Benchmarking the CoW with the TopCoW Challenge: Topology-Aware Anatomical Segmentation of the Circle of Willis for CTA and MRA

Kaiyuan Yang[a,*], Fabio Musio[a,b,*], Yihui Ma[c,d,*], Norman Juchler[b], Johannes C. Paetzold[e], Rami Al-Maskari[f,g], Luciano Höher[f], Hongwei Bran Li[a,h], Ibrahim Ethem Hamamci[a], Anjany Sekuboyina[a], Suprosanna Shit[a], Houjing Huang[a], Chinmay Prabhakar[a], Ezequiel de la Rosa[a], Bastian Wittmann[a], Diana Waldmannstetter[a,g], Florian Kofler[a,g,i,j], Fernando Navarro[a,g,i], Martin J. Menten[g,k,l], Ivan Ezhov[g], Daniel Rueckert[g,i,k,l], Iris N. Vos[m], Ynte M. Ruigrok[n], Birgitta K. Velthuis[o], Hugo J. Kuijf[m], Pengcheng Shi[p], Wei Liu[p], Ting Ma[p,q], Maximilian R. Rokuss[r,s], Yannick Kirchhoff[r,s,u], Fabian Isensee[r,t], Klaus Maier-Hein[r,v], Chengcheng Zhu[w], Huilin Zhao[x], Philippe Bijlenga[y,†], Julien Hämmerli[y,†], Catherine Wurster[y,†], Laura Westphal[z,†], Jeroen Bisschop[aa,†], Elisa Colombo[ab,†], Hakim Baazaoui[z,†], Hannah-Lea Handelsmann[z,†], Andrew Makmur[ac,†], James Hallinan[ac,†], Amrish Soundararajan[ad,†], Bene Wiestler[i,†], Jan S. Kirschke[i,†], Roland Wiest[ae,†], Emmanuel Montagnon[af,#], Laurent Letourneau-Guillon[af,#], Kwanseok Oh[ag,ah,#], Dahye Lee[ag,#], Orhun Utku Aydin[ai,#], Adam Hilbert[ai,#], Jana Rieger[ai,#], Dimitrios Rallios[ai,#], Satoru Tanioka[ai,#], Alexander Koch[ai,#], Dietmar Frey[ai,#], Abdul Qayyum[aj,#], Moona Mazher[ak,#], Steven Niederer[aj,#], Nico Disch[r,s,u,#], Julius Holzschuh[r,#], Dominic LaBella[al,#], Francesco Galati[am,#], Daniele Falcetta[am,#], Maria A. Zuluaga[am,#], Chaolong Lin[an,#], Haoran Zhao[an,#], Zehan Zhang[ao,#], Minghui Zhang[ap,aq,#], Xin You[ap,aq,#], Hanxiao Zhang[ap,#], Guang-Zhong Yang[ap,#], Yun Gu[ap,aq,#], Sinyoung Ra[ar,#], Jongyun Hwang[ar,#], Hyunjin Park[as,#], Junqiang Chen[at,#], Marek Wodzinski[au,av,#], Henning Müller[au,#], Nesrin Mansouri[aw,ax,#], Florent Autrusseau[aw,ax,#], Cansu Yalçin[ay,#], Rachika E. Hamadache[ay,#], Clara Lisazo[ay,#], Joaquim Salvi[ay,#], Adrià Casamitjana[ay,#], Xavier Lladó[ay,#], Uma Maria Lal-Trehan Estrada[ay,#], Valeriia Abramova[ay,#], Luca Giancardo[az,#], Arnau Oliver[ay,#], Paula Casademunt[ba,#], Adrian Galdran[ba,#], Matteo Delucchi[b,bb,#], Jialu Liu[bc,bd,#], Haibin Huang[bc,bd,#], Yue Cui[bc,bd,#], Zehang Lin[be,#], Yusheng Liu[bf,#], Shunzhi Zhu[be,#], Tatsat R. Patel[bg,bi,#], Adnan H. Siddiqui[bg,bi,#], Vincent M. Tutino[bg,bh,#], Maysam Orouskhani[w,#], Huayu Wang[w,#], Mahmud Mossa-Basha[w,#], Yuki Sato[bj,#], Sven Hirsch[b,**], Susanne Wegener[z,**], Bjoern Menze[a,**]

[a] *Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland*
[b] *Institute of Computational Life Sciences, Zurich University of Applied Sciences (ZHAW), Waedenswil, Switzerland*
[c] *Department of Neuroradiology, University Hospital of Zurich, Zurich, Switzerland*
[d] *Department of Neurosurgery, Zhongnan Hospital of Wuhan University, Wuhan, China*
[e] *Department of Radiology at Weill Cornell Medicine, Cornell University, New York, USA*
[f] *Institute for Tissue Engineering and Regenerative Medicine (iTERM), Helmholtz Munich, Neuherberg, Germany*
[g] *School of Computation, Information and Technology, Technical University of Munich, Germany*
[h] *Athinoula A. Martinos Center for Biomedical Imaging, Harvard Medical School, Boston, USA*
[i] *School of Medicine and Health, TUM Klinikum, Technical University of Munich, Germany*
[j] *Helmholtz AI, Helmholtz Munich, Neuherberg, Germany*
[k] *Munich Center for Machine Learning, Munich, Germany*
[l] *Department of Computing, Imperial College London, London, UK*
[m] *Image Sciences Institute, UMC Utrecht, Utrecht, The Netherlands*
[n] *Department of Neurology and Neurosurgery, University Medical Center Utrecht, Utrecht, The Netherlands*
[o] *Department of Radiology, University Medical Center Utrecht, Utrecht, The Netherlands*
[p] *Electronic & Information Engineering School, Harbin Institute of Technology (Shenzhen), China*
[q] *Peng Cheng Laboratory, Shenzhen, China*
[r] *Division of Medical Image Computing, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[s] *Faculty of Mathematics and Computer Science, Heidelberg University, Germany*
[t] *Helmholtz Imaging, German Cancer Research Center, Heidelberg, Germany*
[u] *HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany*
[v] *Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital*
[w] *Department of Radiology, University of Washington, Seattle, WA, USA*
[x] *Department of Radiology, Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China*
[y] *Department of Clinical Neurosciences, Division of Neurosurgery, Geneva University Hospitals, Geneva, Switzerland*
[z] *Department of Neurology, University Hospital of Zurich, Zurich, Switzerland*
[aa] *Department of Physiology, University of Toronto, Canada*
[ab] *Department of Neurosurgery, University Hospital of Zurich, Zurich, Switzerland*
[ac] *Department of Diagnostic Imaging, National University Hospital, Singapore*
[ad] *University of Chicago, USA*
[ae] *Department of Diagnostic and Interventional Neuroradiology, University Hospital Berne and University of Berne, Berne, Switzerland*
[af] *Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM), Montréal, Québec, Canada*

[*] K.Y., F.M., and Y.M. contributed equally
[**] S.H., S.W., and B.M. are co-corresponding authors
e-mail: `hirc@zhaw.ch` (Sven Hirsch), `susanne.wegener@usz.ch` (Susanne Wegener), `bjoern.menze@uzh.ch` (Bjoern Menze)
[†] Clinical committee
[#] Participant of the challenge, ordered alphabatically by team name

[ag] *DEEPNOID Inc., Seoul, South Korea*
[ah] *Department of Artificial Intelligence, Korea University, Seoul, South Korea*
[ai] *Charité Lab for AI in Medicine (CLAIM), Charité Universitätsmedizin Berlin, Berlin, Germany*
[aj] *National Heart and Lung Institute, Faculty of Medicine, Imperial College London, London, UK*
[ak] *Centre for Medical Image Computing, Department of Computer Science, University College London, London, UK*
[al] *Department of Radiation Oncology, Duke University Medical Center, Durham, NC, USA*
[am] *EURECOM, Biot, France*
[an] *Institute of Medical Technology, Peking University Health Science Center, Beijing, China*
[ao] *Hangzhou Genlight MedTech Co., Ltd., China*
[ap] *Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China*
[aq] *Department of Automation, Shanghai Jiao Tong University, Shanghai, China*
[ar] *Department of Artificial Intelligence, Sungkyunkwan University, Seoul, South Korea*
[as] *Department of Electrical and Computer Engineering, Sungkyunkwan University, Seoul, South Korea*
[at] *Shanghai MediWorks Precision Instruments Co., Ltd., China*
[au] *Institute of Informatics, HES-SO Valais-Wallis, Switzerland*
[av] *Department of Measurement and Electronics, AGH University of Krakow, Poland*
[aw] *Institut du Thorax (ITX), Université Nantes, Nantes, France*
[ax] *Laboratoire de Thermique et Energie de Nantes (LTeN), Université Nantes, Polytech'Nantes, Nantes, France*
[ay] *Research Institute of Computer Vision and Robotics (ViCOROB), Universitat de Girona, Catalonia, Spain*
[az] *Center for Precision Health, McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, USA*
[ba] *Physense, BCN-Medtech, Department of Communication and Information Technologies, Universitat Pompeu Fabra, Barcelona, Spain*
[bb] *Department of Mathematical Modeling and Machine Learning, University of Zurich, Zurich, Switzerland*
[bc] *Laboratory of Brain Atlas and Brain-inspired Intelligence, Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[bd] *School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China*
[be] *School of Computer and Information Engineering, Xiamen University of Technology, Xiamen, China*
[bf] *Department of Automation, Shanghai Jiao Tong University, Shanghai, China*
[bg] *Canon Stroke and Vascular Research Center, University at Buffalo, NY, USA*
[bh] *Department of Pathology and Anatomical Sciences, University at Buffalo, NY, USA*
[bi] *Department of Neurosurgery, University at Buffalo, NY, USA*
[bj] *LPIXEL Inc., Tokyo, Japan*

ARTICLE INFO

Article history:

Keywords: Circle of Willis, Vessel Segmentation, Variant Classification, Brain CT Angiography, Brain MR Angiography, Virtual Reality, Fetal PCA, Aneurysm Location

ABSTRACT

The Circle of Willis (CoW) is an important network of arteries connecting major circulations of the brain. Its vascular architecture is believed to affect the risk, severity, and clinical outcome of serious neurovascular diseases. However, characterizing the highly variable CoW anatomy is still a manual and time-consuming expert task. The CoW is usually imaged by two non-invasive angiographic imaging modalities, magnetic resonance angiography (MRA) and computed tomography angiography (CTA), but there exist limited datasets with annotations on CoW anatomy, especially for CTA. Therefore, we organized the TopCoW challenge with the release of an annotated CoW dataset. The TopCoW dataset is the first public dataset with voxel-level annotations for 13 CoW vessel components, enabled by virtual reality technology. It is also the first large dataset using 200 pairs of MRA and CTA from the same patients. As part of the benchmark, we invited submissions worldwide and attracted over 250 registered participants from six continents. The submissions were evaluated on both internal and external test datasets of 226 scans from over five centers. The top performing teams achieved over 90% Dice scores at segmenting the CoW components, over 80% F1 scores at detecting key CoW components, and over 70% balanced accuracy at classifying CoW variants for nearly all test sets. The best algorithms also showed clinical potential in classifying fetal-type posterior cerebral artery and locating aneurysms with CoW anatomy. TopCoW demonstrated the utility and versatility of CoW segmentation algorithms for a wide range of downstream clinical applications with explainability. The annotated datasets and best performing algorithms have been released as public Zenodo records to foster further methodological development and clinical tool building.

1. Introduction

The Circle of Willis (CoW) is an important anastomotic network of arteries connecting the anterior and posterior circulations of the brain, as well as the left and right cerebral hemispheres [1]. Due to its centrality, the CoW is commonly involved in pathologies like aneurysms and stroke. Clinically, the vascular architecture of the CoW is believed to impact the occurrence and severity of stroke [2, 3, 4, 5], pose a potential risk for aneurysm formation [6], and affect the neurologic events and clinical outcomes of neurosurgeries [7, 8]. An accurate characterization of the CoW is therefore of great clinical relevance.

However, clinicians have articulated an unmet demand for efficient software tools to analyze the angio-architecture of the CoW. Assessing the anatomy and vascular components of the CoW from angiography images is still a manual and time-consuming task requiring specialist judgment. The CoW anatomy involves multiple connections and branches of different cerebral vessels. These vessels vary in diameters from around 1 to 4 mm [9]. CoW vessel components are difficult to identify accurately in isolation and often require subtle spatial relationship to distinguish them anatomically. The vessels also have curvatures and turns along their courses. This can result in vessels crossing paths on the angiography images but can be difficult to differentiate whether the vessels are just touching or there are communicating blood flows at the crossing points. Furthermore, the CoW naturally has many variants of which certain principal artery components are hypoplastic or absent. It is estimated that only less than around half of our population has a complete CoW [9, 10]. It is common to see the CoW anatomies vary markedly from person to person. Characterizing the CoW anatomies can therefore be a challenging task due to the complexity and heterogeneity of the anatomy.

The brain arteries, including the CoW, are commonly diagnosed and imaged by two non-invasive angiographic imaging modalities, namely magnetic resonance angiography (MRA) and computed tomography angiography (CTA). There have been a number of publicly available datasets on MRA modality. Earlier MRA datasets include the CASILab (also known as TubeTK or MIDAS) [11] and the IXI [12] datasets, which were acquired from scanners before 2006 and with limited vessel annotations. Recently, more MRA datasets have been published and some were annotated with binary vessel masks, such as the CAS [13] dataset, the SMILE-UHURA [14] dataset on 7T MRA, and the COSTA [15] dataset that offered a subset of annotated CASIL and IXI images. However, the vessel annotation was in binary and there were no anatomical annotations on the CoW. Furthermore, to our knowledge, annotated dataset on the other important modality, CTA, did not exist.

Previously, there has been a high barrier to entry for annotating the CoW anatomy: one would not only need expert-level neuroanatomical knowledge to label or verify the complex and variable CoW anatomy, but also have to overcome the laborious and time-consuming process of 2D annotation for multiclass CoW vessels. To address such annotation obstacles, we turned to virtual reality (VR), which helped attract clinicians' interest and engage them in the annotation process via its appealing visualization and gamification aspects. VR also significantly accelerated the annotation and verification process for tortuous and intertwined vessels via an intuitive 3D workflow.

Prior work on the CoW anatomy characterization task has been developed mainly as a labeling task built upon binary vessel masks, skeletons or graphs [16, 17, 18, 19, 20, 21], and with two recent studies that directly tackled the problem as a multiclass segmentation task [22, 23]. However, only private annotated in-house data or public data without verified CoW annotations were used, and the studies were restricted to only the MRA modality. Furthermore, given the complex and highly heterogeneous anatomies of the CoW in clinical settings, the difficulties associated with the CoW anatomy characterization task in past studies have not been sufficiently conveyed or discussed. We thus identify the following contributions we can make to the field: 1) We provide open data with verified annotations for CoW segmentation benchmarking. A public annotated CoW dataset can benefit algorithm development and comparison. 2) We include the CTA modality. Clinically, CTA is an equally important angiography modality as MRA for CoW anatomy diagnosis. 3) We shed light on the anatomical complexities of CoW variants and evaluate the clinical relevance of current algorithms in handling such complexities.

To this end, we organized a benchmark on "Topology-Aware Anatomical Segmentation of the Circle of Willis for CTA and MRA", or "TopCoW" for short, as registered and included in the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference held in 2023 and 2024. TopCoW was the first public challenge on CoW anatomical segmentation featuring voxel-level vessel annotations on two common non-invasive angiographic imaging modalities, MRA and CTA. The main tasks of the challenge were to automatically segment the CoW vessels and to classify the CoW variants on 3D angiographic images. We collected submissions from global participants and evaluated the performance on both internal and external multi-center test datasets. In addition, we evaluated the CoW segmentation algorithms on two clinical downstream tasks that were routinely performed in practice: classifying fetal-type posterior cerebral artery (PCA) and locating intracranial aneurysms using the CoW anatomy, with the aim of assessing the ability of the submissions to address real-world clinical needs. For transparent benchmark and continued development, we made our annotated datasets and the best algorithm submissions publicly available on Zenodo records[3].

2. Challenge Dataset and Study Design

2.1. TopCoW Dataset

The TopCoW data cohort was composed of patients admitted to the Stroke Center of the University Hospital Zurich (USZ) in 2018 and 2019. The data were acquired at the hospital during routine examinations following standard procedures of the respective imaging. Siemens scanners were used for both modalities. MRA scans were imaged with magnetic field strength

[3] `https://zenodo.org/records/15692630` and `https://zenodo.org/records/15665435`

of 3 Tesla or 1.5 Tesla. In total, 200 pairs of MRA and CTA scans from unique patients were curated for the TopCoW challenge and subsequently split into train and test sets arbitrarily by patient. The inclusion criteria for the TopCoW data were: 1) both MRA and CTA scans were available and in good quality for that patient; 2) at least the MRA or CTA allowed for an assessment of the CoW anatomy; 3) no large aneurysms inside the CoW ROI; 4) certain rare CoW variants that could not be characterized by our 13 annotated CoW components were excluded. All TopCoW data were anonymized, defaced and cropped to the braincase region. Training, validation, and internal test cases all had the MRA and CTA joint-modality pairs, with one scan for each modality. Between the two years of the challenge, the training dataset grew from 90 to 125 patients, and the test dataset grew from 35 to 70 patients. The final 2024 training dataset had 125 patients with the images and annotations released to the public. The validation set included 5 patients whose annotations were not publicly released but were used on the submission website to allow participants to dry run their submissions, which was not included in the final benchmark evaluation. The internal test set had 70 patients and was hidden from the public. The TopCoW data used had been approved by the local ethical committee. The anonymized image data were approved to be released under the "Open use. Must provide the source. Use for commercial purposes requires permission of the data owner." license from the OpenData Swiss [24].

More information on the data cohort, inclusion and exclusion of CoW variants, the anonymization pre-processing, and dataset changelog are in the Supplementary S1, S2, S3, and S4.

2.2. Data Annotation

For each 3D angiography image, we provided three types of annotation regarding the CoW: the voxel-level multiclass segmentation mask of the CoW, a 3D bounding box for the CoW region of interest (ROI), and the CoW variant graph. Virtual reality (VR) was used to efficiently annotate and verify the CoW anatomy in 3D. Fig. 1a shows the workflow and view from VR. The VR annotation setup followed the method as described in [25]. There were 13 CoW vessel components for the multiclass segmentation annotation: left and right internal carotid artery (ICA), left and right anterior cerebral artery (ACA), left and right middle cerebral artery (MCA), anterior communicating artery (Acom), left and right posterior communicating artery (Pcom), left and right posterior cerebral artery (PCA), and basilar artery (BA). Occasionally the anterior part of the CoW can have a third A2 artery arising from the Acom, and we labeled it with class 3rd-A2. Fig. 1a right shows an MRA example with all 13 CoW vessel class labeled. The CoW annotation protocol was designed by a senior neurosurgeon (Y.M., over 10 years of experience) and reviewed by a senior neurosurgeon (P.B., over 15 years of experience) and a senior neurologist (S.W., over 15 years of experience). Y.M. used around 35 initial patients to educate and train the annotators (K.Y. and F.M.) on the CoW anatomical knowledge and the annotation protocol. Around another 40 patients from subsequent cases that the annotators were uncertain of were reviewed and verified by Y.M.. Second opinions and verifications were also obtained from other neurosurgeons (P.B., J.H., C.W., E.C.) and neurologists (S.W., L.W., H.B.) for around 15 patients. All annotated data used in the benchmark were manually verified by at least one annotator. Further details on the CoW annotation protocol can be found in Supplementary S5.

