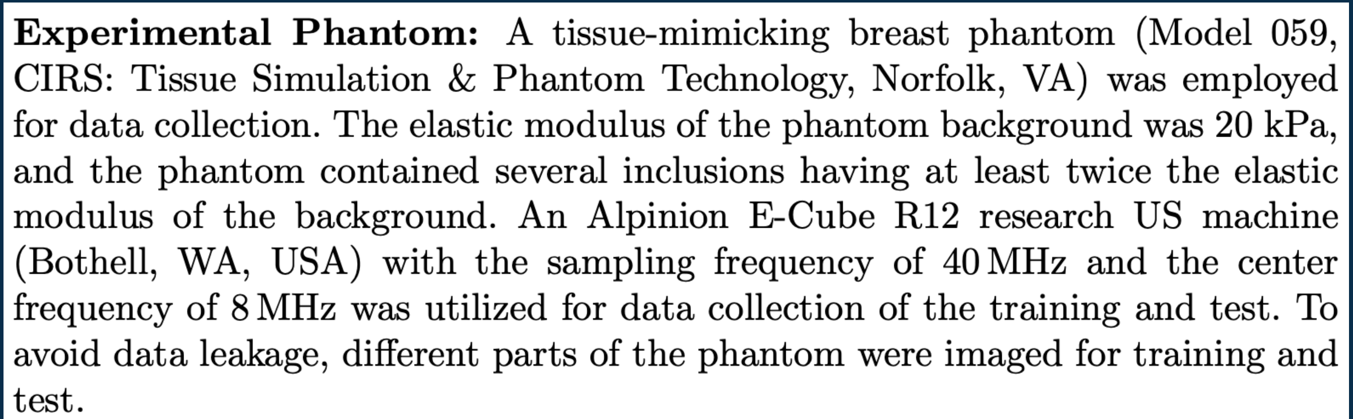
The first 23 cases are from one center [23], and there are 606 images associated with them. The image size is 2000 × 2000. For convenience, we refer to these images as domain A, and they were used as the source dataset to train the cell detector. The remaining 50 cases are from two different centers [22], where 25 cases (cases 24–48) were collected at one center and the other 25 cases (cases 49–73) were collected at another center. Each of these 50 cases is associated with an image, the size of which is 5657 × 5657. We refer to the images associated with cases 24–48 and 49–73 as domain B and domain C, respectively, and they were used as two separate test sets (target domains) to evaluate the performance

[1] https://tupac.grand-challenge.org/.

THE MICCAI REPRODUCIBILITY CHECKLIST

Datasets

For new data collected, a complete description of the data collection process, such as descriptions of the experimental setup, device(s) used, image acquisition parameters, subjects/objects involved, instructions to annotators, and methods for quality control.

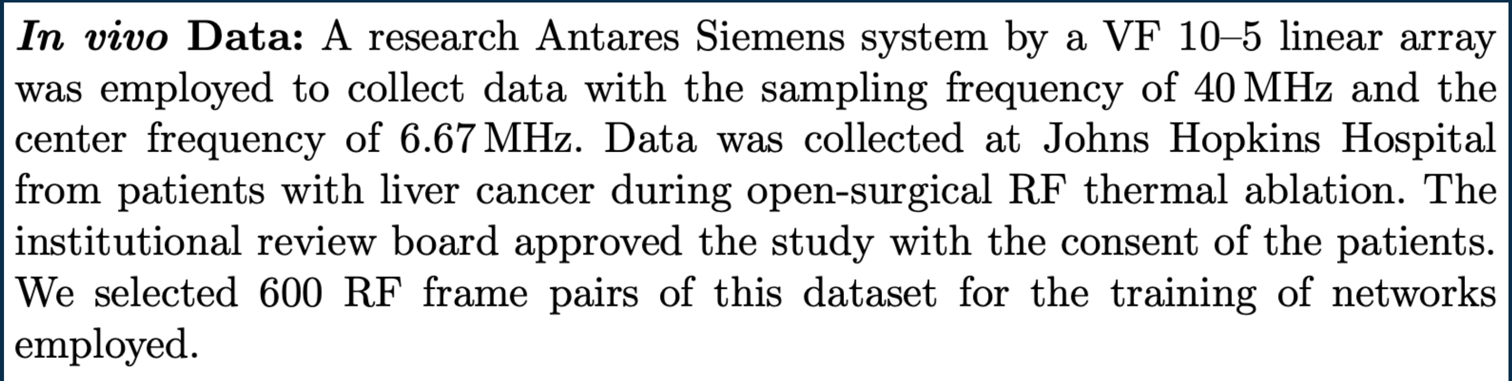


Experimental Phantom: A tissue-mimicking breast phantom (Model 059, CIRS: Tissue Simulation & Phantom Technology, Norfolk, VA) was employed for data collection. The elastic modulus of the phantom background was 20 kPa, and the phantom contained several inclusions having at least twice the elastic modulus of the background. An Alpinion E-Cube R12 research US machine (Bothell, WA, USA) with the sampling frequency of 40 MHz and the center frequency of 8 MHz was utilized for data collection of the training and test. To avoid data leakage, different parts of the phantom were imaged for training and test.

THE MICCAI REPRODUCIBILITY CHECKLIST

Datasets

Whether ethics approval was necessary for the data.

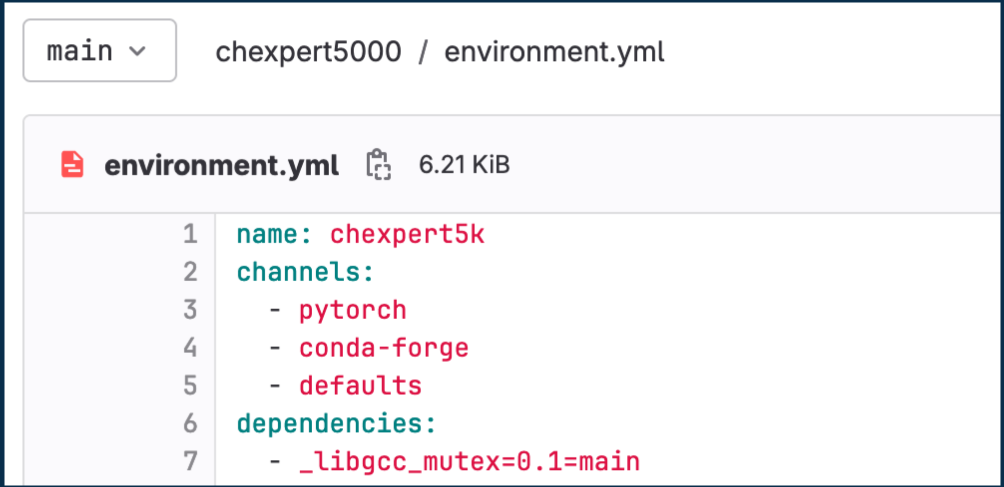


***In vivo* Data:** A research Antares Siemens system by a VF 10–5 linear array was employed to collect data with the sampling frequency of 40 MHz and the center frequency of 6.67 MHz. Data was collected at Johns Hopkins Hospital from patients with liver cancer during open-surgical RF thermal ablation. The institutional review board approved the study with the consent of the patients. We selected 600 RF frame pairs of this dataset for the training of networks employed.

