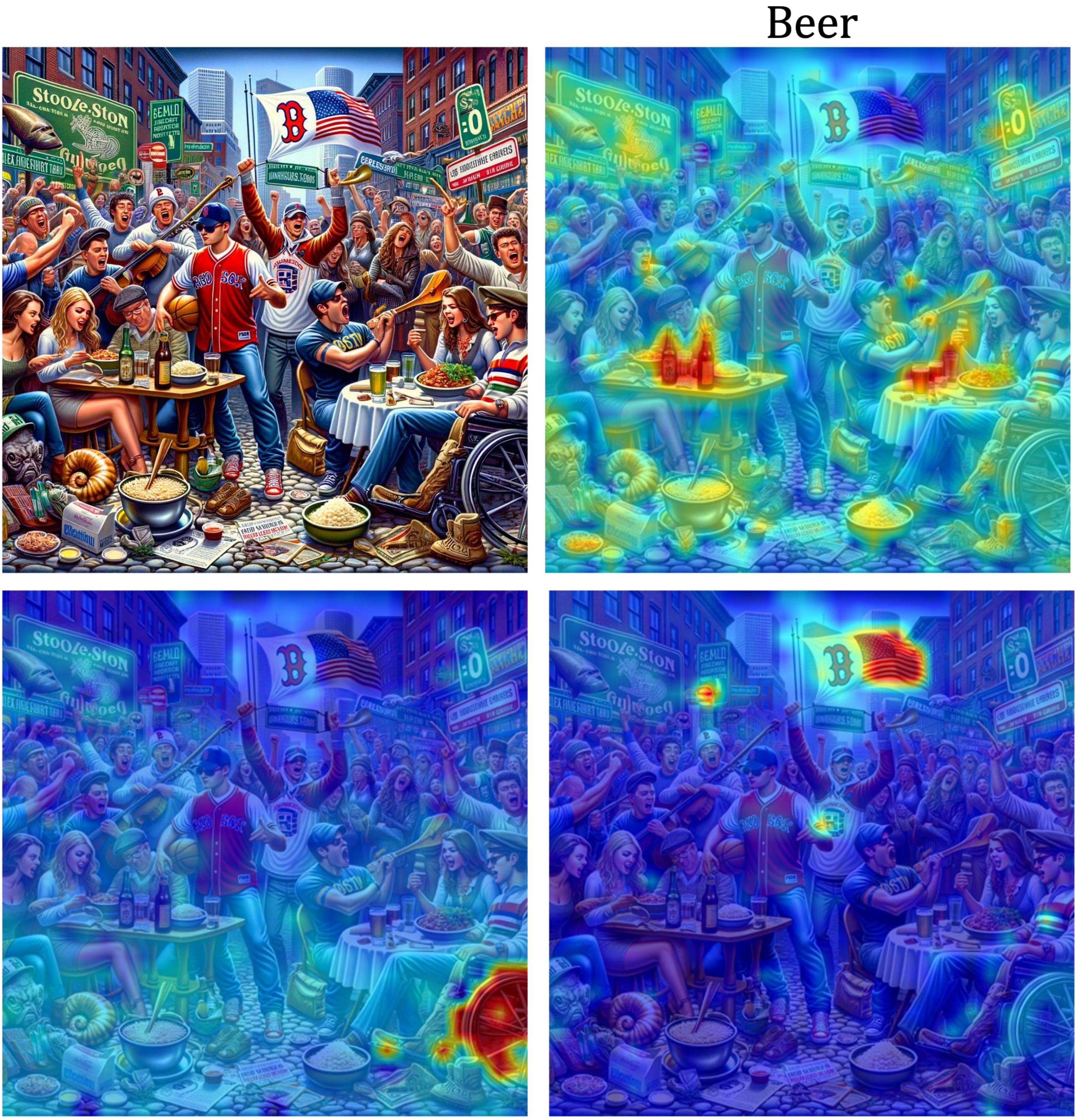
January 2024

Computer Vision News

& Medical Imaging News

The Magazine of the Algorithm Community



A publication by
RSIP VISION

Estimating Generic 3D Room Structures from 2D Annotations

Matias Valdenegro

Denys Rozumnyi is a fifth-year PhD student at ETH Zurich under the supervision of Marc Pollefeys.