Fig. 1b shows an example of segmentation mask and ROI annotations for both MRA and CTA modalities from a TopCoW patient. The CoW ROI was defined as the 3D bounding box containing the volume required for the diagnosis of the CoW variant with a padding. For higher sensitivity, we pad the bounding box with roughly the diameter of the ICA to include slightly more regions in the ROI.

The third type of annotation, the CoW variant graph, was derived from the segmentation mask and encompasses the anterior variant (AV) and posterior variant (PV) graphs. Fig. 1c shows the AV and PV graph composition. The AV graph was determined by four edges: L-A1, Acom, 3rd-A2, and R-A1. The corresponding edge-list was defined by 0 or 1 according to the edge presence. For example, AV-1001 is the anterior variant that has L-A1 present, Acom and 3rd-A2 absent, and R-A1 present. Similarly, the PV graph was determined by a four-element edge-list of L-Pcom, L-P1, R-P1, and R-Pcom.

The TopCoW training data of 250 MRA and CTA and their annotations have been released in a public Zenodo repository at `https://zenodo.org/records/15692630`.

2.3. Inter-Rater Agreement

To estimate the variability of the annotations used for benchmarking and the upper bounds of algorithmic performance, we analyzed inter-rater agreement in two aspects:

CoW Variant Classification Agreement. Senior neurosurgeon P.B. labeled the AV and PV classes for 40 CTA patients from the TopCoW test set, using the images in a dedicated 2-hour session. Labels from P.B. were compared with the annotations used in the benchmark. The selected cases covered all available CoW variants in the internal test data, including 4 AV classes and 6 PV classes, as shown in Fig. 1d. Balanced accuracies between the raters were 88% for AV and 78% for PV. Cohen's Kappa scores were 83% for AV and 72% for PV, suggesting good agreement.

Voxel-Level Segmentation Agreement. Voxel-level annotations were done on a subset of 5 patients from the TopCoW test set by the two manual annotators (K.Y. and F.M.). These five patients were selected because they each contained most or all of the CoW multiclass labels. The CoW anatomical annotations from both annotators were evaluated for Dice scores. Many CoW component classes had Dice scores of around 90% or above, while R-Pcom, L-Pcom, Acom, and 3rd-A2 had slightly lower Dice at 76-89%. Detailed results on the voxel-level segmentation agreement can be found in Supplementary S6.

2.4. External Multi-Center Test Data

In addition to the internal test sets from the TopCoW dataset, we gathered and annotated 86 MRA and CTA scans from four external multi-center test datasets for evaluation on the robustness of the algorithms. These external test datasets were from

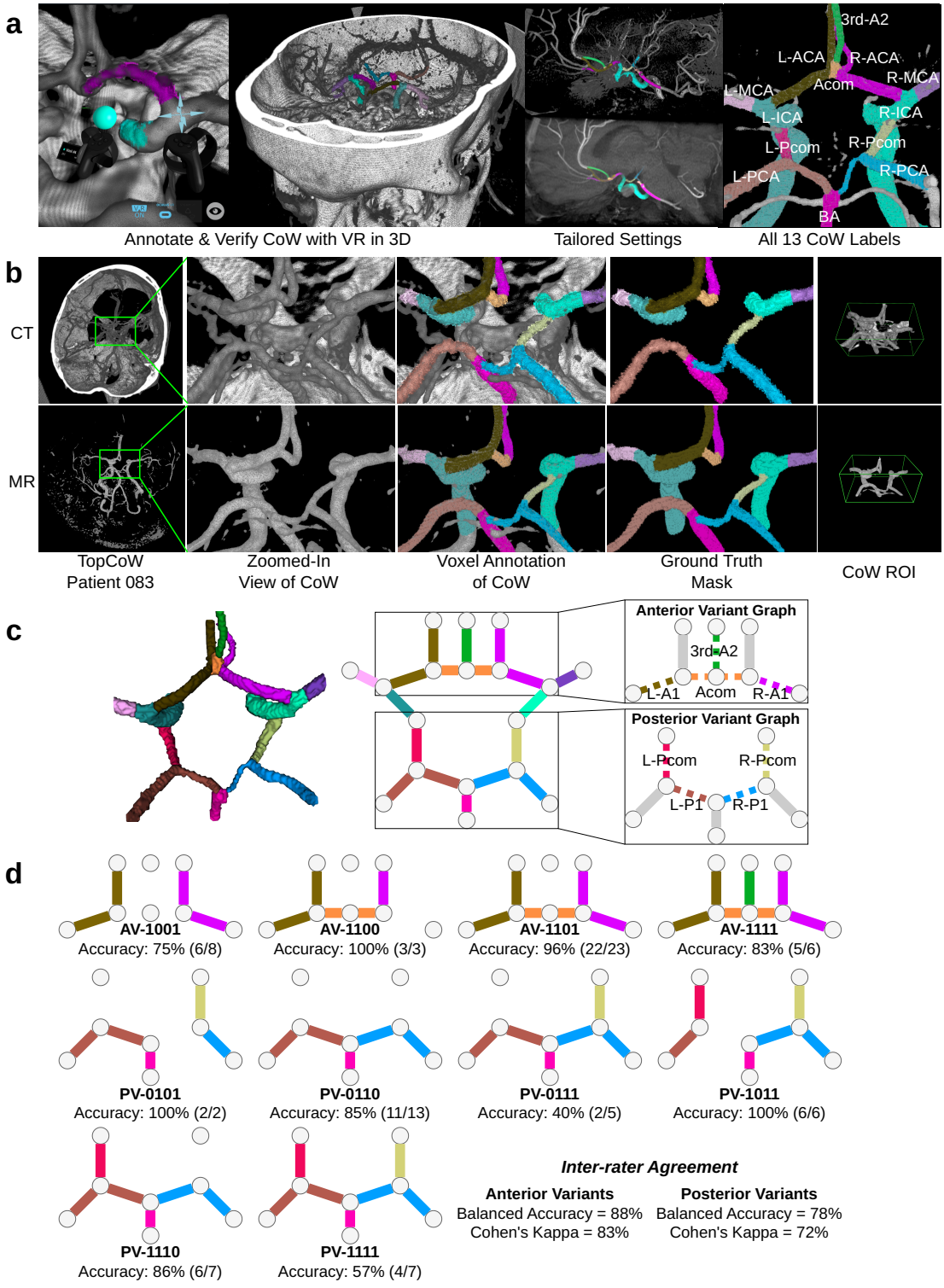

Fig. 1. CoW annotation. (a) Left shows data annotation using VR. Middle shows tailored settings by adjusting opacity, threshold, and window dynamically for suitable visualization during annotation. The right image shows the 13 anatomical labels for the CoW anatomy. (b) TopCoW dataset has paired modalities, CTA and MRA, from the same patient. Voxel annotations of the CoW vessels and a 3D bounding box of the CoW ROI are labelled for each modality. (c) CoW variant graph annotation was converted from the CoW segmentation mask. Each CoW can be classified by an anterior variant (AV) and a posterior variant (PV) graph. Both AV and PV are identified by a four-edge graph, with 0 being absent and 1 being present in the edge-list. (d) Inter-rater agreement for CoW variant classification on 40 TopCoW CTA test cases. Accuracy for each variant is shown along with the balanced accuracy and Cohen's Kappa score.

existing public datasets without CoW annotations. Two external CTA test sets were from the TUM University Hospital in Germany of the public ISLES'24 challenge training set (ISLES) [26, 27] and various hospitals in China of the public

Large IA Segmentation dataset (LargeIA) [28, 29]. Two external MRA test sets were from the Lausanne University Hospital in Switzerland of a public OpenNeuro dataset (Lausanne) [30, 31] and the Hammersmith Hospital in UK of the public IXI dataset (IXI-HH) [12]. The inclusion criteria were similar to that of the TopCoW dataset. For ISLES, we chose 26 CTA patients whose CoW were not occluded within the ROI. For LargeIA dataset, we chose 20 CTAs that do not have aneurysms inside the CoW ROI. For Lausanne and IXI-HH datasets, we chose 20 MRAs each from the healthy control group. All external datasets were annotated in the same fashion as the TopCoW dataset for the CoW benchmark.

Fig. 2 shows the statistical summary of the image information of the training, internal test, and external test data within the ROI. We compared the voxel dimension, entropy, and image intensity. TopCoW data had similar training and test distribution. ISLES and LargeIA datasets had much thinner slice thickness than TopCoW CTA. Lausanne and IXI-HH datasets had much bigger pixel spacing in the X-Y dimension compare with TopCoW MRA. Voxel dimensions were quite different among the external datasets. IXI-HH had a marked lower entropy, which may be due to its dated nature as the images were acquired from around 20 years ago. IXI-HH also had a very different mean intensity distribution compared to other datasets, with many MR images having ultra-high intensity values. Overall, the TopCoW internal test images were in-distribution while the external test datasets were out-of-distribution, which is useful for evaluating generalizability.

Distributions of the CoW variants in all our datasets are shown in Supplementary S7.

The multi-center test sets of 86 images with our CoW annotations used for the benchmark can be accessed from a public Zenodo repository at `https://zenodo.org/records/15692630`.

2.5. Algorithm Submission

Our challenge had two tracks for algorithm submissions, namely a CTA track and an MRA track. The main tasks were to multiclass segment the anatomical components of the CoW and to classify the CoW variant graph. The input to the algorithm was the whole 3D image volume, and the evaluation was conducted within the CoW ROI. For all tasks, the input to the submitted algorithm was initially intended to be a pair of CTA and MRA images from a patient due to the paired-modality feature of the TopCoW dataset. Algorithms that only needed one of the modalities could simply ignore the other modality input. In practice, all of the participating algorithms worked with single-modality input, and thus the submissions were also able to be evaluated with the single-modality external test datasets.

The submitted algorithms must be fully-automatic in the form of isolated Docker containers. For internal test data, the Docker containers were run in the cloud on the submission platform that provided an Nvidia T4 GPU with 16GB GPU memory. Submitted algorithms were limited to a runtime of 12-15 minutes per test case for inference on the cloud. Each team was given only one opportunity to upload their containers for the hidden test set. For external test sets, the Docker containers were run locally on a laptop with an RTX 3080 GPU with 16GB GPU memory.

2.6. Evaluation of Algorithms

Voxel-Level Multiclass Segmentation. For voxel-level metrics, the multiclass CoW segmentation predictions were evaluated using Dice similarity coefficient (Dice score), centerline Dice (clDice) [32], Hausdorff distance at 95% percentile, and connected component (zero-th Betti number) error.

More details on the TopCoW tasks and the evaluation metrics are in Supplementary S8.

Beyond Segmentation I: Key CoW Component Detection. The first beyond segmentation metric was the average F1 score, which is the harmonic mean of the precision and recall, for detection of Acom, Pcoms, and 3rd-A2. Positive detection was defined as at least 25% intersection over union between the predicted and ground truth masks.

Beyond Segmentation II: CoW Variant Classification. The second beyond segmentation metric was variant balanced accuracy (VarBalAcc) for CoW variant graph classification. The VarBalAcc was calculated for both anterior and posterior variants. The variant class was determined using the AV and PV edge-list of the the variant graph, as shown in Fig. 1c. The segmentation mask was converted to edge-list based on presence of the Acom, Pcoms, and 3rd-A2 labels, and whether ACA and PCA were connected to the relevant neighbour labels of ICA and BA for A1 and P1 edges.

The separate CoW variant classification task required the algorithms to output the variant graph directly instead of a segmentation mask. The CoW variant classification task shared the same evaluation metric as the second beyond segmentation metric, which was the VarBalAcc for both anterior and posterior variants.

Ranking. The algorithms were evaluated on the internal test sets using all the metrics, and their rank positions for each metric were averaged to reach a ranking for the leaderboards. To evaluate the ranking stability, we also created 10 bootstraps of the internal test sets and calculated the rankings on the bootstrapped test sets. We refer to the top performing teams on the averaged rank for CTA or MRA tracks as "top teams". The top teams were further evaluated on the external multi-center test sets for generalizability. Details on the ranking analysis can be found in Supplementary S9.

Clinical Application I: Fetal PCA Classification. We extracted diameters and centerline of CoW segmentation masks using the workflow from [33], with Supplementary S10 to provide more details. The diameters along the Pcom and P1 segments were used to determine the CoW anatomical variant called the fetal PCA variant [1]. We compared the diameters of the Pcom and P1 at the 25% percentile. If the Pcom was slightly larger in diameter ($\geq$ 1.05x the diameter of P1), the CoW was classified as having a fetal PCA variant type. Fetal PCA was assessed separately for the left and right sides. The

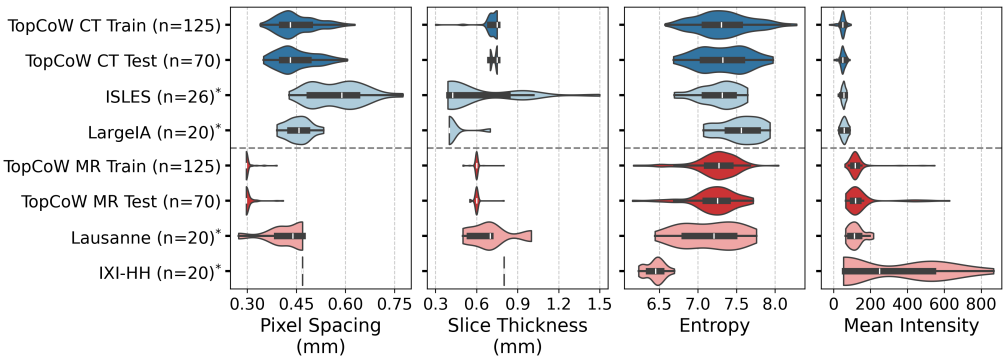

Fig. 2. Data characteristics of the TopCoW training, internal test, and external multi-center test datasets. CTA datasets are in blue violin plots, and MRA datasets are in red. Datasets were compared in terms of pixel spacing, slice thickness, entropy, and mean intensity inside the ROI. One extreme outlier in mean intensity was removed from TopCoW MRA training set for visualization purposes. 'n' is number of cases. '*' indicates external test datasets.

same set of 40 TopCoW CTA cases used in inter-rater agreement for variant classification were also labeled for fetal PCA class by the senior neurosurgeon (P.B.), who labeled the fetal PCA by visually inspecting the images. The fetal PCA labels from P.B. were treated as ground truth.

Clinical Application II: Locating Aneurysm. We selected 12 patients with intracranial aneurysms along their CoW vessels from the aforementioned external LargeIA dataset, which included aneurysm ground truth annotations. The aneurysm locations were then reviewed and labeled by a senior neurosurgeon (Y.M.). CoW segmentation algorithms were applied to the images of the aneurysm patients, and the resulting CoW predictions were overlaid with the provided aneurysm ground truth. The aneurysm's location was determined by identifying the CoW labels adjacent to or overlapping with the aneurysm mask.

3. Results

3.1. Progression of Submissions

Our two iterations of the TopCoW challenge received more than 250 registrations from participants from six continents. Over 25 teams made submissions to the benchmark. Notably, five teams participated in both years. As shown in Fig. 3a, by comparing the performance of the five teams on the common 34 patients present in both years' test sets, we observed a marked improvement in their results, especially for the CTA modality. While an increase in our training data might have contributed to the performance improvement, two findings from Fig. 3a revealed that the progress came more from the innovation of algorithms. Firstly, team 'DKFZ' was the only team that did not make any major changes to their algorithms, and the increased training data had little effect on the performance change, especially on the CTA modality. Secondly, teams like 'UZH', 'NIC-VICOROB', and 'junqiangchen' that made major design changes to their algorithms resulted in more drastic improvement. Thus, the performance improvement was largely propelled by new methodological breakthroughs.

Fig. 3b summarizes the key design choices of the segmentation algorithms from 2024 top teams. All top teams used single-modality input even though TopCoW challenge provided a pair of test images from both modalities to the algorithms as input for the internal test sets. All top teams trained with both modalities in a mixed modality training pool, thus making the algorithm modality-agnostic. Only one team, 'CLAIM', used additional training data prepared independently, which included some external MRA test images without our ground truth labels. There were a mix of strategies for number of stages used in the pipeline: More than half of the top teams went for a two-stage approach, such as with first a localization stage to crop the CoW ROI followed by a segmentation stage on the zoomed-in ROI. All based their network architecture on nnUNet [34]. All used cross-entropy (CE) and Dice loss. All but one used topology-based loss. All but one employed topological optimizations in their methods. The topological optimizations came in three aspects: centerline or skeleton; connected components (CCs); relation among the labels such as neighborhood adjacency. Most top teams considered at least one aspect of the topological optimizations.

Two methodological breakthroughs stood out between 2023 and 2024: significantly more teams trained their algorithms using mixed modality and topological optimizations. As shown in Fig. 3c, only one team ('DKFZ') trained their model with mixed modalities in 2023. The same team stood out by winning most benchmark tasks, inspiring seven other teams to adopt mixed-modality training in 2024. Only two teams ('DKFZ' and 'HITSZ') employed any topological optimizations in 2023. These two teams achieved good performance, and seven additional teams followed suit by incorporating topological optimizations into their models in 2024.

For a full description of the submitted algorithms and their teams, please refer to Supplementary materials S11.