THE MICCAI REPRODUCIBILITY CHECKLIST

Code

Specification of dependencies.

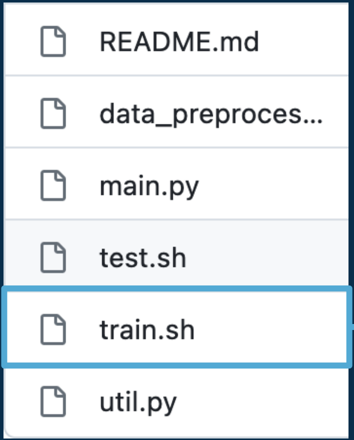


```
name: chexpert5k
channels:
  - pytorch
  - conda-forge
  - defaults
dependencies:
  - _libgcc_mutex=0.1=main
```

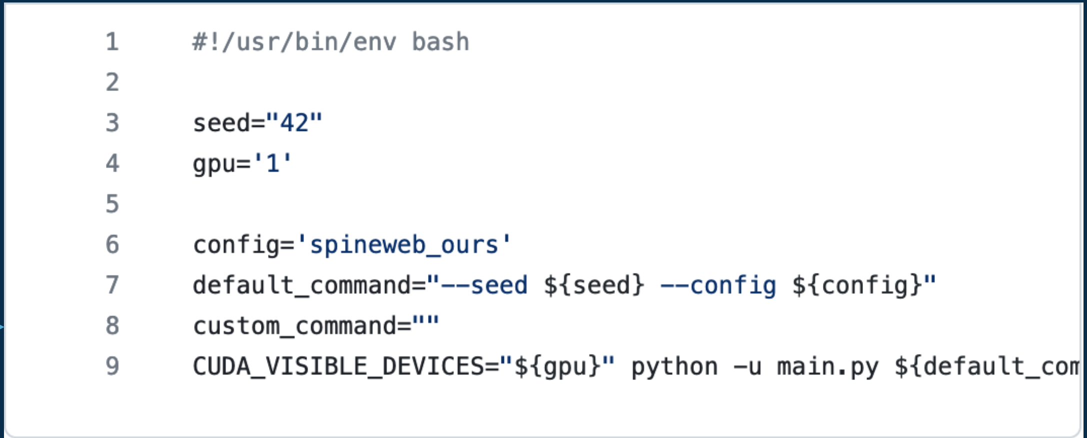
THE MICCAI REPRODUCIBILITY CHECKLIST

Code

Training code.



README.md
data_preproces...
main.py
test.sh
train.sh
util.py

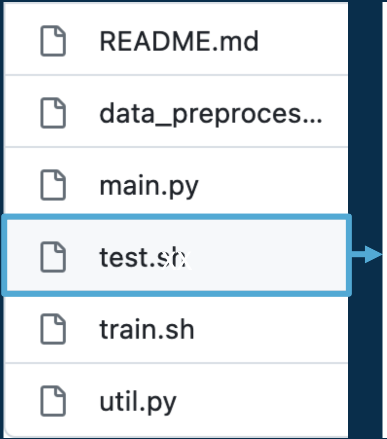


```
#!/usr/bin/env bash

seed="42"
gpu='1'

config='spineweb_ours'
default_command="--seed ${seed} --config ${config}"
custom_command=""
CUDA_VISIBLE_DEVICES="${gpu}" python -u main.py ${default_com
```

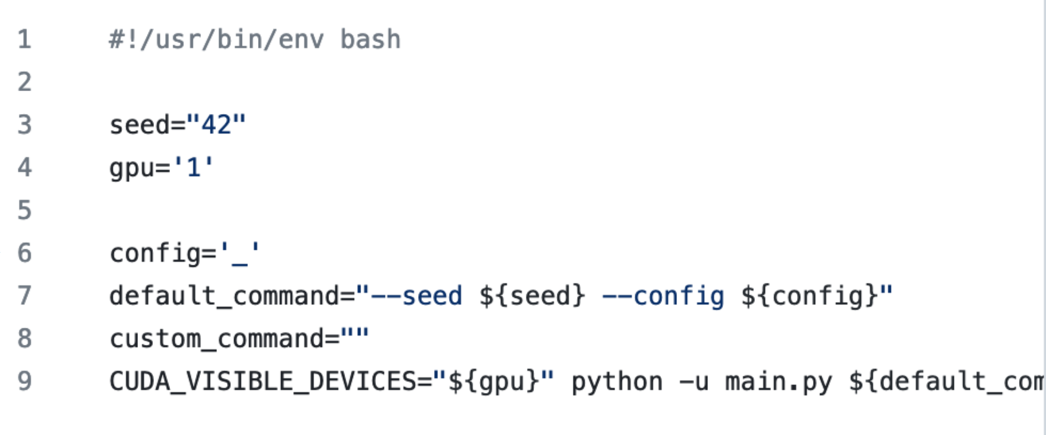
THE MICCAI REPRODUCIBILITY CHECKLIST

Code

Evaluation code.