In his research, he works on object-centric 3D reconstruction, motion estimation, deblurring, and detection. In particular, he focused a lot on motion-blurred objects.

by Denys Rozumnyi

Estimating **3D structures from video** is one of the key tasks in computer vision. The key ingredient of state-of-the-art methods are large annotated datasets for training. Some datasets even offer 3D room layout, which is defined as a set of 3D structural elements such as wall, floor, and ceiling, but there are just a few real datasets with annotated 3D room layouts. However, they require the images/videos to be acquired with special devices or sensors such as RGB-D cameras or panoramic captures.

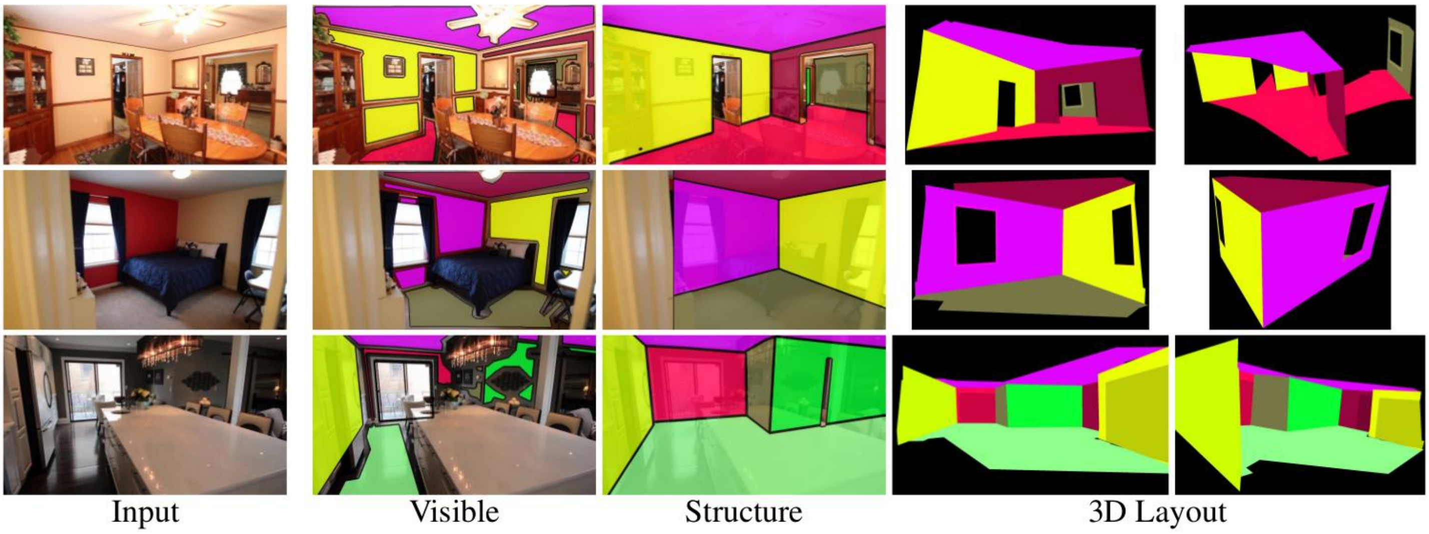


Figure 1: **From 2D annotations to 3D room layouts.** Given an RGB video, we ask human annotators to draw amodal segmentation masks and visible parts of each structural element, *e.g.* wall, floor. Then, our method automatically derives high-quality 3D layouts from these 2D annotations.