3.2. Voxel-Level Multiclass Segmentation

We show the qualitative segmentation results for TopCoW internal test sets from one of the top teams in Fig. 4a. Two patients for each modality were selected, representing a wide range of CoW variants and class-average Dice scores. Predictions were able to segment various complex CoW anatomies accurately, capturing the curvatures of various vessels and boundaries between classes. Patient 116 from TopCoW MR had a

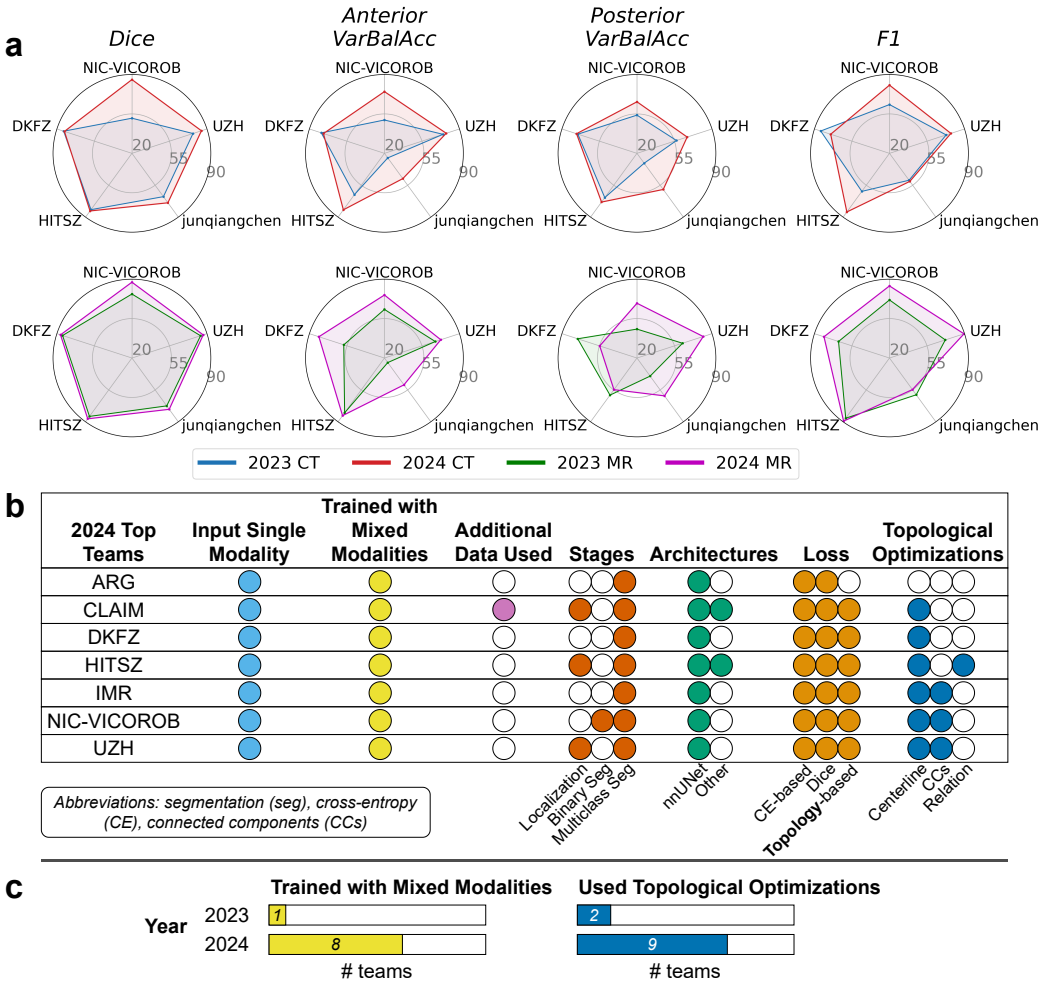

Fig. 3. Progression of TopCoW submissions and winning strategies. (a) Performance of the five teams that participated in both years on the same 34 patients in both years' test sets. The metrics shown are class-average Dice, variant-balanced accuracy (VarBalAcc) of the anterior and posterior CoW variant classifications, and average F1 score for detection of Acom, Pcoms, and 3rd-A2. Upper rows are for the CTA track. Bottom rows are for the MRA track. (b) Key characteristics of the segmentation algorithms from the top teams in alphabetical team name order. (c) Two methodological breakthroughs in 2023 got picked up by many more teams in the following year.

lower class-average Dice score of 79% because Acom was a false-positive detection, resulting in a 0% Dice for that label. For TopCoW internal test sets, top teams had a median class-average Dice of around 90% for both CTA and MRA.

Fig. 4b shows the segmentation performance in class-average Dice by the top 6 teams on both internal and external test sets. The top teams were able to generalize to external test sets for both modalities, with above 80% median Dice for all test sets. The results for other voxel-level metrics also showed good performance and similar generalization pattern as the Dice metric. Detailed results of all the 2024 teams for both internal and external test datasets can be found in Supplementary S12 and S13. We also report the inference time per test image for best algorithms on the external test set in Supplementary S14.

3.3. Beyond Segmentation I: Key CoW Component Detection

The presence and absence of four CoW components directly determine many of the CoW variant types. The four key CoW components are the communicating arteries (Acom, R-Pcom and L-Pcom) and the 3rd-A2 segment. Correct detection of these components is a desirable and clinically relevant feature of segmentation algorithms. We investigated the detection performance of the aforementioned four key CoW components for all test sets in Fig. 5a. F1 scores, which are the mean of precision and recall, from the top 6 teams of each modality were shown in boxplots. The detection of the communicating arteries were consistent across test sets and modalities, and were able to be detected at above 75% F1 scores by top teams for most of the test sets. The 3rd-A2 had poorer detection performance, with CTA datasets above 50% and MRA datasets above 65% F1 scores for most teams. The 3rd-A2 is a rarer vessel component, and its smaller sample size in CTA test sets (6/70, 1/26, and 2/20 occurrences) and in MRA test sets (7/70, 2/20, and 2/20 occurrences) likely contributed to the less consistent detection results.

3.4. Beyond Segmentation II: CoW Variant Classification

CoW variant classification has long been a highly demanded task driven by its great clinical potential. In fact, concurrent to the first iteration of the TopCoW challenge, another CoW-

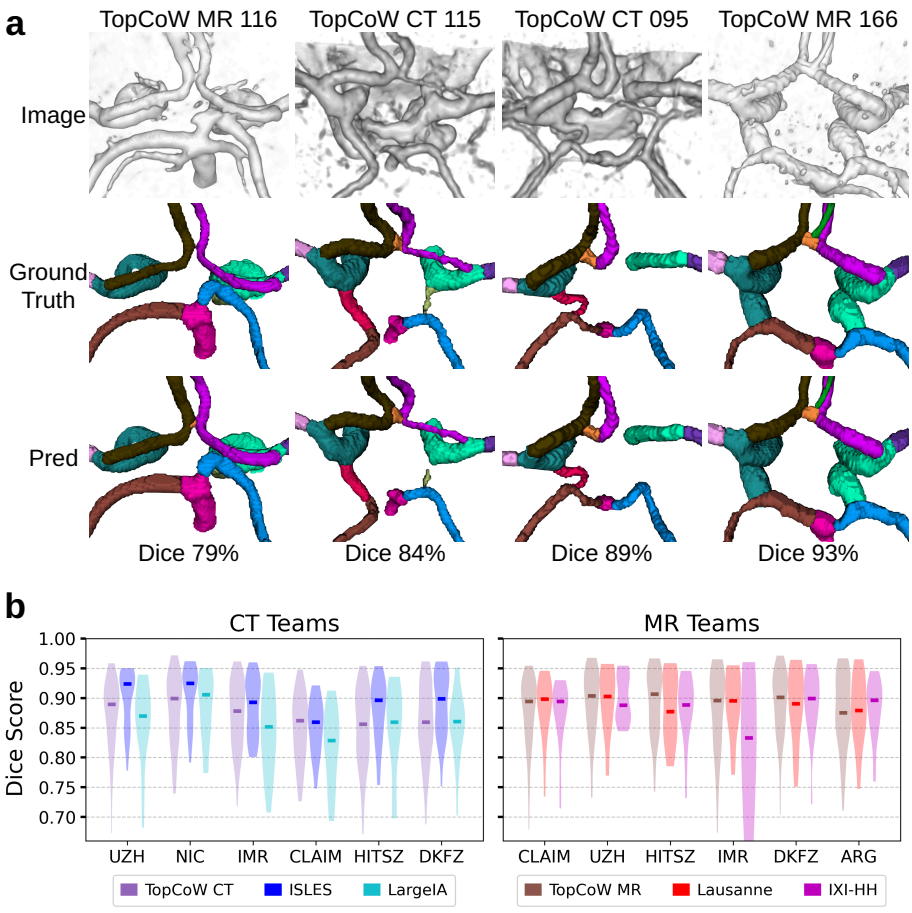

Fig. 4. Voxel-level multiclass segmentation performance. (a) Qualitative results for multiclass segmentation task. The ground truth was compared with predictions from team 'UZH' on four selected patients from the internal test data. Note the increasing class-average Dice scores for the selected four cases. (b) Class-average Dice scores from the top 6 teams on CTA and MRA internal and external test sets. Team 'NIC-VICOROB' is abbreviated as 'NIC'. Team 'CLAIM' used additional training data which included some external MRA test images from Lausanne and IXI-HH without our ground truth labels. Team 'IMR' had five cases in IXI-HH with Dice below the shown range. Charts showing violin plots with the middle bar being the median.

related challenge, the CROWN challenge [21], was held in 2023 MICCAI with a main task on CoW variant classification. Inspired by the CROWN challenge, in 2024 our second iteration, we created a classification task where participants submit algorithms that output CoW variant classes. Interestingly, the best performing algorithms for the classification task came from the segmentation algorithms with an extra post-processing step to convert the segmentation masks to the CoW variant edge-lists. Notably, four teams took part in both the segmentation task and the classification task with segmentation-based and classification-based algorithms respectively. Fig. 5b shows that the segmentation-based algorithms out-performed the classification-based methods by a factor of at least 2x on the internal test sets. The classification-based approaches listed here had explored various strategies ranging from graph learning to self-attention, but they all trained the algorithms as a classification model that optimized for the variant classifier. On the other hand, the segmentation-based approaches focused sorely on the multiclass CoW segmentation. The performance gap between segmentation-based and classification-based algorithms prompted us to focus on the top segmentation algorithms for the external test sets.

Fig. 5c shows the VarBalAcc for both anterior and posterior variants for all test sets from the top 6 segmentation teams. The top teams were able to generalize well to external MRA datasets: On average, both the anterior and posterior VarBalAcc were at around 80% for MRA test sets. CTA datasets had good generalizability except for posterior variants for the LargeIA dataset. This is likely because two PV classes in LargeIA had small sample sizes comprising difficult cases: PV-1011 had only one case and PV-1110 had only two cases (see supplementary Table S3), and they were wrongly classified by most teams, resulting in low a balanced accuracy for posterior variants. Apart from the posterior VarBalAcc performance from LargeIA, CTA datasets also had fairly good classification performance and generalizability with above 70% VarBalAcc on average.

3.5. Clinical Application I: Fetal PCA Classification

One potential clinical use of the CoW segmentation model is to classify fetal PCA variant. Fetal PCA variant plays important roles in surgical planning and interpretation of perfusion imaging [1]. Neurosurgeons and neuroradiologists routinely classify patients as having fetal PCA or not based on CTA and MRA

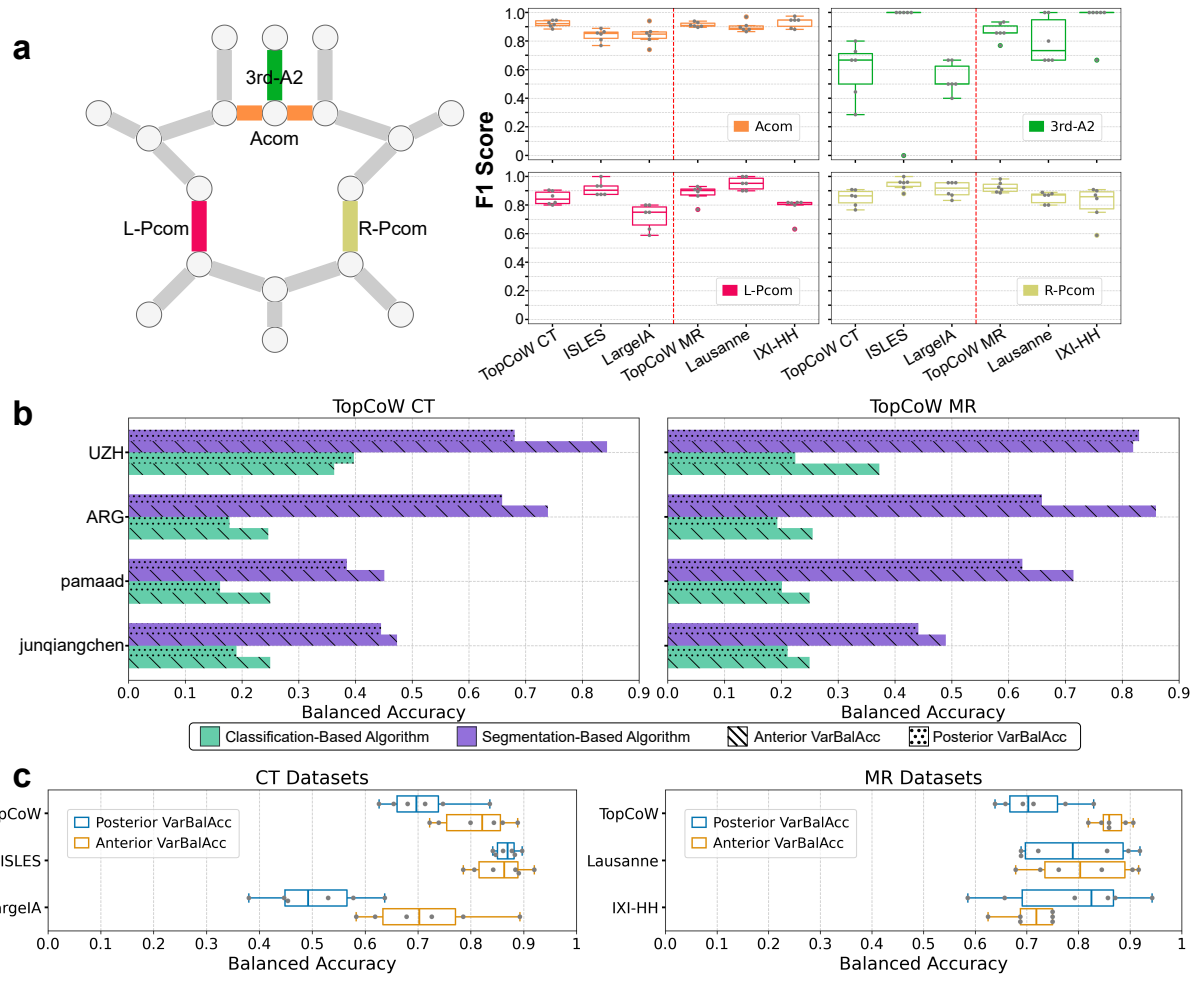

Fig. 5. Beyond segmentation performance. (a) Detection of key CoW components for all test datasets by the top 6 teams from each modality. F1 scores from the six teams were displayed in box plots. (b) Four teams submitted both classification-based and segmentation-based algorithms for classifying the CoW variants. Classification-based algorithms were trained to predict CoW variant classes directly. Segmentation-based algorithms were purely for CoW segmentation and later evaluated for CoW variant classification performance. (c) CoW variant classification performance from the top 6 segmentation teams on internal and external test sets. Anterior and posterior VarBalAcc scores from the top 6 teams were displayed in box plots.

findings. Here, we evaluated the clinical applicability of CoW segmentation algorithms for fetal PCA classification. As shown in Fig. 6a, diameters along the ipsilateral Pcom and P1 segmentation masks, if any, were used to predict the fetal PCA class. This simple heuristic allowed us to convert segmentation masks by top teams into fetal PCA labels. Fig. 6b shows that the segmentation output from the top teams were able to accurately classify fetal PCA variants with around 80% and above precision and recall for both fetal L-PCA and fetal R-PCA.

3.6. Clinical Application II: Locating Aneurysm

Another potential important clinical use is locating the aneurysms with the CoW vessel segments as reference. Clinicians routinely need to locate intracranial aneurysms relative to the CoW anatomy in MRA and CTA images, as aneurysms occur most frequently along the CoW vessels. Here, we evaluated the ability of our CoW segmentation models to help locate intracranial aneurysms automatically. 12 aneurysm patients with various locations of their aneurysms were segmented by the top 4 teams. Fig.6c shows the aneurysm ground truth overlaid with the CoW segmentation predictions from team 'UZH' for four representative patients. CoW vessel labels that overlapped with or were adjacent to the aneurysm were used to describe the location of the aneurysm. Team 'UZH' were able to correctly locate 12/12 aneurysm patients in relation to the CoW vessel segments. Team 'NIC-VICOROB' and 'CLAIM' correctly located 11/12 patients and team 'IMR' located 10/12 patients, although the two mistakes were only minor and explainable. One common mistake was for an aneurysm from patient Tr0004 which was located on R-ACA but was very near where left and right ACAs touched, causing the location of this aneurysm to be wrongly predicted to be adjacent to both ACAs. Another mistake was for patient Tr0019 who had an aneurysm on the ICA from where Pcom tends to originate, and thus a part of the aneurysm was wrongly predicted as Pcom. Overall the CoW segmentation predictions from the top teams were robust against the presence of large aneurysms, and could be applied to locate the aneurysm for most of the patients. We highlight that these results also showed the robustness and generalizablity of the best CoW algorithms when there were no aneurysm cases

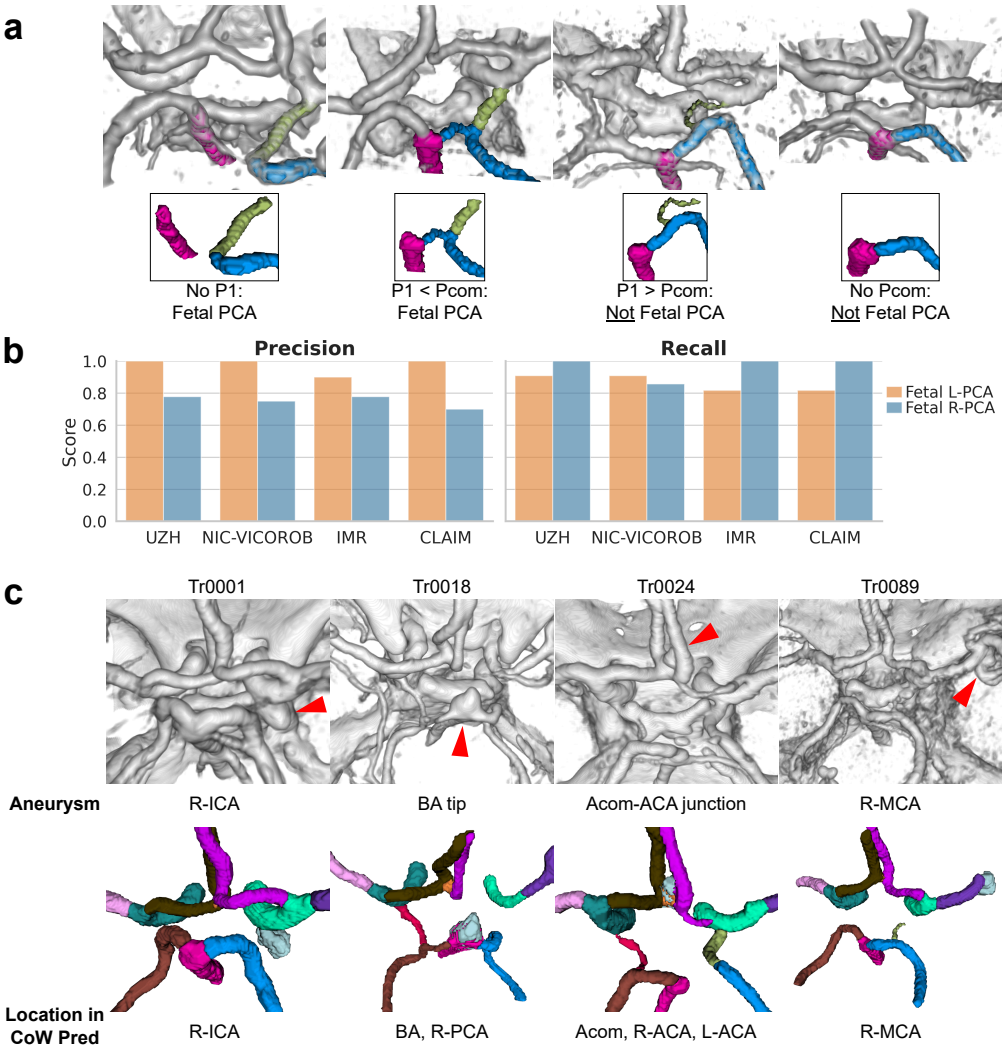

Fig. 6. Two clinical applications of CoW segmentation. (a) Segmentation masks of the ipsilateral Pcom and P1 segments were used to determine whether a fetal PCA class was present on that side. Upper row shows CTA images overlaid with segmentation masks related to fetal PCA. Bottom row shows zoomed-in view of the segmentation masks. Examples shown illustrate the fetal R-PCA classification. (b) Fetal PCA classification performance in terms of precision and recall scores from the top 4 teams. (c) Upper row shows CTA images with aneurysms indicated by red arrowheads. Bottom row shows the aneurysm masks in silver color overlaid with predicted CoW segmentation masks from team 'UZH'. The "Location in CoW Pred" listed the predicted CoW labels that the aneurysm overlapped with or was adjacent to in the overlay.

in the training data and the aneurysm patients came from an external test data. Detailed results for locating the 12 aneurysm patients can be found in Supplementary S15.