README.md
data_preproces...
main.py
test.sh
train.sh
util.py

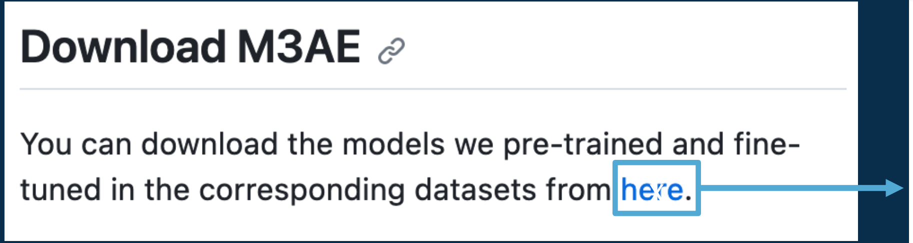


```
#!/usr/bin/env bash

seed="42"
gpu='1'

config='_'
default_command="--seed ${seed} --config ${config}"
custom_command=""
CUDA_VISIBLE_DEVICES="${gpu}" python -u main.py ${default_com
```

THE MICCAI REPRODUCIBILITY CHECKLIST

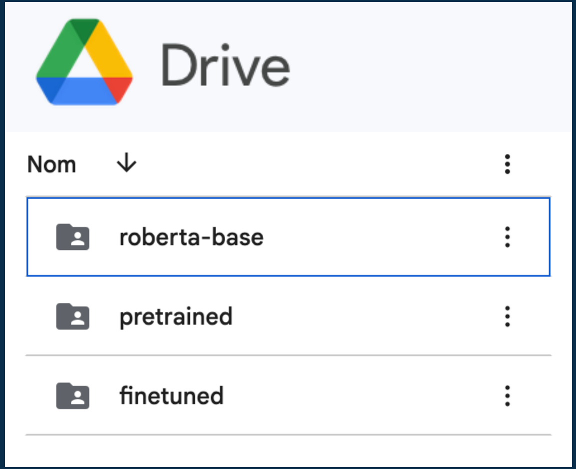
Code

(Pre-)trained model(s).



Download M3AE

You can download the models we pre-trained and fine-tuned in the corresponding datasets from here.

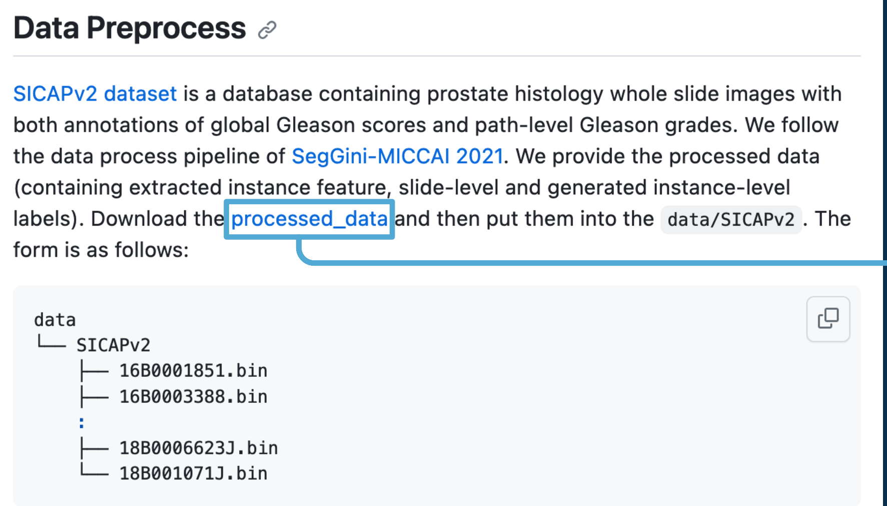


- roberta-base
- pretrained
- finetuned

THE MICCAI REPRODUCIBILITY CHECKLIST

Code

Dataset or link to the dataset needed to run the code.



Data Preprocess

SICAPv2 dataset is a database containing prostate histology whole slide images with both annotations of global Gleason scores and path-level Gleason grades. We follow the data process pipeline of SegGini-MICCAI 2021. We provide the processed data (containing extracted instance feature, slide-level and generated instance-level labels). Download the processed_data and then put them into the `data/SICAPv2`. The form is as follows:

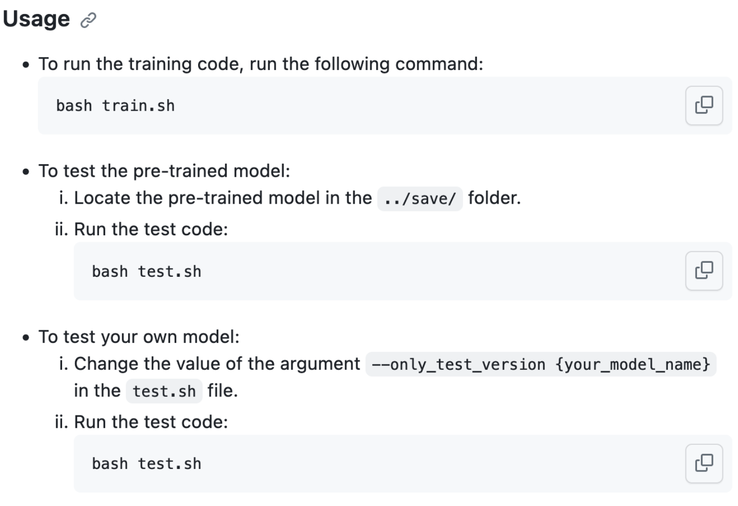
```
data
└── SICAPv2
    ├── 16B0001851.bin
    ├── 16B0003388.bin
    :
    ├── 18B0006623J.bin
    └── 18B001071J.bin
```

Drive

THE MICCAI REPRODUCIBILITY CHECKLIST

Code

README file including a table of results accompanied by precise command to run to produce those results.



Usage

- To run the training code, run the following command:

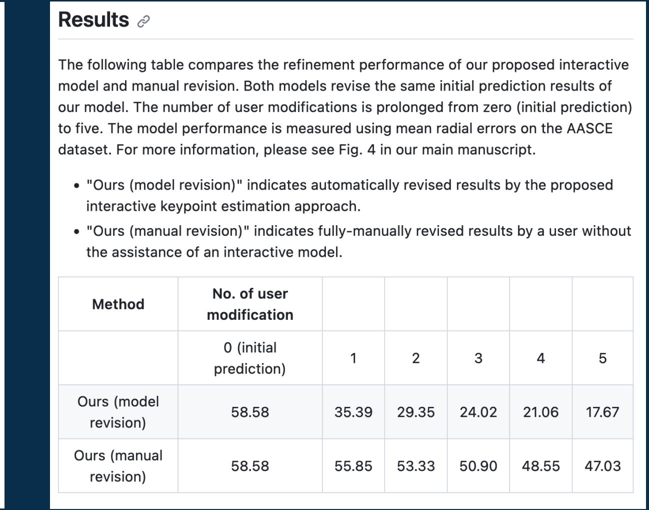
```
bash train.sh
```

- To test the pre-trained model:
  i. Locate the pre-trained model in the `../save/` folder.
  ii. Run the test code:

```
bash test.sh
```

- To test your own model:
  i. Change the value of the argument `--only_test_version {your_model_name}` in the `test.sh` file.
  ii. Run the test code:

```
bash test.sh
```



Results

The following table compares the refinement performance of our proposed interactive model and manual revision. Both models revise the same initial prediction results of our model. The number of user modifications is prolonged from zero (initial prediction) to five. The model performance is measured using mean radial errors on the AASCE dataset. For more information, please see Fig. 4 in our main manuscript.