4. Discussions

4.1. Benefits of Mixed Modality Training for CTA

Training with mixed modalities was an effective strategy, particularly for the CTA modality. This strategy was pioneered by team 'DKFZ' in the first iteration of our challenge, and was picked up by all top teams in the second iteration. The improvement was especially obvious for the more difficult modality CTA. Between MRA and CTA modalities, CTA tended to have lower metric scores even when both modalities had the same number of training data. This could be due to the extra veins and bones surrounding the CoW not seen in MRA, and also the less detailed brain soft tissues in the background. CTA benefited much more from the mixed modality training strategy with bigger improvement in performance than MRA. Given that the annotation efforts were also lighter in MRA images, investing in an easier-to-annotate modality such as MRA can be a good way to reduce the annotation burden and to boost the performance for CTA modality. This is a key finding that can help solve other CTA segmentation tasks in the future.

4.2. Importance of Topological Optimizations

Topological optimizations enabled CoW segmentation algorithms to be used in topology-dependent downstream clinical tasks, such as CoW variant classification and fetal PCA classification. These downstream applications require the segmented vessels to capture key topological properties of the underlying anatomy such as centerline, connected components, and adjacency relation. A wide range of topological optimizations were used by the top submissions. Two of these methods were newly

developed by two teams while taking part in the first iteration of our challenge, with our challenge being one of the first venues to test them. Both methods were centerline-based loss functions that improve vessel segmentation, and they were the "skeleton recall (SkelRecall) loss" [35] by team 'DKFZ' and "centerline boundary Dice (cbDice) loss" [36] by team 'HITSZ'. SkelRecall loss was subsequently picked up and used by three other top teams, 'CLAIM', 'NIC-VICOROB', and 'UZH' in the second iteration in 2024. Topology optimization techniques from other medical scenarios were used as well: Team 'IMR' built upon a loss previously designed to optimize topology for lung airway segmentation [37]. Top teams also optimized other aspects of topology such as joining disconnected components, removing small isolated components, and handling adjacency relation. Collectively, the various topological optimizations allowed the CoW segmentation to have improved connectivity and centerline, and more accurate topology in general, to be effectively applied to topology-dependent downstream tasks like the CoW variant classification and fetal PCA classification.

4.3. Comparison with CROWN Results

Since both the CROWN and TopCoW challenges included a CoW variant classification task, we compare the performance of our algorithms with those reported in the CROWN challenge. Based on the merged common set of variants, the top two teams from the CROWN challenge achieved 24-30% anterior and 28-50% posterior balanced accuracy, whereas the top two teams from the TopCoW MR submissions achieved 82-89% anterior and 75-82% posterior balanced accuracy. The main difference was that the CROWN challenge did not provide any CoW segmentation annotations and the task was formulated purely as an image classification problem. Similar observations were made in our challenge in Fig. 5b, where we found the segmentation-based algorithms that focused on the CoW multiclass segmentation task performed much better on the variant classification task than classification-based algorithms that optimized for the classification task. Solving the CoW multiclass segmentation task can lead to better solutions for downstream tasks like CoW variant classification.

4.4. Explainability via Segmentation

CoW segmentation algorithms not only can solve downstream tasks, but also provide explainability to the solutions—a feature that is important in clinical settings. We showed that the best CoW segmentation algorithms could be used to effectively detect key CoW components, classify CoW variants and fetal-type PCA, and locate intracranial aneurysms. But more importantly, the intermediate steps to transform the segmentation masks into relevant downstream outputs were fully transparent and interpretable, in contrast to black-box approaches. When clinicians need a "confidence" score on why a certain prediction was made for the detection, classification, or localization by the model, they can easily interpret and explain the results by simply inspecting the CoW segmentation prediction.

4.5. VR to Handle Complex Anatomy

As one of the first challenges to use VR generated annotations at-scale, we believe this challenge has successfully shown that VR-based annotation/verification workflow can overcome the otherwise too time-consuming annotation process for a complex multiclass anatomical segmentation problem. The depth dimension as viewed in 3D in VR offered efficient and powerful annotation/verification capabilities, which proved to be uniquely suitable for curvilinear structures like the CoW vessels that can have complicated spatial orientations and relations among the multiclass tortuous vessels. VR enabled us to quickly and accurately produce annotations and check predictions for even complex CoW anatomies and rare variants. This allowed us to prepare such a densely-annotated large-scale dataset that covered many clinically relevant CoW anatomies.

4.6. Limitations and Future Work

Due to the heterogeneity of the CoW anatomy, not all CoW variants were included in our annotation scheme. In future, we can expand the multiclass labels to accommodate other rare CoW variants excluded in our dataset, which will make the trained model applicable and more robust to a broader range of the population.

Our existing CoW variant graph can be further sub-divided into more fine-grained variant types involving hypo-plasticity and vessel diameters, such as whether the A1 segment of ACA or P1 segment of PCA is hypoplastic [38]. This can be done by applying the same workflow used in our fetal PCA classification where we extracted the diameters along the centerlines of P1 and Pcom.

5. Conclusion

The two iterations of the TopCoW challenge attracted over 250 registered participants from six continents, which resulted in over 25 submitted algorithms. We evaluated the performance on internal and external test sets from multi-centers. The top segmentation algorithms showed good performance on multiclass CoW segmentation, detection of key CoW components and classification of CoW variants, and generalized well to new test data. We conducted additional evaluations on the ability of the segmentation model to classify fetal PCA and locate intracranial aneurysms, with results showing promising potential for clinical applications.

TopCoW has demonstrated the power, potential, and versatility of CoW multiclass segmentation for a wide range of tasks beyond segmentation. TopCoW released the first dataset on paired CTA and MRA with annotations, and thus enabled some of the first anatomical segmentation models for CoW, especially for the CTA modality. TopCoW led to several methodological insights on how to best solve the CoW anatomical segmentation and variant classification task, and it catalyzed a few algorithmic breakthroughs from the participants. TopCoW was the first challenge to use a VR-based annotation workflow, which was crucial in preparing the multiclass annotations. As a first benchmark for such a CoW segmentation task, TopCoW gathered strong baseline results for further algorithm development

and comparison. The annotated datasets and the best performing Docker submissions have been released in Zenodo records for public access.

Ultimately we want to solve real clinical problems, and one of them is to prototype an automated CoW characterization tool for diagnosis, screening, and treatment. An accurate characterization of the CoW is of great clinical relevance, and we hope TopCoW challenge has piqued the interest of the community on this worthwhile endeavor.

Data Availability

The TopCoW training data of 250 annotated images and our multi-center test sets of 86 annotated images are released in our public Zenodo repository (15 GB) at `https://zenodo.org/records/15692630`.

Code Availability

The Docker images from best performing teams and the scripts to help run them locally are released in our public Zenodo repository (45 GB) at `https://zenodo.org/records/15665435`.

For code availability of individual team submission, please see Supplementary S11.

The implementation of our evaluation metric code is open sourced at `https://github.com/CoWBenchmark/TopCoW_Eval_Metrics`.

Acknowledgments

The challenge is supported by the Digitalization Initiative of the Zurich Higher Education Institutions (DIZH) and the Helmut Horten Foundation. We thank Hrvoje Bogunović for helpful discussions and suggestions during the early planning stages. We also thank Nathan Spencer and Michael Morehead from syGlass for the technical assistance for the VR setup, and James Meakin and Chris van Run from grand-challenge.org for the technical support for the challenge infrastructure.

Ynte Ruigrok has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 852173). Hakim Baazaoui received funding from the Koetser Foundation and the "Young Talents in Clinical Research" program of the SAMS and of the G. & J. Bangerter-Rhyner Foundation.

Team **2i_mtl** was supported by Grants from the Quebec BioImaging Network (Project No. 21.24) and start-up funds from the Centre de Recherche du CHUM and Departement de radiologie, radio-oncologie et medecine nucleaire, Universite de Montreal/Bayer. Laurent Letourneau-Guillon is supported by a Clinical Research Scholarship-Junior 1 Salary Award (311203) from the Fonds de Recherche du Quebec en Sante and Fondation de l'Association des Radiologistes du Quebec. Team **CLAIM** acknowledges funding from the German Federal Ministry of Education and Research (ANONYMED Project, coordinator DF). Computation has been performed on the HPC for Research cluster of the Berlin Institute of Health. They also acknowledge the contribution of MRCLEAN investigators by providing access to data from the MRCLEAN trial. Team **DKFZ** was supported by the Helmholtz Association under the joint research school "HIDSS4Health - Helmholtz Information and Data Science School for Health" and part of their work was funded by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science. Team **EURECOM** was partially funded by the French government, through the 3IA Cote d'Azur Investments in the Future project managed by the ANR (ANR-19-P3IA-0002) and by the ANR JCJC project I-VESSEG (22-CE45-0015-01). Team **IMR** was supported in part by National Key R&D Program of China (Grant Number: 2022ZD0212400), Natural Science Foundation of China (Grant Number: 62373243) and the Science and Technology Commission of Shanghai Municipality, China (Grant Number: 20DZ2220400), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0102). Team **IWantToGoToCanada** was supported by the National Research Foundation (NRF-2020M3E5D2A01084892), Institute for Basic Science (IBS-R015-D1), ITRC support program (IITP-2023-2018-0-01798), AI Graduate School Support Program (2019-0-00421), ICT Creative Consilience program (IITP-2023-2020-0-01821), and the Artificial Intelligence Innovation Hub program (2021-0-02068). Team **IWM** wants to acknowledge the Polish HPC infrastructure PLGrid support (No. PLG/2023/016239). Team **NantesU** was partially supported by the French ANR project "eCAN" and INSERM CoPoC #MAT-PI-22155-A-01 (RVF23037NSA). Team **NIC-VICOROB** was supported by the Ministerio de Ciencia e Innovacion (DPI2020-114769RB-I00) as well as by ICREA under the ICREA Academia programme, and also partly supported the Ministerio de Ciencia e Innovacion (DPI2020-114769RB-I00). Members of the 2024 team received the support from the PID2020-114769RBI00 and the PID2023-146187OB-I00 projects funded by the Ministerio de Ciencia, Innovación y Universidades. Team **Pamaad** P. Casademunt is supported by the European Union's Horizon 2020 grant agreement No.101136438 (GEMINI project), by the Agència de Gestió d'Ajuts Universitaris i de Recerca (grant No. 2024 FI-1 00419), and the Maria de Maeztu grant of excellence. A. Galdran is supported by grant RYC2022-037144-I, funded by MCIN/AEI/10.13039/501100011033 and by FSE+. They would like to thank the HPC team from ZHAW, particularly Pascal Häussler and Stefan Weber, for their generous allocation of computational resources and technical assistance. Team **UB-VTL** wants to acknowledge the computational resources provided by the Center of Computational Research (CCR) at University of Buffalo. Team **UW** was supported by the United States National Institute of Health (grants R01HL162743 and R00HL136883).

References

[1] A. G. Osborn, Osborn's Brain: Imaging, Pathology, and Anatomy, Amirsys, 2013.
[2] D. S. Liebeskind, Collateral circulation, Stroke 34 (2003) 2279–2284.
[3] Y.-M. Chuang, L. Chan, Y.-J. Lai, K.-H. Kuo, Y.-H. Chiou, L.-W. Huang, Y.-T. Kwok, T.-H. Lai, S.-P. Lee, H.-M. Wu, et al., Configuration of the

circle of willis is associated with less symptomatic intracerebral hemorrhage in ischemic stroke patients treated with intravenous thrombolysis, Journal of Critical Care 28 (2013) 166–172.

[4] T. van Seeters, J. Hendrikse, G. J. Biessels, B. K. Velthuis, W. P. Mali, L. J. Kappelle, Y. van der Graaf, S. S. Group, Completeness of the circle of willis and risk of ischemic stroke in patients without cerebrovascular disease, Neuroradiology 57 (2015) 1247–1251.

[5] K. M. Kim, H.-S. Kang, W. J. Lee, Y. D. Cho, J. E. Kim, M. H. Han, Clinical significance of the circle of willis in intracranial atherosclerotic stenosis, Journal of Neurointerventional Surgery 8 (2016) 251–255.

[6] L. Rinaldo, B. A. McCutcheon, M. E. Murphy, M. Bydon, A. A. Rabinstein, G. Lanzino, Relationship of a1 segment hypoplasia to anterior communicating artery aneurysm morphology and risk factors for aneurysm formation, Journal of Neurosurgery 127 (2016) 89–95.

[7] F. Yang, H. Li, J. Wu, M. Li, X. Chen, P. Jiang, Z. Li, Y. Cao, S. Wang, Relationship of a1 segment hypoplasia with the radiologic and clinical outcomes of surgical clipping of anterior communicating artery aneurysms, World Neurosurgery 106 (2017) 806–812.

[8] P. V. Banga, A. Varga, C. Csobay-Novák, M. Kolossváry, E. Szántó, G. S. Oderich, L. Entz, P. Sótonyi, Incomplete circle of willis is associated with a higher incidence of neurologic events during carotid eversion endarterectomy without shunting, Journal of Vascular Surgery 68 (2018) 1764–1771.

[9] M. J. Krabbe-Hartkamp, J. Van der Grond, F. De Leeuw, J. C. de Groot, A. Algra, B. Hillen, M. Breteler, W. Mali, Circle of willis: morphologic variation on three-dimensional time-of-flight mr angiograms., Radiology 207 (1998) 103–111.

[10] S. Iqbal, A comprehensive study of the anatomical variations of the circle of willis in adult human brains, Journal of Clinical and Diagnostic Research: JCDR 7 (2013) 2423.

[11] E. Bullitt, D. Zeng, G. Gerig, S. Aylward, S. Joshi, J. K. Smith, W. Lin, M. G. Ewend, Vessel tortuosity and brain tumor malignancy: a blinded study, Academic Radiology 12 (2005) 1232–1240.

[12] IXI, IXI dataset - brain development, `https://brain-development.org/ixi-dataset/`, 2022. Accessed: 2022-09-30.

[13] CAS2023, Cerebral artery segmentation challenge (CAS) 2023, `https://codalab.lisn.upsaclay.fr/competitions/9804`, 2023. Accessed: 2023-10-01.

[14] S. Chatterjee, H. Mattern, M. Dörner, A. Sciarra, F. Dubost, H. Schnurre, R. Khatun, C.-C. Yu, T.-L. Hsieh, Y.-S. Tsai, et al., Smile-uhura challenge–small vessel segmentation at mesoscopic scale from ultra-high resolution 7t magnetic resonance angiograms, arXiv preprint arXiv:2411.09593 (2024).

[15] L. Mou, Q. Yan, J. Lin, Y. Zhao, Y. Liu, S. Ma, J. Zhang, W. Lv, T. Zhou, A. F. Frangi, et al., Costa: A multi-center tof-mra dataset and a style self-consistency network for cerebrovascular segmentation, IEEE transactions on medical imaging (2024).

[16] H. Bogunović, J. M. Pozo, R. Cárdenes, L. San Román, A. F. Frangi, Anatomical labeling of the circle of willis using maximum a posteriori probability estimation, IEEE Transactions on Medical Imaging 32 (2013) 1587–1599.

[17] D. Robben, S. Sunaert, V. Thijs, G. Wilms, F. Maes, P. Suetens, Anatomical labeling of the circle of willis using maximum a posteriori graph matching, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part I 16, Springer, 2013, pp. 566–573.

[18] D. Robben, E. Türetken, S. Sunaert, V. Thijs, G. Wilms, P. Fua, F. Maes, P. Suetens, Simultaneous segmentation and anatomical labeling of the cerebral vasculature, Medical Image Analysis 32 (2016) 201–215.

[19] L. Chen, T. Hatsukami, J.-N. Hwang, C. Yuan, Automated intracranial artery labeling using a graph neural network and hierarchical refinement, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI 23, Springer, 2020, pp. 76–85.

[20] S.-W. Hong, H.-N. Song, J.-U. Choi, H.-H. Cho, I.-Y. Baek, J.-E. Lee, Y.-C. Kim, D. Chung, J.-W. Chung, O.-Y. Bang, et al., Automated in-depth cerebral arterial labelling using cerebrovascular vasculature reframing and deep neural networks, Scientific Reports 13 (2023) 3255.

[21] I. N. Vos, Y. M. Ruigrok, E. Bennink, M. R. Velthuis, B. Paic, M. E. Ophelders, M. A. Buser, B. H. van der Velden, G. Chen, M. Coupet, et al., Evaluation of techniques for automated classification and artery quantification of the circle of willis on tof-mra images: The crown challenge, Medical Image Analysis (2025) 103650.

[22] F. Dumais, M. P. Caceres, F. Janelle, K. Seifeldine, N. Arès-Bruneau, J. Gutierrez, C. Bocti, K. Whittingstall, eicab: A novel deep learning pipeline for circle of willis multiclass segmentation and analysis, NeuroImage 260 (2022) 119425.

[23] A. Hilbert, J. Rieger, V. I. Madai, E. M. Akay, O. U. Aydin, J. Behland, A. A. Khalil, I. Galinovic, J. Sobesky, J. Fiebach, et al., Anatomical labeling of intracranial arteries with deep learning in patients with cerebrovascular disease, Frontiers in Neurology 13 (2022) 1000914.

[24] OpenData Swiss, Terms of use — OpenData Swiss, `https://opendata.swiss/en/terms-of-use`, 2024. [Online; accessed 1-March-2024].

[25] D. Kaltenecker, R. Al-Maskari, M. Negwer, L. Hoeher, F. Kofler, S. Zhao, M. Todorov, Z. Rong, J. C. Paetzold, B. Wiestler, M. Piraud, D. Rueckert, J. Geppert, P. Morigny, M. Rohm, B. H. Menze, S. Herzig, M. Berriel Diaz, A. Ertürk, Virtual reality-empowered deep-learning analysis of brain cells, Nature Methods (2024) 1–10.

[26] E. de la Rosa, R. Su, M. Reyes, R. Wiest, E. O. Riedel, F. Kofler, K. Yang, H. Baazaoui, D. Robben, S. Wegener, et al., Isles'24: Improving final infarct prediction in ischemic stroke using multimodal imaging and clinical data, arXiv preprint arXiv:2408.10966 (2024).

[27] E. O. Riedel, E. de la Rosa, T. A. Baran, M. H. Petzsche, H. Baazaoui, K. Yang, D. Robben, J. O. Seia, R. Wiest, M. Reyes, et al., Isles 2024: The first longitudinal multimodal multi-center real-world dataset in (sub-) acute stroke, arXiv preprint arXiv:2408.11142 (2024).

[28] Z.-H. Bo, Large ia segmentation dataset, 2021. URL: `https://doi.org/10.5281/zenodo.6801398`. doi:`10.5281/zenodo.6801398`.

[29] Z.-H. Bo, H. Qiao, C. Tian, Y. Guo, W. Li, T. Liang, D. Li, D. Liao, X. Zeng, L. Mei, et al., Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network, Patterns 2 (2021) 100197.

[30] T. D. Noto, G. Marie, S. Tourbier, Y. Alemán-Gómez, O. Esteban, G. Saliou, M. B. Cuadra, P. Hagmann, J. Richiardi, "lausanne tof-mra aneurysm cohort", 2022. doi:`doi:10.18112/openneuro.ds003949.v1.0.1`.