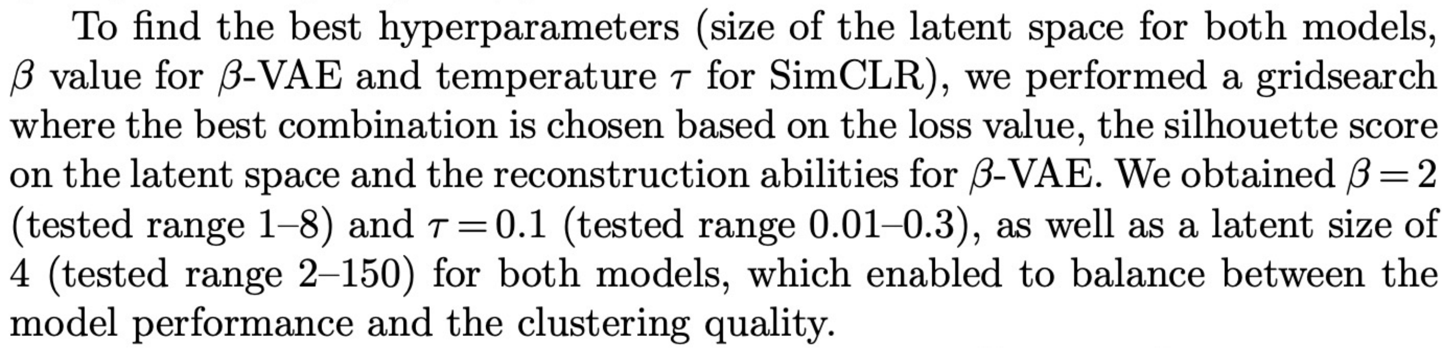
- "Ours (model revision)" indicates automatically revised results by the proposed interactive keypoint estimation approach.
- "Ours (manual revision)" indicates fully-manually revised results by a user without the assistance of an interactive model.

Method	No. of user modification					
	0 (initial prediction)	1	2	3	4	5
Ours (model revision)	58.58	35.39	29.35	24.02	21.06	17.67
Ours (manual revision)	58.58	55.85	53.33	50.90	48.55	47.03

THE MICCAI REPRODUCIBILITY CHECKLIST

Experimental results

The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results.

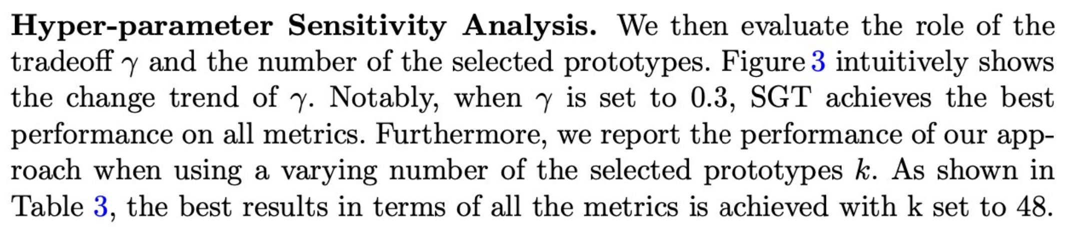


To find the best hyperparameters (size of the latent space for both models, β value for β-VAE and temperature τ for SimCLR), we performed a gridsearch where the best combination is chosen based on the loss value, the silhouette score on the latent space and the reconstruction abilities for β-VAE. We obtained $\beta = 2$ (tested range 1–8) and $\tau = 0.1$ (tested range 0.01–0.3), as well as a latent size of 4 (tested range 2–150) for both models, which enabled to balance between the model performance and the clustering quality.

THE MICCAI REPRODUCIBILITY CHECKLIST

Experimental results

Information on sensitivity regarding parameter changes.



Hyper-parameter Sensitivity Analysis. We then evaluate the role of the tradeoff γ and the number of the selected prototypes. Figure 3 intuitively shows the change trend of γ. Notably, when γ is set to 0.3, SGT achieves the best performance on all metrics. Furthermore, we report the performance of our approach when using a varying number of the selected prototypes k. As shown in Table 3, the best results in terms of all the metrics is achieved with k set to 48.

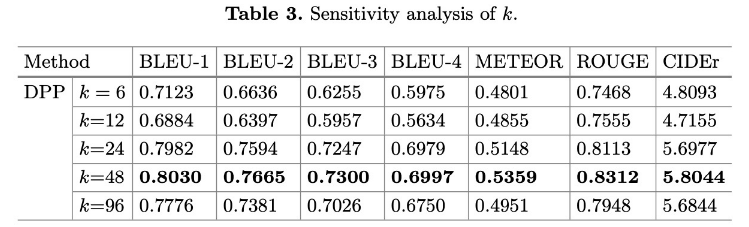
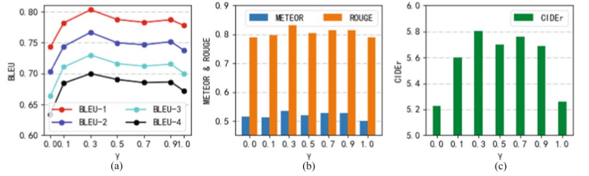


Table 3. Sensitivity analysis of k.

Method		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
DPP	$k = 6$	0.7123	0.6636	0.6255	0.5975	0.4801	0.7468	4.8093
	k=12	0.6884	0.6397	0.5957	0.5634	0.4855	0.7555	4.7155
	k=24	0.7982	0.7594	0.7247	0.6979	0.5148	0.8113	5.6977
	k=48	**0.8030**	**0.7665**	**0.7300**	**0.6997**	**0.5359**	**0.8312**	**5.8044**
	k=96	0.7776	0.7381	0.7026	0.6750	0.4951	0.7948	5.6844



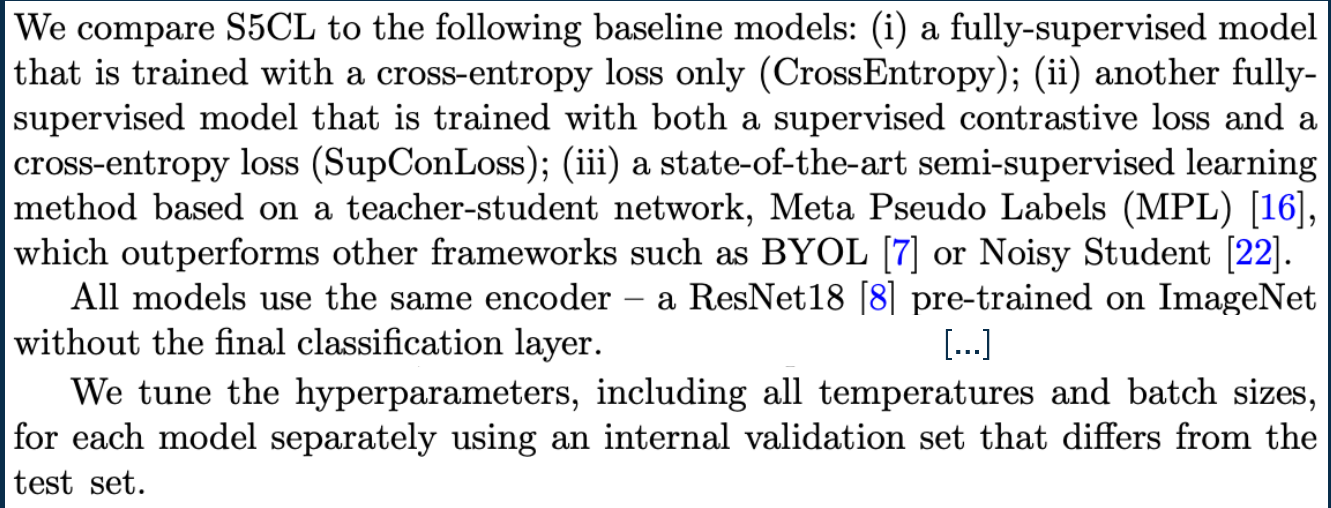
(a) (b) (c)

Fig. 3. Impact of the tradeoff γ on BLEU, METEOR, ROUGE and CIDEr.

THE MICCAI REPRODUCIBILITY CHECKLIST

Experimental results

Details on how baseline methods were implemented and tuned.



We compare S5CL to the following baseline models: (i) a fully-supervised model that is trained with a cross-entropy loss only (CrossEntropy); (ii) another fully-supervised model that is trained with both a supervised contrastive loss and a cross-entropy loss (SupConLoss); (iii) a state-of-the-art semi-supervised learning method based on a teacher-student network, Meta Pseudo Labels (MPL) [16], which outperforms other frameworks such as BYOL [7] or Noisy Student [22].

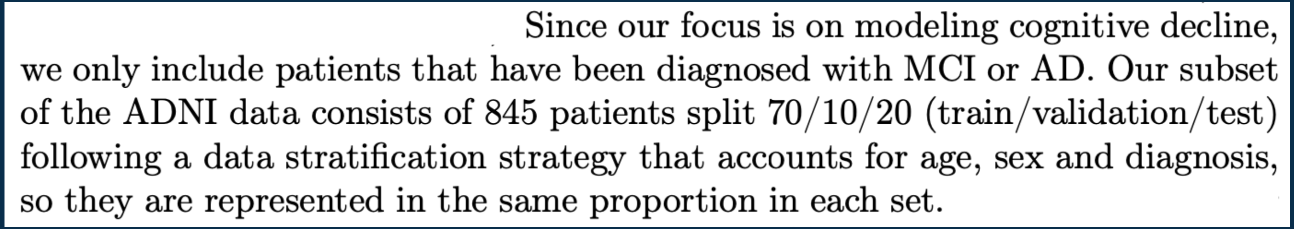
All models use the same encoder – a ResNet18 [8] pre-trained on ImageNet without the final classification layer. […]

We tune the hyperparameters, including all temperatures and batch sizes, for each model separately using an internal validation set that differs from the test set.

THE MICCAI REPRODUCIBILITY CHECKLIST

Experimental results

The details of train / validation / test splits.

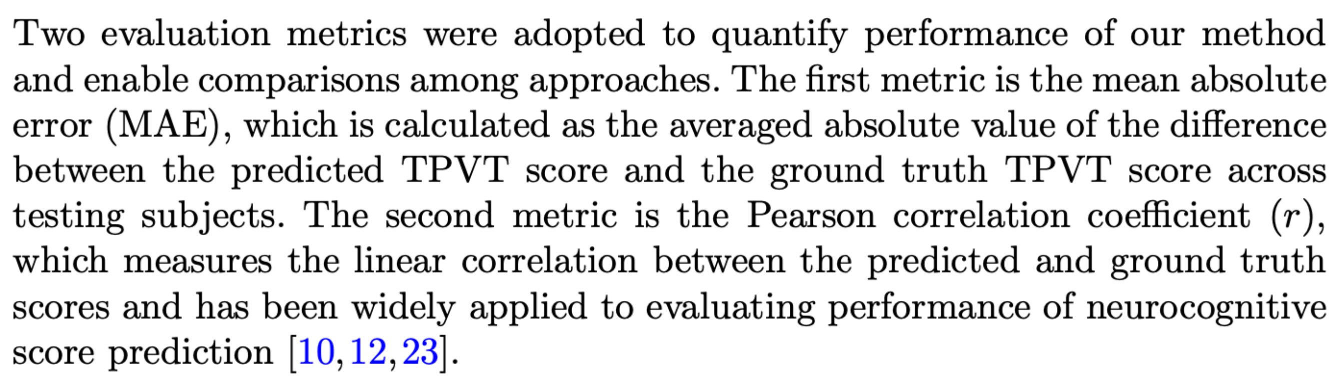


Since our focus is on modeling cognitive decline, we only include patients that have been diagnosed with MCI or AD. Our subset of the ADNI data consists of 845 patients split 70/10/20 (train/validation/test) following a data stratification strategy that accounts for age, sex and diagnosis, so they are represented in the same proportion in each set.

THE MICCAI REPRODUCIBILITY CHECKLIST

Experimental results

A clear definition of the specific evaluation metrics and/or statistics used to report results.



Two evaluation metrics were adopted to quantify performance of our method and enable comparisons among approaches. The first metric is the mean absolute error (MAE), which is calculated as the averaged absolute value of the difference between the predicted TPVT score and the ground truth TPVT score across testing subjects. The second metric is the Pearson correlation coefficient (r), which measures the linear correlation between the predicted and ground truth scores and has been widely applied to evaluating performance of neurocognitive score prediction [10,12,23].

THE MICCAI REPRODUCIBILITY CHECKLIST

Experimental results

A description of results with central tendency (e.g. mean) & variation (e.g. error bars).