[31] T. Di Noto, G. Marie, S. Tourbier, Y. Alemán-Gómez, O. Esteban, G. Saliou, M. B. Cuadra, P. Hagmann, J. Richiardi, Towards automated brain aneurysm detection in tof-mra: open data, weak labels, and anatomical knowledge, Neuroinformatics 21 (2023) 21–34.

[32] S. Shit, J. C. Paetzold, A. Sekuboyina, I. Ezhov, A. Unger, A. Zhylka, J. P. Pluim, U. Bauer, B. H. Menze, clDice-a novel topology-preserving loss function for tubular structure segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16560–16569.

[33] F. Musio, K. Yang, S. Shit, C. Prabhakar, N. Juchler, B. Menze, S. Hirsch, Quantitative evaluation of the circle of willis vascular architecture in 3d ct and mr angiography, in: 8th International Conference on Computational and Mathematical Biomedical Engineering (CMBE24), Arlington, VA, USA, 24-26 June 2024, volume 2, Computational & Mathematical Biomedical Engineering, 2024, pp. 563–566.

[34] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, K. H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2021).

[35] Y. Kirchhoff, M. R. Rokuss, S. Roy, B. Kovacs, C. Ulrich, T. Wald, M. Zenk, P. Vollmuth, J. Kleesiek, F. Isensee, et al., Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures, in: European Conference on Computer Vision, Springer, 2024, pp. 218–234.

[36] P. Shi, J. Hu, Y. Yang, Z. Gao, W. Liu, T. Ma, Centerline boundary dice loss for vascular segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 46–56.

[37] M. Zhang, Y. Gu, Towards connectivity-aware pulmonary airway segmentation, IEEE Journal of Biomedical and Health Informatics 28 (2023) 321–332.

[38] L. P. Westphal, N. Lohaus, S. Winklhofer, C. Manzolini, U. Held, K. Steigmiller, J. M. Hamann, M. El Amki, T. Dobrocky, L. D. Panos, et al., Circle of willis variants and their association with outcome in patients with middle cerebral artery-m1-occlusion stroke, European Journal of Neurology 28 (2021) 3682–3691.

[39] X. Li, P. S. Morgan, J. Ashburner, J. Smith, C. Rorden, The first step

for neuroimaging data analysis: Dicom to nifti conversion, Journal of Neuroscience Methods 264 (2016) 47–56.

[40] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, et al., Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images, Radiology: Artificial Intelligence 5 (2023).

[41] N. Schimke, J. Hale, Quickshear defacing for neuroimages, in: Proceedings of the 2nd USENIX Conference on Health Security and Privacy, HealthSec'11, USENIX Association, USA, 2011, p. 11.

[42] F. Isensee, M. Schell, I. Pflueger, G. Brugnara, D. Bonekamp, U. Neuberger, A. Wick, H.-P. Schlemmer, S. Heiland, W. Wick, et al., Automated brain extraction of multisequence mri using artificial neural networks, Human Brain Mapping 40 (2019) 4952–4964.

[43] A. Bouthillier, H. R. Van Loveren, J. T. Keller, Segments of the internal carotid artery: a new classification, Neurosurgery 38 (1996) 425–433.

[44] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, A. Kirillov, Boundary iou: Improving object-centric image segmentation evaluation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 15334–15342.

[45] A. Reinke, M. D. Tizabi, M. Baumgartner, M. Eisenmann, D. Heckmann-Nötzel, A. E. Kavur, T. Rädsch, C. H. Sudre, L. Acion, M. Antonelli, et al., Understanding metric-related pitfalls in image analysis validation, Nature methods 21 (2024) 182–194.

[46] J. Meyer-Spradow, T. Ropinski, J. Mensmann, K. Hinrichs, Voreen: A rapid-prototyping environment for ray-casting-based volume visualizations, IEEE Computer Graphics and Applications 29 (2009) 6–13.

[47] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999 (2018).

[48] T. M. Consortium, Project monai, `https://doi.org/10.5281/zenodo.4323059`, 2020. Zenodo.

[49] M. J. Cardoso, W. Li, R. Brown, N. Ma, E. Kerfoot, Y. Wang, B. Murrey, A. Myronenko, C. Zhao, D. Yang, et al., Monai: An open-source framework for deep learning in healthcare, arXiv preprint arXiv:2211.02701 (2022).

[50] A. Celaya, B. Riviere, D. Fuentes, A generalized surface loss for reducing the hausdorff distance in medical imaging segmentation, arXiv preprint arXiv:2302.03868 (2023).

[51] G. Jocher, J. Qiu, A. Chaurasia, Ultralytics YOLO, 2023. URL: `https://github.com/ultralytics/ultralytics`.

[52] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).

[53] P. Dutta, S. Bose, S. K. Roy, S. Mitra, Are vision xlstm embedded unet more reliable in medical 3d image segmentation?, arXiv (2024). URL: `https://arxiv.org/abs/2406.16993`.

[54] A. Myronenko, 3d mri brain tumor segmentation using autoencoder regularization, in: International MICCAI brainlesion workshop, Springer, 2018, pp. 311–320.

[55] F. Galati, D. Falcetta, R. Cortese, B. Casolla, F. Prados, N. Burgos, M. A. Zuluaga, A2v: A semi-supervised domain adaptation framework for brain vessel segmentation via two-phase training angiography-to-venography translation, arXiv preprint arXiv:2309.06075 (2023).

[56] B. Billot, D. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. Dalca, J. Iglesias, Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining, Medical Image Analysis 86 (2023) 102789.

[57] S. Roy, G. Koehler, C. Ulrich, M. Baumgartner, J. Petersen, F. Isensee, P. F. Jaeger, K. H. Maier-Hein, Mednext: transformer-driven scaling of convnets for medical image segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2023) 405–415.

[58] H. H. Lee, S. Bao, Y. Huo, B. A. Landman, 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation, arXiv preprint arXiv:2209.15076 (2022).

[59] P. Shi, X. Guo, Y. Yang, C. Ye, T. Ma, Nextou: Efficient topology-aware u-net for medical image segmentation, arXiv preprint arXiv:2305.15911 (2023).

[60] M. Zhang, X. You, H. Zhang, Y. Gu, Topology-aware exploration of circle of willis for cta and mra: Segmentation, detection, and classification, arXiv preprint arXiv:2410.15614 (2024).

[61] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, D. Xu, Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, in: International MICCAI BrainLesion Workshop, Springer, 2021, pp. 272–284.

[62] F. Milletari, N. Navab, S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 fourth International Conference on 3D Vision (3DV), IEEE, 2016, pp. 565–571.

[63] F. Autrusseau, R. Nader, A. Nouri, V. L'Allinec, R. Bourcier, Toward a 3d arterial tree bifurcation model for intra-cranial aneurysm detection and segmentation, in: 2022 26th International Conference on Pattern Recognition (ICPR), IEEE, 2022, pp. 4500–4506.

[64] R. Nader, F. Autrusseau, V. L'Allinec, R. Bourcier, A vascular synthetic model for improved aneurysm segmentation and detection via deep neural networks, arXiv preprint arXiv:2403.18734 (2024).

[65] C. Acebes, A. H. Moustafa, O. Camara, A. Galdran, The centerline-cross entropy loss for vessel-like structure segmentation: Better topology consistency without sacrificing accuracy, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 710–720.

[66] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video swin transformer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 3202–3211.

[67] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

[68] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[69] A. Hilbert, V. I. Madai, E. M. Akay, O. U. Aydin, J. Behland, J. Sobesky, I. Galinovic, A. A. Khalil, A. A. Taha, J. Wuerfel, et al., Brave-net: fully automated arterial brain vessel segmentation in patients with cerebrovascular disease, Frontiers in artificial intelligence 3 (2020) 552258.

[70] J. Muschelli, A publicly available, high resolution, unbiased ct brain template, in: Information Processing and Management of Uncertainty in Knowledge-Based Systems: 18th International Conference, IPMU 2020, Lisbon, Portugal, June 15–19, 2020, Proceedings, Part III 18, Springer, 2020, pp. 358–366.

[71] S. K. Warfield, K. H. Zou, W. M. Wells, Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation, IEEE transactions on medical imaging 23 (2004) 903–921.

[72] M. Baumgartner, P. F. Jäger, F. Isensee, K. H. Maier-Hein, nndetection: a self-configuring method for medical object detection, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, Springer, 2021, pp. 530–539.

[73] D. Drees, A. Scherzinger, R. Hägerling, F. Kiefer, X. Jiang, Scalable robust graph and feature extraction for arbitrary vessel networks in large volumetric datasets, BMC bioinformatics 22 (2021) 346.

Supplementary Material

For the paper *"Benchmarking the CoW with the TopCoW Challenge: Topology-Aware Anatomical Segmentation of the Circle of Willis for CTA and MRA"*

S1. TopCoW Data Cohort

The patients of the challenge cohort were admitted for or recovering from a stroke-related neurological disorder, including ischemic stroke, transient ischemic attack, stroke mimic, retinal infarct or amaurosis fugax, intracerebral hemorrhage, and cerebral sinus vein thrombosis.

Other relevant clinical characteristics of the TopCoW dataset include sex (men 59.5%, women 40.0%) and age (mean 70.9 years, standard deviation 14.9 years).

S2. Inclusion and Exclusion of CoW Variants

We tried to include as many diverse CoW variants as possible in our challenge dataset. From our observation, the following variants were **included** and annotated in our training and test dataset:

- ✓ with or without Acom (*Note: Acom determines A1/A2*)
- ✓ double Acom
- ✓ with or without Pcom (*Note: Pcom determines P1/P2*)
- ✓ the triple ACA variant or 3rd-A2
- ✓ aplastic or hypoplastic A1 or P1 segments
- ✓ fetal PCA variants
- ✓ when CoW vessels (e.g. ACA, PCA, Acom) have fenestrations

However, there are some rare variants of which the topology cannot be characterized by our CoW multiclass labels. These variants are much less common than the ones we have included, and our thirteen CoW segment labels are insufficient to describe the complex anatomy of these rare variants. Here is a list of such CoW variants that we had observed and **excluded** from our dataset:

- ✗ azygos ACA or when the left and right ACAs are fused
- ✗ anterior choroidal artery (AChA) course and supply replacing an ipsilateral fetal PCA
- ✗ duplicated PCA
- ✗ persistent primitive trigeminal artery between ICA and BA

S3. Anonymization, Defacing, and Pre-processing

The data were anonymized (removal and anonymization of relevant DICOM patient information). Additional de-facing and cropping procedures were performed to ensure patient privacy in the image data after converting the DICOM to NIfTI format using dcm2niix [39]. Specifically, we masked out the face using TotalSegmentator [40] for CTA and shear-cutted the facial regions with quickshear method [41] for MRA, and then cropped the image data using brain mask from TotalSegmentator for CTA or HD-BET [42] for MRA to include only the braincase region.

Other than the defacing and cropping-to-braincase of the nifti image data, the only pre-processing of the data was to re-orient the image to LPS+ orientation. No further pre-processing of the data was performed to keep the data as close to the original clinical setting as possible.

S4. Notes on the Updates to the 2023 TopCoW Data

In 2023, due to technical reasons related to image data type, a few test images frequently caused inference runtime error in participant Docker containers. In order for the participants to be aware of such technical issues, we moved one representative patient to the validation set in 2024 so participants could pick up on related technical mistakes early on before submitting for the final test set.

In 2023, in addition to the TopCoW challenge training data, we also released a small specific subset (20 MRA scans) of the CROWN challenge [21] data with our annotations (multiclass CoW voxel mask and CoW ROI). This data was not included in the 2024 data release, but it is available upon request.

In 2024, we also refined the segmentation mask and bounding box labels for some of the 2023 data, both for training and the test data, and re-released the updated 2023 data in the 2024 release. We trimmed floating blobs of disconnected components smaller than 13 voxels for each CoW mask labels, which leads to more accurate ground-truth Betti-0 numbers for segmentation masks. We slightly adjusted some CoW ROI 3D bounding box labels for some 2023 cases, making them more harmonized and consistent with our bounding box definition.

S5. Details on Annotation & Verification Protocol

The annotation protocol on how to segment vessel components and boundaries at bifurcation points such as ACA-ICA-MCA, ACA-Acom, PCA-Pcom, Pcom-ICA, etc., were discussed and agreed upon by the clinical experts. For example, we marked the superior tip of the ACA-ICA-MCA bifurcations to be part of ICA, and similarly for BA-PCA bifurcation, we marked the tip to be of BA. We also included the infundibulum as the origin of certain vessel components such as Pcom. The annotation protocol also covered CoW variants such as fetal PCA, triple ACA etc.

For the TopCoW dataset annotation, since the TopCoW data had both CTA and MRA modalities for the same patients, the anatomy of the CoW was first inspected in both CTA and MRA

to diagnose the anatomical components. Then the CoW vessels were annotated or verified for each modality.

CoW annotations for the initial 260 images were manually labeled. The remaining data used in the benchmark were pre-labeled using a model that was based on the 2023 submission from team 'DKFZ' and trained on the initial 260 annotated images. All images were manually corrected and verified.

For vessels extending beyond the CoW ROI such as the ACA, MCA, and PCA, we typically only labelled until the first major bifurcation occurs, and we only labelled the main vessel instead of any minor branches. For the CTA modality, the ICAs were not labelled through the anterior clinoid and sphenoid bone regions, but were labeled starting from the C7 segment in Bouthillier classification system [43]. For MRA, we labelled the entire curvature of the ICA in the CoW region even in bone regions.

S6. Inter-Rater Agreement for Voxel-Level Segmentation

Fig. S1 shows the Dice scores for all 13 CoW labels for the CoW anatomical annotations from both annotators (K.Y. and F.M.).

Table S1 shows the per case performance of the inter-rater agreement. The 90% class-average Dice per case indicated good agreement in raters' multiclass voxel annotations. The merged binary mask had around 95% Dice which showed good agreement for binary segmentation annotations. Besides the Dice similarity coefficient, we also compared the raters' annotations in terms of centerline Dice or clDice [32] and errors in connected components using zero-th Betti number β_0. The clDice and β_0 errors both had very good inter-rater agreement at near maximum scores as shown in Table S1.

S7. CoW Variant Distributions

We document the prevalence of all present CoW anterior and posterior variants from our training and test data in Table S2 and Table S3.

S8. Evaluation Metrics for All Tasks

For the 2023 iteration, there were two tasks and both were for segmentation: multiclass segmentation of the anatomical components of the CoW and binary segmentation of the CoW vessels. For binary segmentation task, the binary vessel label was generated by merging the multiclass CoW labels. In 2024, we kept the multiclass segmentation task but discontinued the binary segmentation task from 2023. We introduced two new tasks in 2024: a detection task for the CoW ROI and a classification task for the CoW variant graphs.

Here we describe in details the metrics used for all the tasks. The segmentation tasks in 2023 were evaluated using three metrics: Dice similarity coefficient, centerline Dice (clDice) [32], and zero-th Betti number (β_0) error. In 2024, we added the following segmentation metrics: Hausdorff distance at 95% percentile (HD95), average F1 score for detection of Acom, Pcoms, and 3rd-A2, variant balanced accuracy (VarBalAcc) of CoW variant graph classification, and balanced CoW topology match rate (TMR). The last metric was a custom composite metric combining detection, connectivity, and graph classification metrics: For labels in anterior or posterior variants, the predicted segmentation needs to satisfy correct detection, correct neighborhood connectivity (connected to valid vessel classes), no 0-th Betti number errors, and not left-right flipped in order to be counted as a match in the topology.

The CoW ROI detection task was evaluated using two metrics: intersection over union (IoU) and boundary IoU with boundary distance of 20% of the box dimension [44, 45] between the ground-truth bounding box and the predicted box.

Our evaluation code with documentation was open sourced at `https://github.com/CoWBenchmark/TopCoW_Eval_Metrics`.

S9. Ranking Analysis for 2024 Segmentation Algorithms

The segmentation algorithms were ranked by averaging the rank positions for each metric. The ranking on the internal test data was used to select the best algorithms. We also created 10 bootstraps of the internal test sets and calculated the rankings on the bootstrapped sets. Table S4 shows the ranking on the original non-bootstrapped set and the average ranking of 10 bootstrapped test sets. The rankings after bootstraps were stable, and supported the selection of the top teams.

S10. Details on Fetal PCA Classification

Fig. S2 shows the intermediate workflow in *Voreen* [46] for computing diameters along Pcom and P1 centerlines from segmentation masks for fetal PCA classification.

S11. Descriptions of Submitted Algorithms and Teams

In this section we summarize the methods and algorithms of all the participating teams ordered alphabetically by team names:

2i_mtl. The team included Emmanuel Montagnon and Laurent Letourneau-Guillon. The team took part in the CT binary task in 2023. They employed a two-stage approach consisting of a patch-based 3D AttentionUNet [47] followed by a 3D autoencoder to mitigate false positives. The autoencoder received as input both an image patch and the AttentionUNet mask prediction.

agaldran. The submissions were made by Adrian Galdran. He took part in all four tracks and tasks in 2023. His approach was based on a self-adapting 3D dynamic UNet provided by the MONAI library [48, 49]. Internal cross-validation on the volumetric patch size over various sizes was performed.

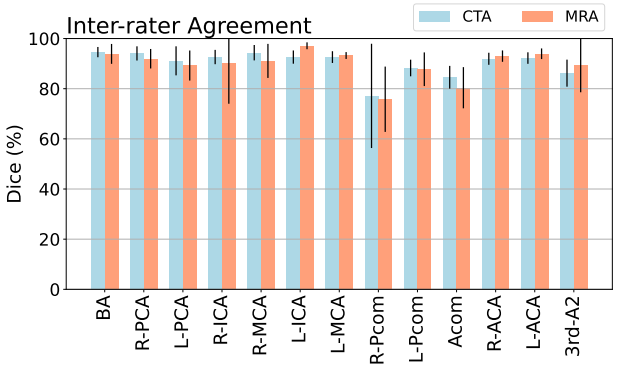

Fig. S1. Inter-rater agreement for voxel-level segmentation on a subset of the internal test set ($n = 5$) for all 13 CoW component classes in Dice scores ± standard deviations. Note that the n for R-Pcom, L-Pcom, and 3rd-A2 are $3, 3, 3$ for CTA and $4, 4, 3$ for MRA, respectively.

Table S1. Inter-rater agreement results (mean ± standard deviation) on segmentation of a subset of the internal test set ($n = 5$) for MRA and CTA in terms of binary and class average Dice scores, centerline Dice (clDice) scores, binary and class average errors of the zero-th Betti number β_0. The binary Dice, binary β_0 errors, and *clDice* scores were computed on the merged binary class. The class-average *Dice* scores and β_0 errors were computed for each class separately and the average was taken per case. The arrow indicates the favorable direction.

Inter-rater agreement on segmentation					
Modality	Binary Dice (%) ↑	Per case class-average Dice (%) ↑	clDice (%) ↑	Binary β_0 error ↓	Per case class-average β_0 error ↓
CTA	94.86 ± 2.24	90.91 ± 3.35	99.72 ± 0.24	0 ± 0	0 ± 0
MRA	96.49 ± 2.01	90.21 ± 4.02	99.49 ± 0.80	0.4 ± 0.55	0.07 ± 0.07

Table S2. Distribution of CoW anterior variants (AV) across the TopCoW train and test sets as well as the external datasets. The AV is identified by a four-edge graph, with 0 being absent and 1 being present in the edge-list. The edges are: L-A1, Acom, 3rd-A2, R-A1. Values are reported as absolute counts, with relative percentages in parentheses.