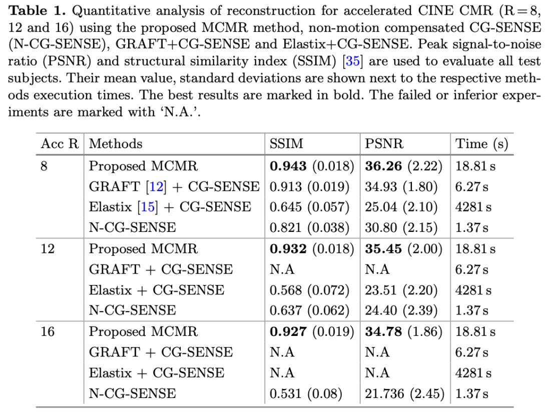


Table 1. Quantitative analysis of reconstruction for accelerated CINE CMR (R = 8, 12 and 16) using the proposed MCMR method, non-motion compensated CG-SENSE (N-CG-SENSE), GRAFT+CG-SENSE and Elastix+CG-SENSE. Peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [35] are used to evaluate all test subjects. Their mean value, standard deviations are shown next to the respective methods execution times. The best results are marked in bold. The failed or inferior experiments are marked with 'N.A.'.

Acc R	Methods	SSIM	PSNR	Time (s)
8	Proposed MCMR	**0.943** (0.018)	**36.26** (2.22)	18.81 s
	GRAFT [12] + CG-SENSE	0.913 (0.019)	34.93 (1.80)	6.27 s
	Elastix [15] + CG-SENSE	0.645 (0.057)	25.04 (2.10)	4281 s
	N-CG-SENSE	0.821 (0.038)	30.80 (2.15)	1.37 s
12	Proposed MCMR	**0.932** (0.018)	**35.45** (2.00)	18.81 s
	GRAFT + CG-SENSE	N.A	N.A	6.27 s
	Elastix + CG-SENSE	0.568 (0.072)	23.51 (2.20)	4281 s
	N-CG-SENSE	0.637 (0.062)	24.40 (2.39)	1.37 s
16	Proposed MCMR	**0.927** (0.019)	**34.78** (1.86)	18.81 s
	GRAFT + CG-SENSE	N.A	N.A	6.27 s
	Elastix + CG-SENSE	N.A	N.A	4281 s
	N-CG-SENSE	0.531 (0.08)	21.736 (2.45)	1.37 s

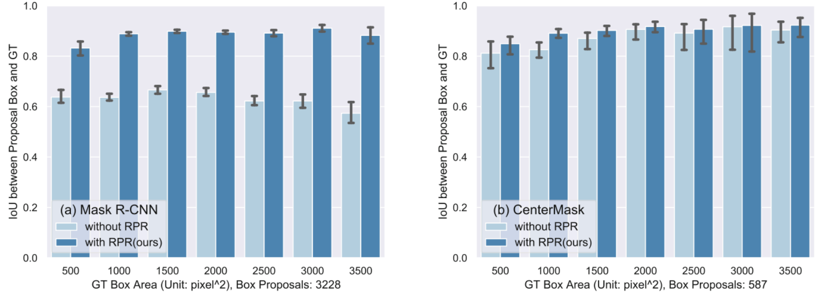


Fig. 4. Histogram of intersection-over-union (IoU) between region proposal boxes and ground truth (GT) boxes along with GT box areas on urothelial cell testing dataset. The box area unit is pixel^2. Error bars (mean ± std) are added.

THE MICCAI REPRODUCIBILITY CHECKLIST

Experimental results

An analysis of statistical significance of reported differences in performance between methods.

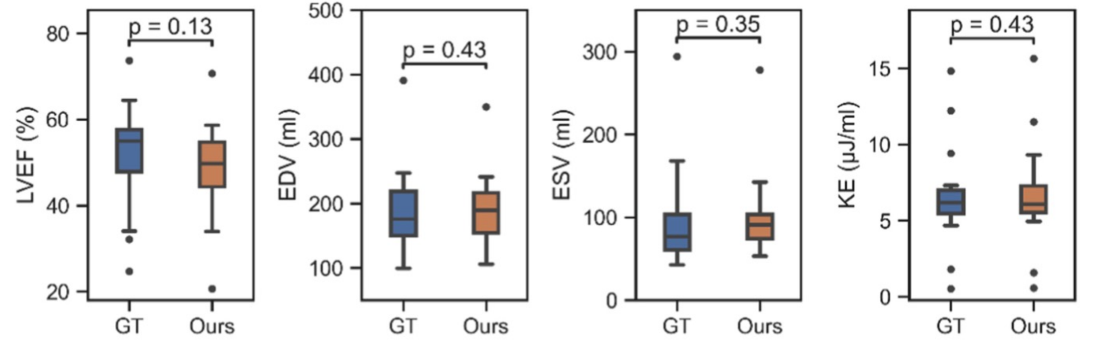


Fig. 5. Box plots comparing four clinical evaluation metrics including EDV, ESV, LVEF and KE derived from the manual segmentation and our prediction. GT represents the ground truth. P-value was computed using Wilcoxon-signed-rank test. $P < 0.05$ indicate a significant difference between two variables.

THE MICCAI REPRODUCIBILITY CHECKLIST

Experimental results

The average runtime for each result, or estimated energy cost.

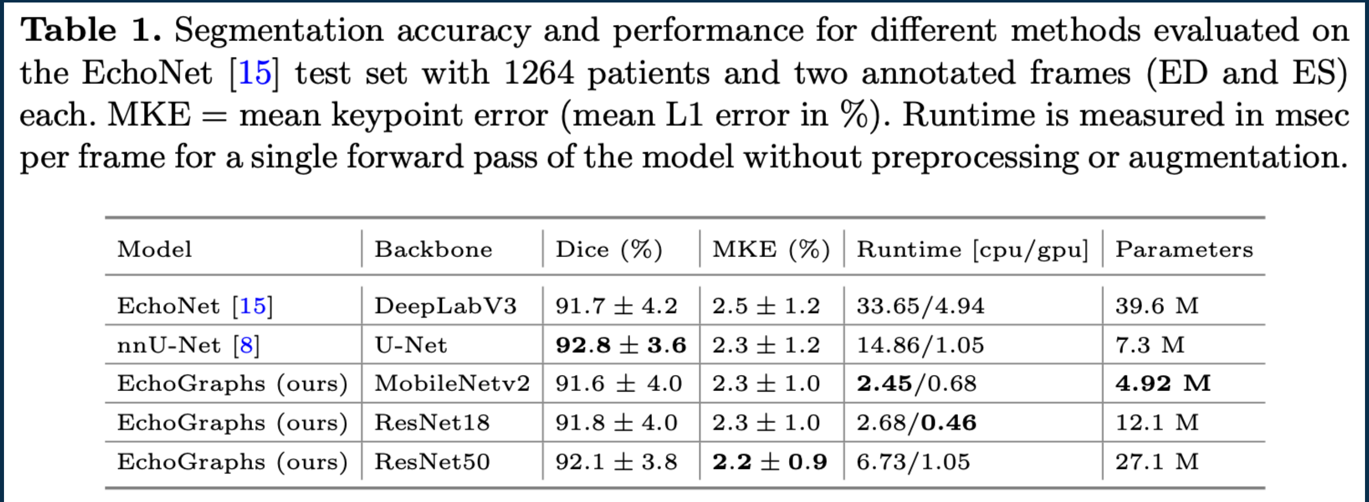


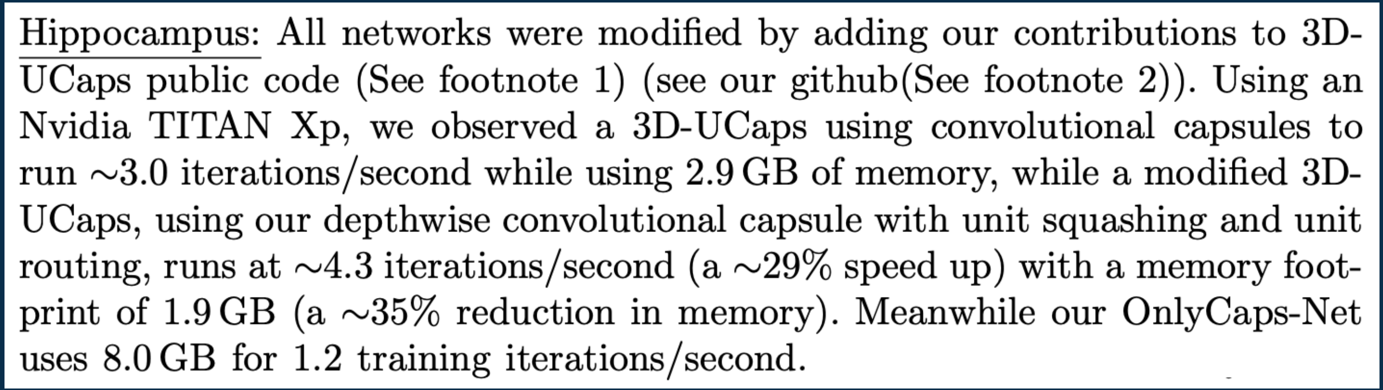
Table 1. Segmentation accuracy and performance for different methods evaluated on the EchoNet [15] test set with 1264 patients and two annotated frames (ED and ES) each. MKE = mean keypoint error (mean L1 error in %). Runtime is measured in msec per frame for a single forward pass of the model without preprocessing or augmentation.

Model	Backbone	Dice (%)	MKE (%)	Runtime [cpu/gpu]	Parameters
EchoNet [15]	DeepLabV3	91.7 ± 4.2	2.5 ± 1.2	33.65/4.94	39.6 M
nnU-Net [8]	U-Net	**92.8 ± 3.6**	2.3 ± 1.2	14.86/1.05	7.3 M
EchoGraphs (ours)	MobileNetv2	91.6 ± 4.0	2.3 ± 1.0	**2.45**/0.68	**4.92 M**
EchoGraphs (ours)	ResNet18	91.8 ± 4.0	2.3 ± 1.0	2.68/**0.46**	12.1 M
EchoGraphs (ours)	ResNet50	92.1 ± 3.8	**2.2 ± 0.9**	6.73/1.05	27.1 M

THE MICCAI REPRODUCIBILITY CHECKLIST

Experimental results

A description of the memory footprint.

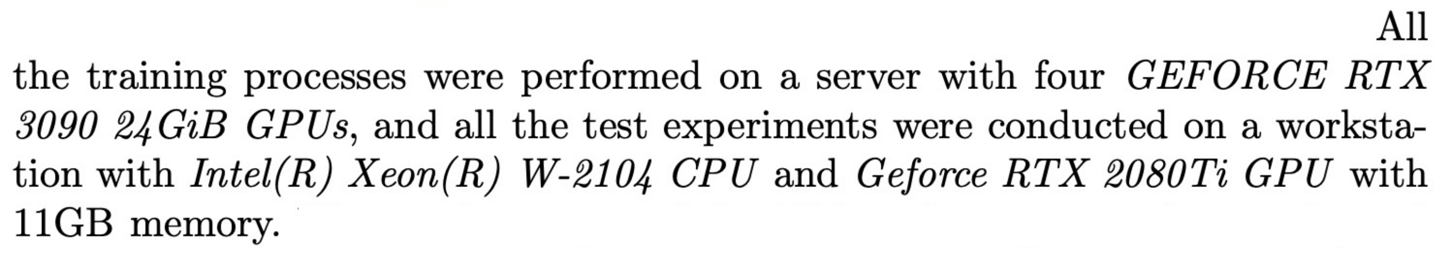


Hippocampus: All networks were modified by adding our contributions to 3D-UCaps public code (See footnote 1) (see our github(See footnote 2)). Using an Nvidia TITAN Xp, we observed a 3D-UCaps using convolutional capsules to run $\sim$3.0 iterations/second while using 2.9 GB of memory, while a modified 3D-UCaps, using our depthwise convolutional capsule with unit squashing and unit routing, runs at $\sim$4.3 iterations/second (a $\sim$29% speed up) with a memory footprint of 1.9 GB (a $\sim$35% reduction in memory). Meanwhile our OnlyCaps-Net uses 8.0 GB for 1.2 training iterations/second.

THE MICCAI REPRODUCIBILITY CHECKLIST

Experimental results

A description of the computing infrastructure used (hardware and software).



All the training processes were performed on a server with four *GEFORCE RTX 3090 24GiB GPUs*, and all the test experiments were conducted on a workstation with *Intel(R) Xeon(R) W-2104 CPU* and *Geforce RTX 2080Ti GPU* with 11GB memory.

THE MICCAI REPRODUCIBILITY CHECKLIST

Experimental results

An analysis of situations in which the method failed.