Distribution of CoW anterior variants						
Dataset	AV-0101	AV-1001	AV-1100	AV-1101	AV-1111	Total
TopCoW CT Train	3 (2.4%)	19 (15.2%)	5 (4.0%)	83 (66.4%)	15 (12.0%)	125
TopCoW CT Test	0	14 (20.0%)	3 (4.3%)	47 (67.1%)	6 (8.6%)	70
ISLES	0	7 (26.9%)	1 (3.8%)	17 (65.4%)	1 (3.8%)	26
LargeIA	0	4 (20.0%)	0	14 (70.0%)	2 (10.0%)	20
TopCoW MR Train	3 (2.4%)	19 (15.2%)	5 (4.0%)	82 (65.6%)	16 (12.8%)	125
TopCoW MR Test	0	12 (17.1%)	3 (4.3%)	48 (68.6%)	7 (10.0%)	70
Lausanne	0	4 (20.0%)	0	14 (70.0%)	2 (10.0%)	20
IXI-HH	1 (5.0%)	1 (5.0%)	0	16 (80.0%)	2 (10.0%)	20

Table S3. Distribution of CoW posterior variants (PV) across the TopCoW train and test sets as well as the external datasets. The PV is identified by a four-edge graph, with 0 being absent and 1 being present in the edge-list. The edges are: L-Pcom, L-P1, R-P1, R-Pcom. Values are reported as absolute counts, with relative percentages in parentheses.

Distribution of CoW posterior variants										
Dataset	PV-0101	PV-0110	PV-0111	PV-1001	PV-1010	PV-1011	PV-1101	PV-1110	PV-1111	Total
TopCoW CT Train	3 (2.4%)	46 (36.8%)	21 (16.8%)	2 (1.6%)	2 (1.6%)	3 (2.4%)	5 (4.0%)	16 (12.8%)	27 (21.6%)	125
TopCoW CT Test	2 (2.9%)	36 (51.4%)	12 (17.1%)	0	0	6 (8.6%)	0	7 (10.0%)	7 (10.0%)	70
ISLES	3 (11.5%)	12 (46.2%)	3 (11.5%)	0	0	1 (3.8%)	0	2 (7.7%)	5 (19.2%)	26
LargeIA	0	7 (35.0%)	5 (25.0%)	3 (15.0%)	0	1 (5.0%)	0	2 (10.0%)	2 (10.0%)	20
TopCoW MR Train	3 (2.4%)	47 (37.6%)	19 (15.2%)	0	2 (1.6%)	3 (2.4%)	4 (3.2%)	15 (12.0%)	32 (25.6%)	125
TopCoW MR Test	2 (2.9%)	33 (47.1%)	12 (17.1%)	0	0	4 (5.7%)	0	6 (8.6%)	13 (18.6%)	70
Lausanne	1 (5.0%)	5 (25.0%)	3 (15.0%)	0	0	1 (5.0%)	0	3 (15.0%)	7 (35.0%)	20
IXI-HH	0	7 (35.0%)	2 (10.0%)	0	0	1 (5.0%)	0	2 (10.0%)	8 (40.0%)	20

Table S4. Mean position (mean±std) of the ranking for the multiclass segmentation task for both CTA and MRA. The ranking shown is either from the non-bootstrapped internal test set (leaderboard) or from the average across 10 bootstraps. The mean position is computed as the average rank position across the 9 evaluation metrics used for the segmentation task. The top three values for each column are marked with gold, silver and bronze colors. If a team only submitted to one of the tracks, the columns of the other track are filled with a '-'.

TopCoW multiclass segmentation ranking				
	CTA		MRA	
Team	Leaderboard mean position	Bootstrapped mean position	Leaderboard mean position	Bootstrapped mean position
ARG	6.00 ± 1.25	6.30 ± 1.39	5.56 ± 1.34	5.72 ± 1.31
CLAIM	3.33 ± 2.16	3.28 ± 1.97	2.22 ± 1.40	2.18 ± 1.29
DeepLearnAI	10.11 ± 1.10	9.92 ± 1.09	7.44 ± 1.26	7.52 ± 1.48
DKFZ	5.44 ± 1.71	5.12 ± 1.64	4.89 ± 1.79	4.57 ± 1.78
DLaBella29	7.33 ± 1.49	7.17 ± 1.56	-	-
HITSZ	4.67 ± 1.70	4.90 ± 1.68	3.00 ± 0.94	3.30 ± 1.21
IMR	3.33 ± 2.00	3.58 ± 1.77	4.56 ± 2.06	4.40 ± 2.29
junqiangchen	9.89 ± 0.99	9.97 ± 0.88	10.00 ± 0.00	9.80 ± 0.54
NIC-VICOROB	3.22 ± 2.15	3.30 ± 2.07	6.11 ± 2.23	6.51 ± 1.97
pamaad	10.00 ± 1.05	10.01 ± 1.21	8.22 ± 0.92	8.03 ± 1.17
UB-VTL	11.78 ± 0.42	11.79 ± 0.44	-	-
UZH	2.78 ± 1.47	2.62 ± 1.79	2.89 ± 2.56	2.87 ± 2.23

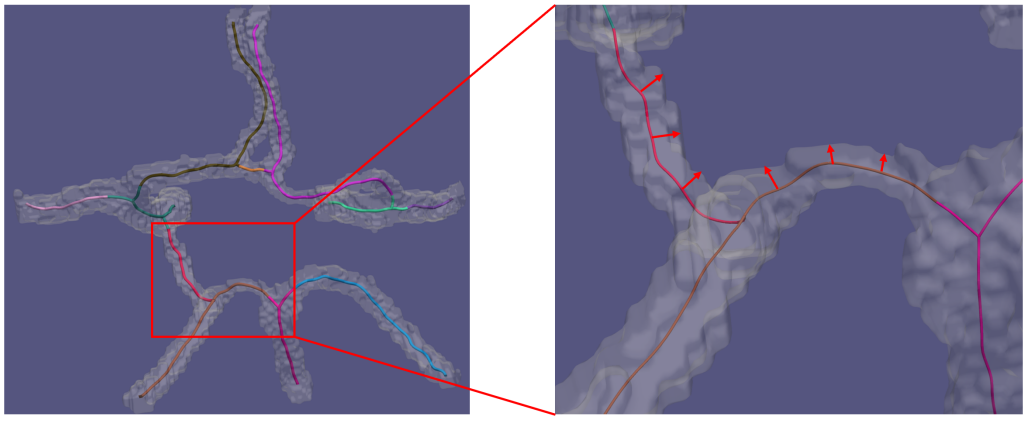

Fig. S2. The radius estimation of the Pcom and the P1 segments was based on the centerline graph and the surface mesh of the segmentation mask. The centerline graphs were extracted using *Voreen* [46] with the radius given as an attribute for each edge of the graph.

ARG. The full team name was 'ARG-DEEPNeuro'. The team consisted of Kwanseok Oh and Dahye Lee. They took part in all three tasks for both MRA and CTA tracks in 2024. For multiclass segmentation task and CoW detection task, they used the 3D nnUNet to segment the CoW vessels. In segmentation task, they added the Generalized Surface Loss [50] to the default loss functions (Dice and CE) and applied class weight multiplication. For detection task, they trimmed disconnected components before extracting bounding box coordinates. For both tasks on the CTA track, models were trained using both MRA and CTA modalities. For variant graph classification task, they used a nnUNet encoder combined with a self-attention mechanism tailored for anterior and posterior predictions. **The segmentation algorithm has been included in our Zenodo Docker release (records/15665435).**

CLAIM. The team was composed of Adam Hilbert, Orhun Aydin, Jana Rieger, Dimitrios Rallios, Satoru Tanioka, and Dietmar Frey. They participated in all three tasks for both MRA and CTA tracks in 2024. All models were trained on mixed MRA and CTA scans. For multiclass segmentation task, they used a combined dataset of TopCoW and 500 additional scans enriched for rare anatomical variants. A two-stage approach was employed: an initial ROI detection step using their task 2 detection model, followed by segmentation with the 3D nnUNet within the enlarged ROI. To improve topological connectedness, they incorporated the Skeleton Recall (SkelRecall) loss [35]. For CoW detection task, they used YOLOv8 by Ultralytics [51] trained on 2D slices to detect ROIs, which were aggregated into 3D boxes. For CoW graph classification task, they implemented a rule-based graph extraction algorithm operating on the output of segmentation task. Their segmentation algorithm and a list of all the additional dataset references are available at `https://github.com/claim-berlin/TopCoW_2024_MRA_winning_solution`. **The segmentation algorithm has been included in our Zenodo Docker release (records/15665435).**

DeepLearnAI. The team was composed of Abdul Qayyum, Moona Mazher, and Steven A. Niederer. They participated in the multiclass segmentation task for both MRA and CTA tracks in 2024. They used a two-stage approach: the first stage em-

ployed self-supervised learning using DINOv2 [52] to pretrain representations, and the second stage used a 3D xLSTM-UNet model [53] for downstream segmentation. The pipeline was implemented using MONAI.

DKFZ. Also known as team 'WilliWillsWissen' in 2023. The submissions were made by Maximilian R. Rokuss, Yannick Kirchhoff, Nico Disch, Julius Holzschuh, Fabian Isensee and Klaus Maier-Hein. The team took part in the segmentation tasks for both MRA and CTA tracks in both 2023 and 2024. They used a patch-based 3D nnUNet-with various adaptations and trained on both modalities. To improve connectedness they incorporated the clDice [32] and a recall on the skeleton of the vessels (SkelRecall); and they employed extensive cross-validation ensemble per model as well as a subsequent ensemble of differently trained models for each submission. **The segmentation algorithm has been included in our Zenodo Docker release (records/15665435).**

DLaBella29. The submissions were made by Dominic LaBella, who participated in the multiclass segmentation task for both MRA and CTA tracks in 2024. He used a 3D SegResNet [54] via MONAI and Auto3DSeg. A connectivity check ensured R/L-Pcom contact with the PCA and ICA, and post-processing included removal of distant or small components.

EURECOM. The team consisted of Francesco Galati, Daniele Falcetta and Maria A. Zuluaga. They took part in both binary segmentation tasks in 2023 and applied a single model strategy for multi-domain vessel segmentation. They employed an adapted version of their A2V framework [55] consisting of a single encoder-generator architecture for image reconstruction, translation, and ultimately segmentation with a shared latent space for both modalities. In a first step they extracted brain masks required by A2V using SynthSeg [56]. The code is available at `https://github.com/i-vesseg/MultiVesSeg`.

gbCoW. The team was made up of Chaolong Lin and Haoran Zhao. They took part in the MRA binary and multiclass tasks in 2023 using the nnUNet framework for patch-based 3D segmentation. They trained a single model on the multiclass labels only; the binary masks were obtained from their multiclass predictions.

gl. The submissions were made by Zehan Zhang. He submitted algorithms to both multiclass segmentation tasks in 2023 following a multi-step approach consisting of 1) the extraction of a custom ROI using a dataset specific atlas and affine registration, 2) binary segmentation and 3) subsequent multiclass segmentation with a 2-channel input (image ROI and binary mask). For the segmentation, both the 3D MedNexT [57] and UX-Net [58] architectures were employed as an esemble. The inference code is available at `https://github.com/zzh980123/TopCoW_Algo_Submission`.

HITSZ. Also known as team 'NexToU' in 2023. The team consisted of Pengcheng Shi, Wei Liu and Ting Ma. They participated in the segmentation tasks for both MRA and CTA tracks in both 2023 and 2024. They used a two-stage pipeline: a low-resolution model trained on binary labels using the 3D nnUNet, followed by a full-resolution model trained on multiclass labels using their own NexToU architecture [59]. Both stages were trained on MRA and CTA data. Besides the default losses, they used the cbDice loss [36], which is both topology-aware and diameter-balanced, and a hierarchical topological interaction loss. The code for NexToU is at `https://github.com/PengchengShi1220/NexToU`; the cbDice loss function is available at `https://github.com/PengchengShi1220/cbDice`. **The segmentation algorithm has been included in our Zenodo Docker release (records/15665435).**

IMR. The team was composed of Minghui Zhang, Xin You, Hanxiao Zhang, Guang-Zhong Yang, and Yun Gu. They participated in all three tasks for both MRA and CTA tracks in 2024. All models were based on the 3D nnUNet and trained on both modalities. For multiclass segmentation task, they further developed the connectivity-aware loss (CAL) based on [37] to improve topological completeness, and applied a topology-aware refinement postprocessing step to repair disconnected vessel components. Detection task was reformulated as segmentation followed by bounding box extraction. Classification task was inferred directly from segmentation outputs without a separate classification head. A detailed method description can be found on their ArXiv pre-print [60]. **The segmentation algorithm has been included in our Zenodo Docker release (records/15665435).**

IWantToGoToCanada. The team consisted of Sinyoung Ra, Jongyun Hwang and Hyunjin Park. They took part in the CTA binary and multiclass tasks in 2023. The nnUNet was used to extract the binary segmentation mask. For the subsequent multiclass segementation a 3D Swin-UNETR [61] architecture was employed with both the image and the binary mask as input.

junqiangchen. The submissions were made by Junqiang Chen, taking part in all tasks and tracks in both 2023 and 2024. In 2023, a two-stage approach was employed using the VNet3D [62] for both stages: In the first stage a custom ROI was extracted based on a binary segmentation, in the second stage the segmentation was performed on the ROI only. In 2024, all models were trained using mixed MRA and CTA data. For segmentation and detection tasks, he used a 3D VNet architecture to segment the CoW vessels. For classification task, a 3D ResNet was used. The code is available on `https://github.com/junqiangchen/PytorchDeepLearing`.

lWM. The team was composed of Marek Wodzinski and Henning Müller. They took part in all four tracks and tasks in 2023 using a patch-based 3D ResidualUNet with a focus on data preprocessing and augmentation.

NantesU. Nesrin Mansouri and Florent Autrusseau participated in the segmentation task for the MRA track only in 2024, using the 3D nnUNet. To address class imbalance, they generated 504 synthetic MRA images focusing on underrepresented arteries, expanding the dataset from 125 to 629 images. Their synthetic modeling approach was based on prior work [63, 64].

NIC-VICOROB. Also known as team 'NIC-VICOROB-1' in 2023. The team was made up of Cansu Yalçin, Rachika E. Hamadache, Clara Lisazo, Joaquim Salvi, Adrià Casamitjana, and Xavier Lladó. In 2023, the team took part in all tracks and tasks using a patch-based 3D nnUNet. Working in two stages with a 2-channel input consisting of both the image and the binary mask improved the segmentation results for both the CTA and MRA multiclass tasks. In 2024, they participated in segmentation and detection tasks for both tracks. Models based on 3D nnUNnet were trained on MRA only for the MRA track, and on both modalities for the CTA track. For multiclass segmentation task, a two-stage pipeline was similar to 2023 was used: multiclass segmentation followed by binary segmentation for postprocessing. For detection task, CoW detection was treated as a segmentation task using binarized masks. For topological optimization they used the Skeleton Recall (SkelRecall) loss and did postprocessing on the predictions including background filling, small component removal, and connectivity restoration. **The segmentation algorithm has been included in our Zenodo Docker release (records/15665435).**

NIC-VICOROB-2. In 2023, NIC-VICOROB had a second team consisting of Uma Maria Lal-Trehan Estrada, Valeriia Abramova, Luca Giancardo and Arnau Oliver, taking part in all four tracks and tasks. For the binary segmentation tasks they employed a patch-based 3D AttentionUNet. For the multiclass segmentation tasks they employed a two-stage approach using a 2D AttentionUNet with full axial slices and binary segmentation masks, obtained from the 3D AttentionUNet, as input. The 2D approach was chosen due to GPU memory limitations.

pamaad. The team was composed of Paula Casademunt, Adrian Galdran, and Matteo Delucchi. They participated in all three tasks for both MRA and CTA tracks in 2024. For segmentation and detection tasks, they used 3D nnUNet with a novel centerlineCE loss [65]. For segmentation task, they used a two stage approach involving an initial segmentation for custom ROI cropping followed by multiclass segmentation on the ROI. For classification task, they used a Video Swin Transformer [66] adapted to 3D medical data.

refrain. The submissions were made by Jialu Liu, Haibin Huang and Yue Cui. They submitted algorithms for the MRA binary and multiclass task, employing a 3D nnUNet for both tasks. A template atlas was used to extract a custom ROI via registration. Furthermore, they used data augmentation to balance the training set with respect underrepresented CoW variants and applied segment specific loss weighting with higher weights for R-Pcom, L-Pcom, Acom and 3rd-A2. The code is available on `https://github.com/Vessel-Segmentation/Topcow_private`.

sjtu_eiee_2-426lab. The submissions were made by Zehang Lin, Yusheng Liu and Shunzhi Zhu, taking part in the CTA binary and multiclass tasks. They used a two-stage approach using the 3D nnUNet: A first binary segmentation for a custom ROI extraction followed by a segmentation on the extracted ROI only.

SynthCLAIM. CLAIM had a second team that included Alexander Koch. They participated in multiclass segmentation task of the MRA track only in 2024. To generate training data, they trained a StyleGAN [67] on the TopCoW MRA scans and synthesized 10,000 TOF-MRA volumes. These were pseudo-labeled using the nnUNet segmentation model from team CLAIM. A two-stage pipeline was then trained exclusively on this synthetic dataset: first, a 2D YOLO-based [68] model was trained to detect the CoW ROI slice-wise; second, a 3D nnUNet performed multiclass segmentation within the detected ROIs. To enhance topological accuracy, the nnUNet incorporated the Skeleton Recall (SkelRecall) loss.

UB-VTL. The team consisted of Tatsat R. Patel, Adnan H. Siddiqui, and Vincent M. Tutino. In 2023, they participated in binary segmentation tasks employing a patch-based 3D BRAVE-NET [69] taking as input a normal patch and a low-res patch for more context. As modifications, they added residual connections, used parametric rectified linear units (PReLu) as activations and worked with the centerline Dice (clDice) as a loss function. In 2024, they participated in the segmentation and detection tasks for the CTA track only. For segmentation task, they used a two-stage pipeline based on BRAVE-NET. Stage 1 performed binary segmentation of CoW vessels using the clDice loss; stage 2 used a modified BRAVE-NET with the stage 1 binary mask as additional input to perform multiclass segmentation. For detection task, they used a two-stage atlas-based method without deep learning: CTA images were registered to a CT atlas [70], and the atlas ROI was computed using the STAPLE [71] algorithm. During inference, the atlas ROI was registered to test images, and Cartesian ray tracing was used to predict CoW coordinates.

UW. The submission was made by Maysam Orouskhani, Huayu Wang, Mahmud Mossa-Basha and Chengcheng Zhu. They took part in the MRA binary task in 2023 using the patch-based 3D nnUNet framework with a modified 3-component loss function consisting of Dice, Cross-Entropy (CE) and TopK loss. The code, models and trained weights can be accessed via `https://github.com/orouskhani/TopCow2023`.