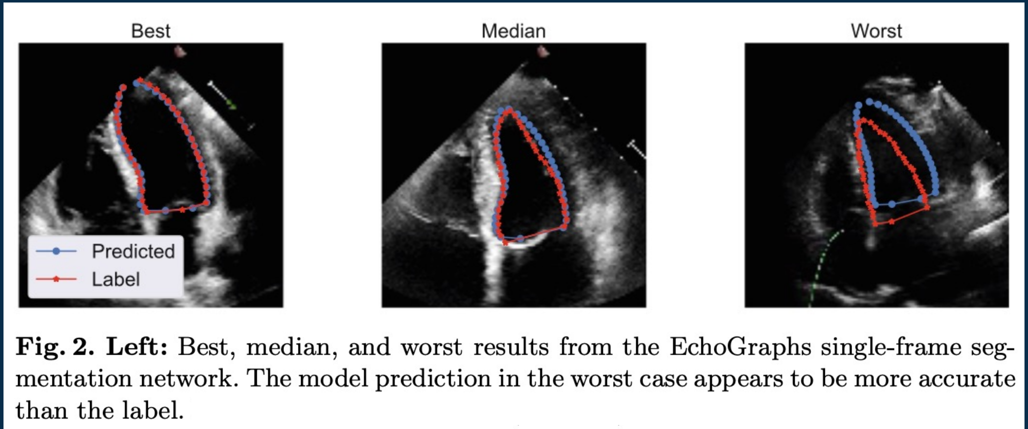
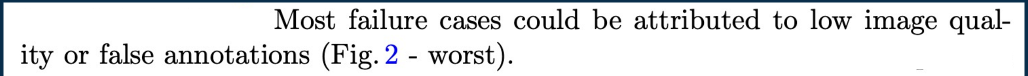


Fig. 2. Left: Best, median, and worst results from the EchoGraphs single-frame segmentation network. The model prediction in the worst case appears to be more accurate than the label.

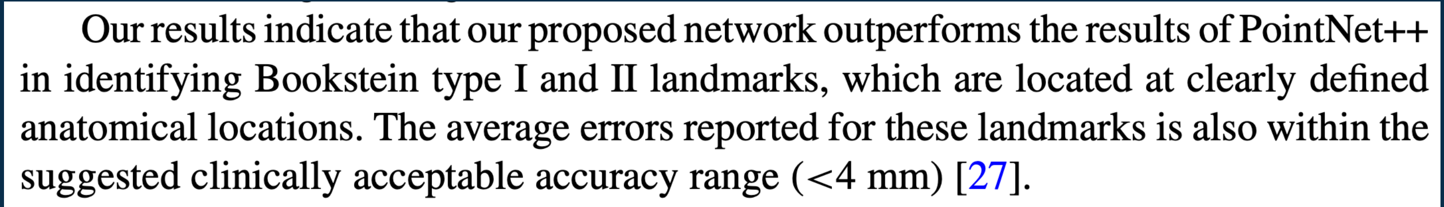


Most failure cases could be attributed to low image quality or false annotations (Fig. 2 - worst).

THE MICCAI REPRODUCIBILITY CHECKLIST

Experimental results

Discussion of clinical significance.



Our results indicate that our proposed network outperforms the results of PointNet++ in identifying Bookstein type I and II landmarks, which are located at clearly defined anatomical locations. The average errors reported for these landmarks is also within the suggested clinically acceptable accuracy range (<4 mm) [27].

THE MICCAI REPRODUCIBILITY CHECKLIST

ON THE REVIEWER SIDE

‘Please comment on the reproducibility of the paper. Note, that authors have filled out a reproducibility checklist upon submission. Please be aware that authors are not required to meet all criteria on the checklist - for instance, providing code and data is a plus, but not a requirement for acceptance.’

THE MICCAI REPRODUCIBILITY CHECKLIST

ON THE REVIEWER SIDE

Reviewer 1	Reviewer 2	Reviewer 3
Limited reproducibility.	Please add the batch size used for training the networks. Please provide the computational time (at training and testing) of the networks. Additional statistical tests are needed to evaluate the significance of the results. Please use statistical tests to compare the performance of the model trained with multi-centric MRIs to those trained with MRIs from a single center. Did data augmentations were performed to improve the method generalization? Please describe the data. What are the MRI scanners used for the data acquisition? How many Teslas? What are the MRI parameters (TE, TR, flip angle, acquisition time, etc)? Did this study consider the whole MRI volumes or only 2D slice MRIs? It will be great to consider the whole MRI volumes during the method evaluation (that is more needed for clinical practice). JIM model was trained with MRIs from institutions 1, 2, and 3 and evaluated on MRIs from institution 4. That seems not enough to conclude that the model trained with multi-centric MRIs had better generalizability and accuracy than those trained with images from a single institution. I think it will be interesting to train and evaluate the JIM model with MRIs from distinct combinations of institutions (e.g., train JIM with MRI from institutions 2, 3, and 4 and evaluate it on MRI from institution 1) and compare the obtained results (evaluation metrics + statistical tests).	high reproducibility.

THE MICCAI REPRODUCIBILITY CHECKLIST

ON THE REVIEWER SIDE

Reviewer 1	Reviewer 2	Reviewer 3
I don't see any limitations on the reproducibility.	Study can be reproduced	Not reproducible - although publicly available datasets are used, implementation code has not been made available
The authors ticked Yes for most items of the reproducibility checklist, which is sometimes in contradiction with what is provided in the paper (e.g. range of hyper-parameters considered). It seems that the code will be made available.	The authors did not provide their statement on code availability. This would be highly desirable.	ok
Its reproducibility is very high. The method and parameter explanations are excellent, and the data-related parts and learning environment are also well explained. I look forward to the code release of this paper.	It is very difficult to reproduce the paper and achieved results from description of the manuscript. The authors use a subset and crop the videos but any of this information is reported.	According to the authors, the code will be made public and the dataset is publicly available.

RESOURCES

- MICCAI 2020 Hackathon: https://2020.miccai-hackathon.com
- Repo-Review: https://github.com/scientific-python/repo-review
- Repo anonymisation:
  - https://anonymous.4open.science
  - https://www.micahsmith.com/blog/2019/10/anonymize-github-repos-double-blind
- List of open-source MICCAI papers: https://github.com/JunMa11/MICCAI-OpenSourcePapers
- Book chapter (the chapter itself and all its references!): Colliot, O., Thibeau-Sutre, E., and Burgos, N.: ‘Reproducibility in Machine Learning for Medical Imaging’. In Machine Learning for Brain Disorders, edited by Olivier Colliot, Neuromethods, Springer, 2023. doi:10.1007/978-1-0716-3195- 9_21