UZH. Also known as team 'Organizers' in 2023. The team was composed of Houjing Huang, Fabio Musio, Chinmay Prabhakar, Suprosanna Shit, and Kaiyuan Yang. In 2023, the team participated in the multiclass tasks of both tracks using a two-stage approach: stage-1 detection of the CoW ROIs with nnDetection [72] based on the binary vessel labels and stage-2 multiclass segmentation on the ROIs with 3D nnUNet. Additionally, inter-modal registration was used as a data augmentation strategy, registering all the image pairs and thereby doubling the size of the training set for both modalities. In 2024, the team upgraded the detection module with a nnUNet segmentation model for the binary 'brick' masks enclosed by the CoW ROI bounding box coordinates. All segmentation models were now trained on mixed modalities. Inter-modal registration for data augmentation was kept and used to double the training set. The stage-2 segmentation also added the Skeleton Recall (SkelRecall) loss. For classification task, again a two-stage method was

applied: binary vessel segmentation using nnUNet, followed by edge prediction. For the edge prediction, the binary segmentation mask was converted into a graph using Voreen [46, 73], and then edge labels were predicted based on topological properties of the graph. **The segmentation algorithm has been included in our Zenodo Docker release (records/15665435).**

ysato. The submission was made by Yuki Sato, taking part in the MRA binary task. The author employed a non-deep learning approach based on recursive algorithm consisting of auto vessel thresholding and region growing with a rule-based automated seed point selection. It was the only non-deep learning based submission to our challenge. Accordingly, the inference time per case was very short (~15s) and the computations could be done on a CPU.

S12. Detailed Results of the 2024 Tasks for Internal Test

Table S5, Table S6, and Table S7 show the results of the 2024 multiclass segmentation task, the detection task for the CoW ROI, and the classification task for the CoW variant graphs on the TopCoW internal test data.

S13. Detailed Results of the 2024 Tasks for External Test

Table S8, Table S9, and Table S10 show the results of the 2024 tasks on the external test data.

S14. Inference Time for Best Segmentation Algorithms

Table S11 shows the inference time in seconds by the top segmentation algorithms.

The top algorithms were selected for further benchmarking on external test sets. During this process, we performed a runtime analysis and measured their inference time using a laptop with GPU. The runtime depended on the dataset, but none of the datasets took more than two and a half minutes on average on a test image for prediction. Teams 'UZH' and 'NIC-VICOROB' had slightly longer inference time of around 2 minutes per test image. The other five teams had average inference time ranging from 35 seconds to 69 seconds per image.

S15. Detail Results on Locating Aneurysm

Table S12 lists in detail the selected 12 patients with intracranial aneurysms from the LargeIA CTA dataset, their aneurysm locations, and the locations in terms of CoW labels as predicted by the top teams.

Fig. S3 shows the aneurysm and CoW vessels of the patient Tr0004 for which most top teams had an error in locating this aneurysm.

S16. Detailed Results of the 2023 Tasks

Table S13 and Table S14 show the results for the 2023 binary segmentation task and the multiclass segmentation task. Fig. S4 shows the qualitative examples of the predictions from one of the 2023 winning teams for a few selected anterior (Fig. S4 top) and posterior (Fig. S4 bottom) variants.

Table S5. Results (mean ± standard deviation) of the CoW multiclass segmentation task from 2024 on the TopCoW internal test set in terms of class-average Dice similarity coefficient, centerline Dice (clDice) on merged binary mask, class-average 0-th Betti number (β_0) error, class-average Hausdorff Distance 95% Percentile (HD95), average F1 score for detection of communicating arteries and 3rd-A2, variant-balanced accuracy (VarBalAcc) of graph classification, and CoW variant topology match rate (TMR) for both the anterior and posterior variants. The arrow indicates the favorable direction. The top three values for each metric are marked in gold, silver, and bronze colors. Teams marked with a '*' were late submissions due to technical issues. The bottom two rows for each modality are the results for the top-3 teams ensemble and the organizer teams ('UZH', 'DKFZ', 'HITSZ') ensemble.

Multiclass segmentation performance on the 2024 TopCoW internal test data

CTA (n=70)

Team	Per case class-avg Dice (%) ↑	clDice (%) ↑	Per case class-avg β_0 error ↓	Per case class-avg HD95 ↓	Average F1 score (%) ↑	Anterior VarBalAcc (%) ↑	Posterior VarBalAcc (%) ↑	Anterior TMR (%) ↑	Posterior TMR (%) ↑
ARG	85.05 ± 6.74	98.75 ± 1.28	0.13 ± 0.14	4.85 ± 5.03	77.19 ± 9.16	73.94	65.87	39.98	42.79
CLAIM	84.92 ± 5.63	99.00 ± 1.15	0.06 ± 0.09	3.03 ± 3.87	83.11 ± 9.63	73.94	83.60	45.21	69.97
DeepLearnAI	73.39 ± 22.52	96.38 ± 3.55	0.23 ± 0.24	11.61 ± 10.71	21.63 ± 37.47	66.45	16.67	24.01	12.96
DKFZ	86.04 ± 6.43	98.90 ± 1.27	0.10 ± 0.10	3.97 ± 4.33	73.76 ± 17.47	79.95	65.41	48.13	41.87
DLaBella29	84.14 ± 6.28	98.16 ± 2.09	0.07 ± 0.08	5.20 ± 5.01	78.09 ± 10.56	60.03	52.78	32.28	43.06
HITSZ	85.03 ± 6.55	98.39 ± 2.22	0.10 ± 0.11	3.76 ± 4.20	79.50 ± 9.44	86.03	71.36	48.51	49.80
IMR	87.13 ± 5.68	98.60 ± 1.64	0.06 ± 0.09	2.88 ± 3.71	86.01 ± 8.17	88.88	62.63	46.59	49.93
junqiangchen	74.81 ± 7.19	96.96 ± 2.56	0.34 ± 0.24	9.72 ± 6.51	48.62 ± 30.14	47.34	44.51	7.98	16.73
NIC-VICOROB	88.01 ± 5.98	98.94 ± 1.41	0.04 ± 0.06	3.45 ± 4.43	74.97 ± 26.96	72.21	74.74	57.05	62.96
pamaad	64.66 ± 29.64	98.23 ± 2.02	0.31 ± 0.34	11.34 ± 10.83	57.44 ± 16.13	45.09	38.49	12.42	13.96
UB-VTL	65.60 ± 9.29	88.87 ± 7.02	1.10 ± 0.72	17.92 ± 7.87	0.00 ± 0.00	19.64	12.04	8.93	10.65
UZH	87.37 ± 5.89	98.57 ± 2.42	0.04 ± 0.05	3.22 ± 4.25	85.78 ± 5.32	84.37	68.06	73.91	57.74
Top-3 ensemble	88.49 ± 5.80	98.84 ± 1.54	0.05 ± 0.08	3.04 ± 4.05	84.11 ± 10.34	84.71	68.72	52.36	56.02
Orgs ensemble	87.56 ± 6.04	98.89 ± 1.44	0.08 ± 0.09	3.23 ± 3.96	80.21 ± 12.40	86.56	63.49	70.15	42.79

MRA (n=70)

Team	Per case class-avg Dice (%) ↑	clDice (%) ↑	Per case class-avg β_0 error ↓	Per case class-avg HD95 ↓	Average F1 score (%) ↑	Anterior VarBalAcc (%) ↑	Posterior VarBalAcc (%) ↑	Anterior TMR (%) ↑	Posterior TMR (%) ↑
ARG	87.69 ± 6.34	98.88 ± 1.51	0.12 ± 0.13	3.72 ± 4.56	87.12 ± 6.15	85.94	65.87	41.00	41.60
CLAIM	87.60 ± 5.98	99.16 ± 1.34	0.05 ± 0.07	1.50 ± 2.53	91.51 ± 4.50	89.14	77.49	60.12	59.75
DeepLearnAI	82.70 ± 20.38	98.74 ± 1.80	0.18 ± 0.23	5.40 ± 9.77	66.91 ± 39.05	66.67	70.65	29.17	35.81
DKFZ	89.07 ± 5.63	99.06 ± 1.40	0.10 ± 0.11	2.48 ± 3.40	86.16 ± 5.62	84.45	63.85	48.81	41.98
DLaBella29*	84.63 ± 10.18	96.13 ± 5.44	0.08 ± 0.10	6.18 ± 7.82	75.14 ± 5.22	71.43	50.67	37.80	50.17
HITSZ	89.36 ± 5.45	99.04 ± 1.44	0.09 ± 0.11	2.46 ± 3.39	90.04 ± 2.59	85.94	71.30	53.27	52.60
IMR	87.88 ± 6.88	98.43 ± 2.23	0.09 ± 0.13	2.57 ± 3.66	91.23 ± 2.03	90.62	69.24	52.31	49.05
junqiangchen	77.06 ± 7.01	98.06 ± 1.98	0.33 ± 0.26	8.67 ± 5.96	49.97 ± 29.93	48.96	44.13	8.33	7.09
NantesU*	79.53 ± 23.84	98.33 ± 2.02	0.21 ± 0.50	5.52 ± 9.02	84.51 ± 3.18	74.55	66.57	42.56	36.00
NIC-VICOROB	80.26 ± 22.14	98.79 ± 1.59	0.08 ± 0.11	6.40 ± 8.58	77.44 ± 14.13	60.64	71.45	35.79	50.41
NIC-VICOROB*	89.58 ± 5.44	98.96 ± 1.50	0.03 ± 0.05	2.31 ± 3.30	89.86 ± 7.51	82.96	72.31	52.38	61.79
pamaad	77.64 ± 22.41	98.65 ± 1.92	0.19 ± 0.21	7.54 ± 10.00	78.11 ± 6.56	71.43	62.42	27.98	39.33
SynthCLAIM*	60.23 ± 22.83	83.52 ± 28.37	0.72 ± 0.44	16.78 ± 25.33	49.37 ± 28.87	52.60	46.87	2.60	3.03
UZH	89.35 ± 5.53	98.58 ± 2.34	0.03 ± 0.05	2.60 ± 3.62	91.74 ± 3.60	81.92	82.96	60.19	70.78
Top-3 ensemble	90.12 ± 5.72	99.05 ± 1.68	0.06 ± 0.09	1.94 ± 3.14	92.13 ± 4.53	83.93	88.70	57.44	61.35
Orgs ensemble	90.08 ± 5.80	99.03 ± 1.51	0.09 ± 0.11	1.96 ± 3.09	88.18 ± 6.62	88.02	65.37	49.78	47.95

Table S6. Results (mean ± standard deviation) of the CoW ROI detection task from 2024 on the TopCoW internal test data in terms of intersection over union (IoU) and boundary IoU. The top three values for each metric are marked in gold, silver and bronze colors. If a team only submitted to one of the tracks the columns of the other track are filled with a '-'.

CoW ROI detection performance on the 2024 TopCoW internal test data

	CTA (n=70)		MRA (n=70)	
Team	IoU (%) ↑	Boundary IoU (%) ↑	IoU (%) ↑	Boundary IoU (%) ↑
ARG	74.20 ± 7.08	61.93 ± 9.53	80.30 ± 6.08	70.08 ± 8.93
CLAIM	72.45 ± 7.79	59.61 ± 10.40	76.58 ± 6.34	65.12 ± 8.62
IMR	74.19 ± 7.63	61.86 ± 10.26	77.40 ± 9.97	66.77 ± 11.43
junqiangchen	72.13 ± 9.59	59.95 ± 11.61	1.06 ± 2.27	0.96 ± 1.59
NIC-VICOROB	76.42 ± 6.52	65.14 ± 8.80	81.54 ± 6.77	71.97 ± 9.56
pamaad	33.59 ± 6.54	16.24 ± 6.48	76.58 ± 9.46	65.63 ± 11.30
UB-VTL	63.14 ± 11.20	48.05 ± 13.47	-	-
UZH	79.31 ± 7.52	69.47 ± 10.38	84.84 ± 5.11	77.08 ± 7.27

Table S7. Results of the CoW variant graph classification task from 2024 on the TopCoW internal test set in terms of variant-balanced accuracy (VarBalAcc) for both the anterior and the posterior variant. The top three values for each metric are marked in gold, silver and bronze colors. Teams marked with * were segmentation-based and # were classification-based submissions.

Variant classification performance on the 2024 TopCoW internal test data				
	CTA (n=70)		MRA (n=70)	
Team	Anterior VarBalAcc (%) ↑	Posterior VarBalAcc (%) ↑	Anterior VarBalAcc (%) ↑	Posterior VarBalAcc (%) ↑
ARG#	24.66	17.79	25.52	19.31
CLAIM*	73.94	88.76	89.14	74.60
IMR*	87.28	62.17	86.46	69.24
junqiangchen#	25.00	19.05	25.00	21.12
pamaad#	25.00	16.14	25.00	20.12
UZH#	36.26	39.75	37.28	22.47

Table S8. Results (mean ± standard deviation) of the CoW multiclass segmentation on the external test data for the top 6 teams in terms of class-average Dice similarity coefficient, centerline Dice (clDice) on merged binary mask, class-average 0-th Betti number (β_0) error, class-average Hausdorff Distance 95% Percentile (HD95), average F1 score for detection of communicating arteries and 3rd-A2, variant-balanced accuracy (VarBalAcc) of graph classification, and CoW variant topology match rate (TMR) for both the anterior and posterior variants. The arrow indicates the favorable direction. The top three values for each metric are marked in gold, silver, and bronze colors. The bottom two rows for each dataset are the results for the top-3 teams ensemble and the organizer teams ('UZH', 'DKFZ', 'HITSZ') ensemble. Team 'CLAIM' used additional training data which included some external MRA test images from Lausanne and IXI-HH without our ground truth labels.

ISLES CTA multiclass segmentation performance (n=26)

Team	Per case class-avg Dice (%) ↑	clDice (%) ↑	Per case class-avg β_0 error ↓	Per case class-avg HD95 ↓	Average F1 score (%) ↑	Anterior VarBalAcc (%) ↑	Posterior VarBalAcc (%) ↑	Anterior TMR (%) ↑	Posterior TMR (%) ↑
CLAIM	85.23 ± 4.64	98.69 ± 1.65	0.07 ± 0.09	2.94 ± 3.62	93.23 ± 9.56	80.67	84.17	21.85	76.11
DKFZ	89.00 ± 5.48	98.94 ± 0.97	0.09 ± 0.09	3.16 ± 4.47	91.65 ± 6.85	78.57	88.33	47.69	59.72
HITSZ	87.72 ± 5.46	98.24 ± 2.32	0.11 ± 0.09	2.97 ± 4.27	92.12 ± 6.11	84.24	86.11	51.89	54.72
IMR	88.39 ± 4.93	98.66 ± 1.20	0.09 ± 0.11	3.26 ± 3.83	90.50 ± 5.51	88.45	84.72	35.50	71.39
NIC-VICOROB	90.72 ± 4.78	98.65 ± 1.57	0.04 ± 0.05	2.55 ± 4.00	68.01 ± 39.39	92.02	89.72	65.55	80.83
UZH	89.98 ± 4.89	98.70 ± 1.27	0.04 ± 0.06	2.53 ± 3.33	94.76 ± 5.89	89.08	87.78	62.61	73.33
Top-3 ensemble	91.51 ± 4.46	99.04 ± 0.99	0.06 ± 0.08	2.13 ± 3.15	94.17 ± 6.82	88.45	91.11	60.50	78.89
Orgs ensemble	89.97 ± 4.55	98.96 ± 0.98	0.08 ± 0.08	2.53 ± 3.55	93.07 ± 6.32	82.14	89.72	51.26	61.11

LargeIA CTA multiclass segmentation performance (n=20)

Team	Per case class-avg Dice (%) ↑	clDice (%) ↑	Per case class-avg β_0 error ↓	Per case class-avg HD95 ↓	Average F1 score (%) ↑	Anterior VarBalAcc (%) ↑	Posterior VarBalAcc (%) ↑	Anterior TMR (%) ↑	Posterior TMR (%) ↑
CLAIM	81.88 ± 6.44	98.79 ± 1.39	0.11 ± 0.13	4.21 ± 4.14	73.68 ± 16.17	89.29	63.73	33.33	49.84
DKFZ	85.56 ± 5.71	98.39 ± 1.97	0.17 ± 0.18	5.31 ± 5.11	74.51 ± 9.88	58.33	45.40	25.00	23.17
HITSZ	84.82 ± 7.57	98.50 ± 1.52	0.13 ± 0.20	4.53 ± 5.69	71.72 ± 18.91	61.90	53.02	48.81	30.08
IMR	84.43 ± 6.86	97.67 ± 2.55	0.15 ± 0.12	6.42 ± 5.50	71.57 ± 17.85	67.86	38.02	16.67	32.46
NIC-VICOROB	88.17 ± 5.70	98.55 ± 1.62	0.04 ± 0.05	3.38 ± 4.48	82.25 ± 10.57	72.62	57.78	46.43	49.44
UZH	86.14 ± 6.76	97.75 ± 1.97	0.03 ± 0.05	3.59 ± 4.57	77.84 ± 17.02	78.57	44.68	71.43	24.13
Top-3 ensemble	87.95 ± 6.32	98.48 ± 1.62	0.09 ± 0.11	4.49 ± 5.18	76.89 ± 16.89	64.29	47.06	44.05	43.73
Orgs ensemble	87.27 ± 6.63	98.36 ± 1.66	0.12 ± 0.15	4.14 ± 5.14	75.13 ± 15.66	64.29	55.40	46.43	32.06

Lausanne MRA multiclass segmentation performance (n=20)

Team	Per case class-avg Dice (%) ↑	clDice (%) ↑	Per case class-avg β_0 error ↓	Per case class-avg HD95 ↓	Average F1 score (%) ↑	Anterior VarBalAcc (%) ↑	Posterior VarBalAcc (%) ↑	Anterior TMR (%) ↑	Posterior TMR (%) ↑
ARG	87.98 ± 5.83	99.02 ± 1.09	0.10 ± 0.13	3.69 ± 4.63	83.23 ± 9.59	67.86	72.22	46.43	46.83
CLAIM	88.38 ± 5.33	98.89 ± 1.40	0.03 ± 0.05	2.16 ± 4.21	96.46 ± 4.55	91.67	85.56	60.71	83.17
DKFZ	88.77 ± 6.22	98.98 ± 1.16	0.09 ± 0.11	3.32 ± 5.26	82.54 ± 10.63	72.62	68.89	29.76	45.87
HITSZ	88.24 ± 5.78	98.79 ± 1.40	0.12 ± 0.12	3.62 ± 4.60	83.20 ± 11.04	76.19	68.89	35.71	43.49
IMR	89.26 ± 5.48	98.64 ± 1.76	0.08 ± 0.11	3.22 ± 4.51	89.15 ± 7.19	90.48	89.68	47.62	73.81
UZH	89.69 ± 5.62	98.71 ± 1.42	0.03 ± 0.05	3.33 ± 4.63	90.91 ± 5.41	84.52	91.90	41.67	65.71
Top-3 ensemble	90.69 ± 5.51	98.98 ± 1.20	0.06 ± 0.10	2.12 ± 4.19	93.08 ± 6.08	89.29	91.11	63.10	65.71
Orgs ensemble	89.98 ± 5.97	98.84 ± 1.24	0.08 ± 0.09	2.90 ± 5.15	91.70 ± 6.39	72.62	74.44	29.76	43.49

IXI-HH MRA multiclass segmentation performance (n=20)

Team	Per case class-avg Dice (%) ↑	clDice (%) ↑	Per case class-avg β_0 error ↓	Per case class-avg HD95 ↓	Average F1 score (%) ↑	Anterior VarBalAcc (%) ↑	Posterior VarBalAcc (%) ↑	Anterior TMR (%) ↑	Posterior TMR (%) ↑
ARG	88.26 ± 4.91	98.10 ± 1.74	0.12 ± 0.09	3.39 ± 5.22	90.29 ± 7.38	75.00	85.71	34.38	30.71
CLAIM	87.62 ± 5.41	98.20 ± 1.63	0.06 ± 0.12	3.67 ± 6.16	91.47 ± 7.56	75.00	65.71	37.50	63.21
DKFZ	89.20 ± 5.49	98.38 ± 1.89	0.08 ± 0.12	2.64 ± 4.26	83.53 ± 10.81	62.50	94.29	21.88	41.79
HITSZ	88.45 ± 5.41	98.52 ± 1.60	0.11 ± 0.13	3.41 ± 5.22	87.89 ± 9.96	75.00	58.57	21.88	36.07
IMR	75.78 ± 23.29	90.17 ± 22.14	0.19 ± 0.24	10.41 ± 19.77	77.55 ± 17.15	68.75	79.29	39.06	49.29
UZH	89.95 ± 3.87	98.11 ± 2.18	0.03 ± 0.04	3.18 ± 3.77	89.72 ± 7.09	68.75	87.14	29.69	54.64
Top-3 ensemble	90.47 ± 5.47	98.70 ± 1.53	0.07 ± 0.11	2.49 ± 4.96	90.88 ± 6.99	75.00	71.43	34.38	48.93
Orgs ensemble	90.44 ± 5.40	98.58 ± 1.67	0.08 ± 0.13	2.22 ± 4.03	91.87 ± 6.64	75.00	94.29	21.88	51.79

Table S9. Results (mean ± standard deviation) of the CoW ROI detection task on the external test set for the top 4 teams in terms of intersection over union (IoU) and Boundary IoU. The top three values for each metric are marked in gold, silver and bronze colors.

Detection on external test data								
	ISLES CTA (n=26)		LargeIA CTA (n=20)		Lausanne MRA (n=20)		IXI-HH MRA (n=20)	
Team	IoU (%) ↑	Boundary IoU (%) ↑	IoU (%) ↑	Boundary IoU (%) ↑	IoU (%) ↑	Boundary IoU (%) ↑	IoU (%) ↑	Boundary IoU (%) ↑
ARG	79.07 ± 6.74	68.79 ± 9.36	76.67 ± 6.87	65.80 ± 9.23	77.71 ± 6.39	65.87 ± 9.14	77.23 ± 10.06	66.03 ± 13.72
IMR	62.51 ± 26.70	52.60 ± 25.33	64.30 ± 27.74	55.43 ± 25.86	76.85 ± 6.82	64.68 ± 9.84	73.06 ± 19.87	61.91 ± 20.09
NIC-VICOROB	80.27 ± 6.54	70.58 ± 9.03	80.74 ± 6.86	71.46 ± 9.55	80.08 ± 6.60	69.93 ± 9.56	83.59 ± 7.85	74.83 ± 11.02
UZH	85.85 ± 7.09	78.33 ± 10.31	89.83 ± 5.21	84.33 ± 7.58	89.70 ± 6.79	84.38 ± 9.53	93.14 ± 5.66	89.48 ± 8.42

Table S10. Results of the CoW variant graph classification task on the external test set for the top 4 teams in terms of variant-balanced accuracy (VarBalAcc) for both the anterior and the posterior variant. The top three values for each metric are marked as gold, silver and bronze colors. Teams marked with * were segmentation-based and # were classification-based submissions. Team 'CLAIM' used additional training data which included some external MRA test images from Lausanne and IXI-HH without our ground truth labels.

Variant graph classification on external data								
	ISLES CTA (n=26)		LargeIA CTA (n=20)		Lausanne MRA (n=20)		IXI-HH MRA (n=20)	
Team	Anterior VarBalAcc (%) ↑	Posterior VarBalAcc (%) ↑	Anterior VarBalAcc (%) ↑	Posterior VarBalAcc (%) ↑	Anterior VarBalAcc (%) ↑	Posterior VarBalAcc (%) ↑	Anterior VarBalAcc (%) ↑	Posterior VarBalAcc (%) ↑
CLAIM*	80.67	84.17	89.29	63.73	91.67	85.56	75.00	65.71
IMR*	86.97	84.72	67.86	34.68	88.10	89.68	68.75	81.79
junqiangchen#	25.00	14.44	33.33	17.62	33.33	9.05	25.00	18.21
UZH#	27.52	13.06	30.95	22.62	30.95	24.44	34.38	8.21

Table S11. Per case inference time in seconds (mean±std) for the CoW multiclass segmentation task on the external test set. The times are reported for the top 6 teams for each track separately. For teams outside the top 6 in a given track, inference times are not included and are indicated with a '-'.

Per case inference time of segmentation algorithms (seconds)				
Team	ISLES CTA	LargeIA CTA	Lausanne MRA	IXI-HH MRA
ARG	-	-	105.0 ± 17.0	65.5 ± 5.1
CLAIM	36.6 ± 4.8	39.9 ± 3.1	34.3 ± 3.9	30.0 ± 0.7
DKFZ	70.7 ± 11.4	75.5 ± 9.6	77.1 ± 14.1	44.1 ± 3.5
HITSZ	65.5 ± 18.0	76.5 ± 11.8	80.6 ± 14.1	56.9 ± 5.8
IMR	51.2 ± 9.0	62.1 ± 8.8	56.3 ± 15.0	32.2 ± 1.4
NIC-VICOROB	123.9 ± 21.8	129.0 ± 20.7	-	-
UZH	145.2 ± 25.1	139.6 ± 20.8	147.4 ± 19.9	93.5 ± 10.2

Table S12. Aneurysm locations given by the predicted CoW labels that the aneurysms overlap with or are adjacent to in the overlay of the masks. The locations in CoW predictions were evaluated for the top teams 'UZH', 'NIC-VICOROB', 'IMR', and 'CLAIM' for 12 cases from the LargeIA CTA dataset containing at least one aneurysm inside or near the CoW ROI. Wrong locations in predictions are marked in magenta color.

Aneurysm location in CoW predictions for LargeIA CTA					
Image	Aneurysm location	Location in CoW prediction			
		UZH	NIC-VICOROB	IMR	CLAIM
Tr0001	R-ICA	R-ICA	R-ICA	R-ICA	R-ICA
Tr0004	R-A2 origin near Acom-complex	R-ACA	R-ACA, L-ACA	R-ACA, L-ACA	R-ACA, L-ACA
Tr0005	Acom-ACA junction R-ACA side	Acom, R-ACA, L-ACA	Acom, R-ACA, L-ACA	Acom, R-ACA, L-ACA	R-ACA, L-ACA
Tr0006	L-ICA-Pcom junction	L-ICA, L-Pcom	L-ICA, L-Pcom	L-ICA, L-Pcom	L-ICA, L-Pcom
Tr0015	L-ICA	L-ICA	L-ICA	L-ICA	L-ICA
Tr0018	BA tip	BA, R-PCA	BA, R-PCA, L-PCA	BA, R-PCA, L-PCA	BA, R-PCA, L-PCA
Tr0019	1. R-ICA bifurcation 2. L-ICA	1. R-ICA 2. L-ICA	1. R-ICA, R-ACA 2. L-ICA	1. R-ICA 2. L-ICA, L-Pcom	1. R-ICA 2. L-ICA
Tr0024	Acom-ACA junction	Acom, R-ACA, L-ACA	Acom, R-ACA, L-ACA	Acom, R-ACA, L-ACA	Acom, R-ACA, L-ACA
Tr0025	1. BA trunk 2. BA tip	1. BA 2. BA, L-PCA	1. BA 2. BA, L-PCA	1. BA 2. BA, L-PCA	1. BA 2. BA, R-PCA, L-PCA
Tr0038	L-ICA-Pcom junction	L-ICA, L-Pcom	L-ICA, L-Pcom	L-ICA, L-Pcom	L-ICA, L-Pcom
Tr0069	L-MCA bifurcation	L-MCA	L-MCA	L-MCA	L-MCA
Tr0089	R-MCA bifurcation	R-MCA	R-MCA	R-MCA	R-MCA

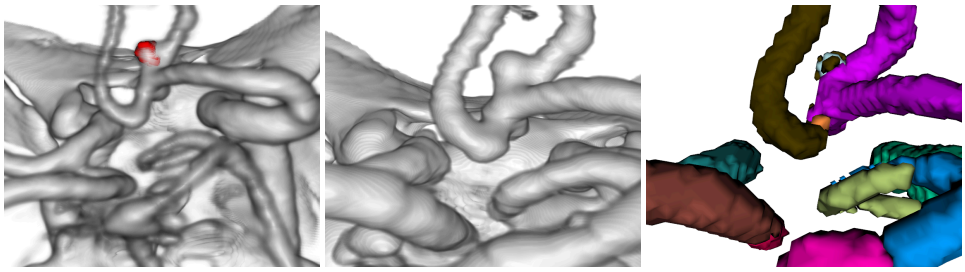

Fig. S3. Tr0004 contains a common mistake for locating the aneurysm due to the tortuous ACAs and the presence of another infundibulum near the Acom-complex. Left: Image overlaid with the ground-truth aneurysm mask in red; Middle: Zoomed-in view of the image; Right: Aneurysm mask in silver color overlaid with predicted CoW masks from team 'CLAIM'.

Table S13. Binary segmentation task results (mean ± standard deviation) from 2023 for MRA and CTA in Dice scores, centerline Dice (clDice) scores and errors in the zero-th Betti number β_0. The arrow indicates the favorable direction. The top three values for each metric and each track are marked as gold, silver and bronze cells in decreasing order. A '*' behind the team name indicates that the segmentation predictions have been converted from the multiclass submissions and inserted here. If a team only submitted to one of the tracks the columns of the other track are filled with a '-'.

2023 binary segmentation performance						
	MRA (n=35)			CTA (n=35)		
Team	Dice (%) ↑	clDice (%) ↑	β_0 error ↓	Dice (%) ↑	clDice (%) ↑	β_0 error ↓
2i_mtl	-	-	-	44.15 ± 11.94	48.04 ± 11.23	28.77 ± 15.12
agaldran	94.33 ± 2.64	96.68 ± 2.79	1.57 ± 1.29	87.73 ± 3.26	95.99 ± 2.97	2.26 ± 1.82
DKFZ	95.54 ± 2.72	98.31 ± 2.14	0.37 ± 0.60	92.16 ± 2.76	98.42 ± 1.83	0.57 ± 0.81
EURECOM	93.84 ± 2.69	94.42 ± 3.10	3.77 ± 2.79	84.79 ± 4.34	89.92 ± 5.43	8.14 ± 4.66
gbCoW	94.95 ± 2.89	98.10 ± 2.33	0.86 ± 0.94	-	-	-
gl*	93.67 ± 4.84	95.81 ± 5.11	0.71 ± 0.86	70.94 ± 24.88	75.80 ± 28.05	2.09 ± 1.48
HITSZ	94.05 ± 2.91	97.29 ± 2.29	0.69 ± 0.68	92.28 ± 2.83	97.70 ± 2.54	0.77 ± 0.97
IWantToGoToCanada	-	-	-	90.06 ± 2.67	96.56 ± 2.65	1.40 ± 0.88
junqiangchen	94.09 ± 2.08	96.86 ± 2.80	3.20 ± 2.96	89.49 ± 2.87	95.91 ± 1.63	3.91 ± 2.84
lWM	94.39 ± 2.48	97.39 ± 2.76	1.43 ± 1.36	91.07 ± 2.50	97.07 ± 2.60	1.74 ± 1.27
NIC-VICOROB	95.60 ± 2.32	98.26 ± 2.20	0.60 ± 0.60	92.07 ± 3.54	97.93 ± 2.47	0.60 ± 0.85
NIC-VICOROB-2	93.13 ± 3.67	94.30 ± 7.03	17.69 ± 57.63	89.68 ± 3.87	95.45 ± 3.28	3.80 ± 2.74
refrain	93.87 ± 2.02	98.20 ± 1.84	0.97 ± 1.20	-	-	-
sjtu_eiee_2-426lab	-	-	-	92.75 ± 3.19	97.81 ± 2.66	0.40 ± 0.55
UB-VTL	91.78 ± 2.41	93.27 ± 3.50	0.91 ± 0.98	72.27 ± 6.89	77.49 ± 7.92	2.37 ± 1.72
UW	95.37 ± 2.20	98.18 ± 1.93	0.89 ± 0.87	-	-	-
UZH*	95.14 ± 2.90	98.06 ± 2.30	0.57 ± 0.65	90.25 ± 5.73	96.98 ± 3.48	1.14 ± 1.44
ysato	88.05 ± 4.94	91.99 ± 3.60	2.60 ± 1.77	-	-	-

Table S14. Results (mean ± standard deviation) of the multiclass segmentation task from 2023 for MRA and CTA in terms of class-average Dice scores, centerline Dice (clDice) scores and class-average errors in the zero-th Betti number β_0. The *clDice* scores were computed on the merged binary class, the *Dice* scores and β_0 errors were computed for each class separately and the average was taken per case. The arrow indicates the favorable direction. The top three values for each metric and each track are marked as gold, silver and bronze cells in decreasing order. If a team only submitted to one of the tracks the columns of the other track are filled with a '-'.

2023 multiclass segmentation performance						
	MRA (n=35)			CTA (n=35)		
Team	Per case class-average Dice (%) ↑	clDice (%) ↑	Per case class-average β_0 error ↓	Per case class-average Dice (%) ↑	clDice (%) ↑	Per case class-average β_0 error ↓
agaldran	0.01 ± 0.07	0.16 ± 0.69	1.01 ± 0.18	0.15 ± 0.66	0.77 ± 1.62	1.07 ± 0.32
DKFZ	84.58 ± 6.47	97.21 ± 3.37	0.06 ± 0.06	83.32 ± 5.65	98.12 ± 1.86	0.11 ± 0.11
gbCoW	80.51 ± 14.69	98.10 ± 2.33	0.28 ± 0.27	-	-	-
gl	81.27 ± 9.16	95.81 ± 5.11	0.09 ± 0.08	54.02 ± 21.76	75.80 ± 28.05	0.24 ± 0.22
HITSZ	83.76 ± 6.95	97.29 ± 2.29	0.14 ± 0.14	81.67 ± 6.81	97.78 ± 1.88	0.16 ± 0.14
IWantToGoToCanada	-	-	-	74.34 ± 7.90	97.61 ± 1.78	0.77 ± 0.50
junqiangchen	71.56 ± 9.72	96.86 ± 2.80	0.87 ± 0.41	67.68 ± 5.97	95.91 ± 1.63	1.09 ± 0.37
lWM	79.03 ± 8.75	96.23 ± 3.11	0.27 ± 0.22	72.49 ± 8.46	92.78 ± 5.13	0.32 ± 0.19
NIC-VICOROB	77.13 ± 20.15	96.31 ± 5.78	0.11 ± 0.09	51.84 ± 31.11	97.86 ± 2.38	0.43 ± 0.38
NIC-VICOROB-2	65.99 ± 8.76	94.12 ± 6.12	2.64 ± 3.00	62.41 ± 8.22	95.65 ± 3.38	2.31 ± 1.37
refrain	83.72 ± 5.79	98.20 ± 1.84	0.19 ± 0.17	-	-	-
sjtu_eiee_2-426lab	-	-	-	67.46 ± 27.78	97.15 ± 3.22	0.23 ± 0.27
UZH	83.98 ± 7.33	98.06 ± 2.30	0.16 ± 0.13	77.00 ± 11.95	96.98 ± 3.48	0.23 ± 0.26

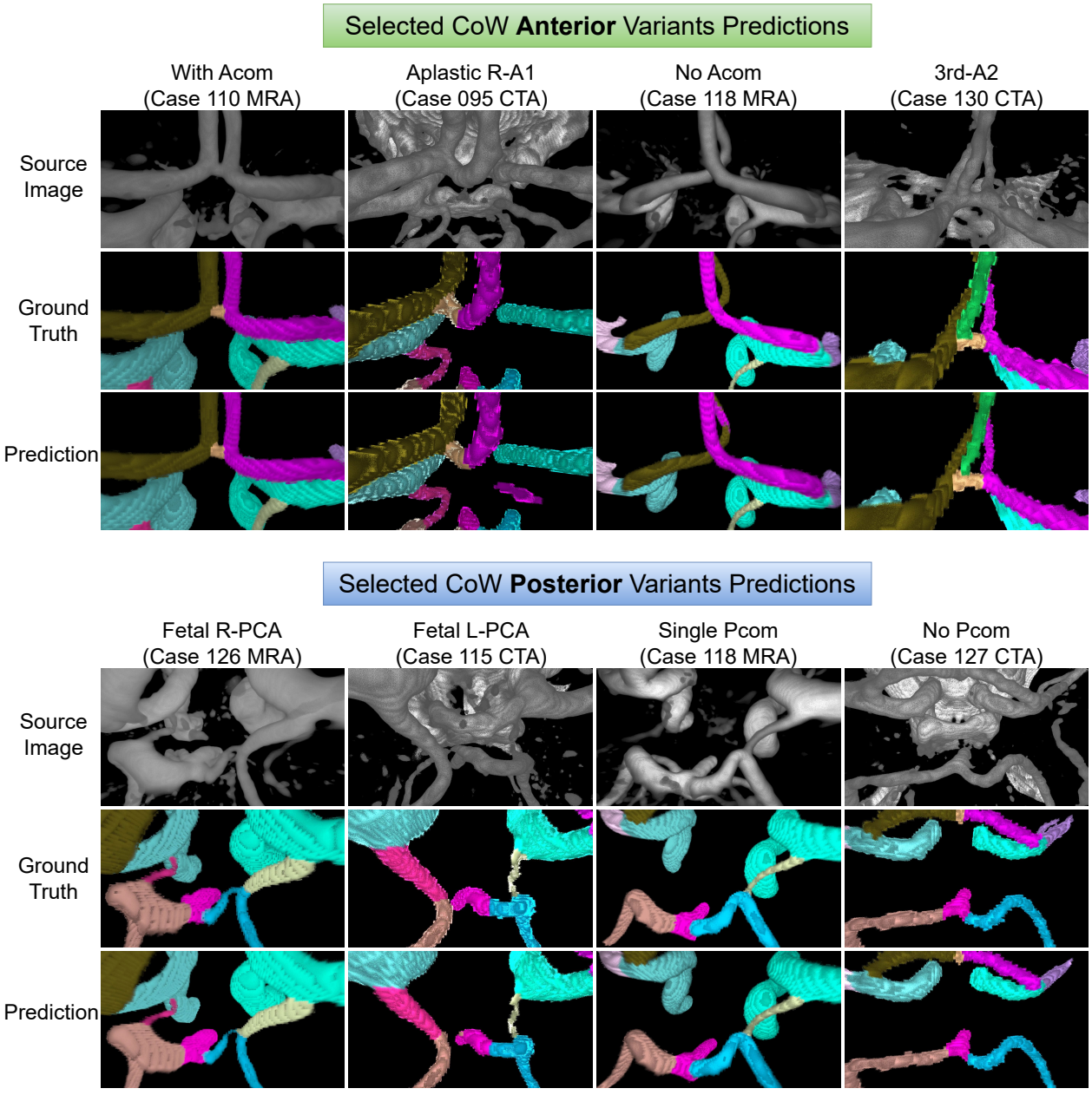

Fig. S4. More qualitative results for anterior (top sub-figure) and posterior (bottom sub-figure) variants in multiclass segmentation task. The predictions are by 2023 team 'DKFZ/WilliWillsWissen'. Alternating columns showcasing MRA and CTA.