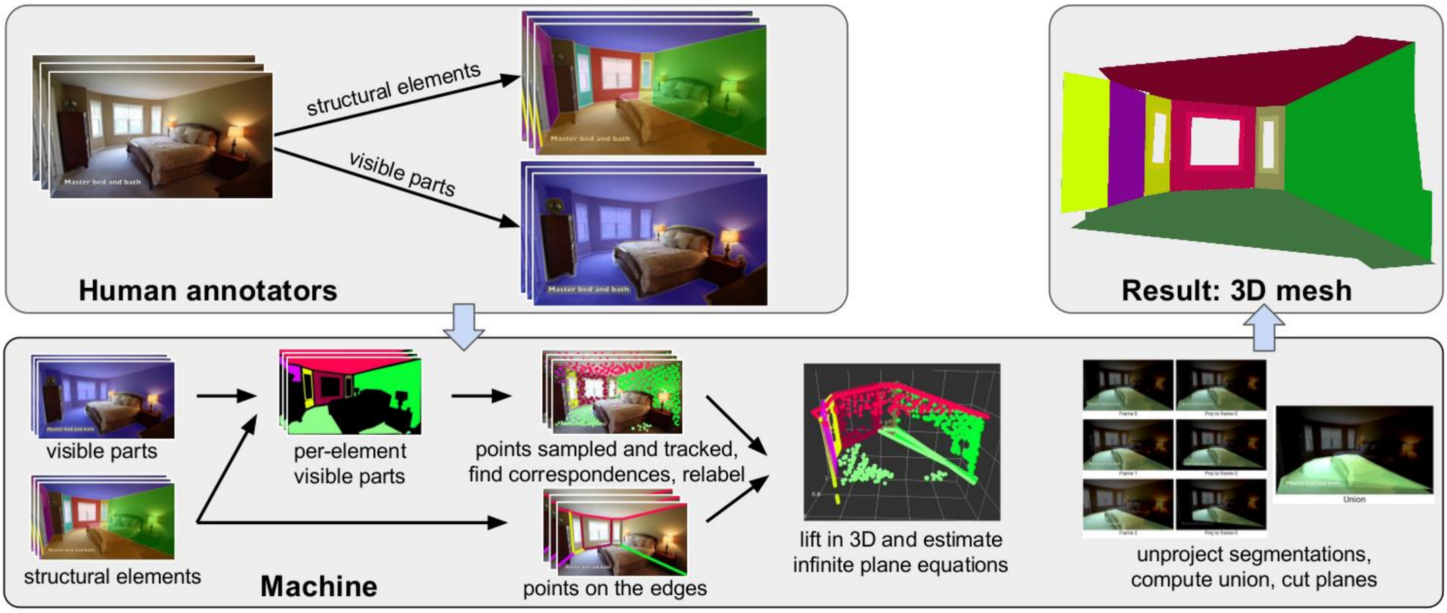


Figure 2: **Pipeline overview.** For each selected video frame independently, humans annotate segmentation masks for structural elements and their visible parts. We track 2D points on visible parts and use the 3D geometry induced by them to estimate 3D plane equations for the structural elements. Finally, we estimate the spatial extent of each structural element as a union of unprojected 2D annotations across all frames.

In contrast, we propose the first method for annotating **generic 3D room layouts** on commonly available RGB videos. We ask human annotators for an easy task: draw an amodal segmentation mask for each structural element **in 2D**, and also approximately mark its visible parts **in 2D**. Moreover, all annotation is performed on each video frame **independently** without requiring the annotator to provide correspondences among video frames.

The simplification of the annotator task is enabled by shifting much of the work to an automatic method we propose, capable of deriving 3D room layouts from these 2D annotations. This method estimates a 3D plane equation for each structural element, as well as a finite spatial extent that captures all parts of the plane that are in the camera's field-of-view at any time during the video. We estimate all elements jointly and connect adjacent elements at the right contact edges in 3D, matching their shared edge as observed in the 2D annotations.

Using the proposed approach, we annotate 2246 scenes from the RealEstate10k dataset that contains YouTube videos of indoor scenes. The rooms are complex and cover generic types, not limited to Cuboid/Manhattan. They can even be composite, such as two rooms connected by a door or a staircase. The method also works when the video does not show the full room (a common case).

We evaluate the quality of the produced 3D layout annotations on RealEstate10k in terms of Reprojection Intersection-over-Union to the input ground-truth 2D segmentation masks, and on ScanNet, which enables measuring 3D depth errors as it features ground-truth depth maps. We find the 3D layouts to be highly accurate, with only about 20cm depth error (out of 7-meter wide scenes), and high Reprojection IoU around 0.9. We also manually inspected the 3D layouts, finding that we reconstructed correctly about 98 % of all structural elements that are actually in the scene. We have released the dataset publicly at https://github.com/google-research/cad-estate

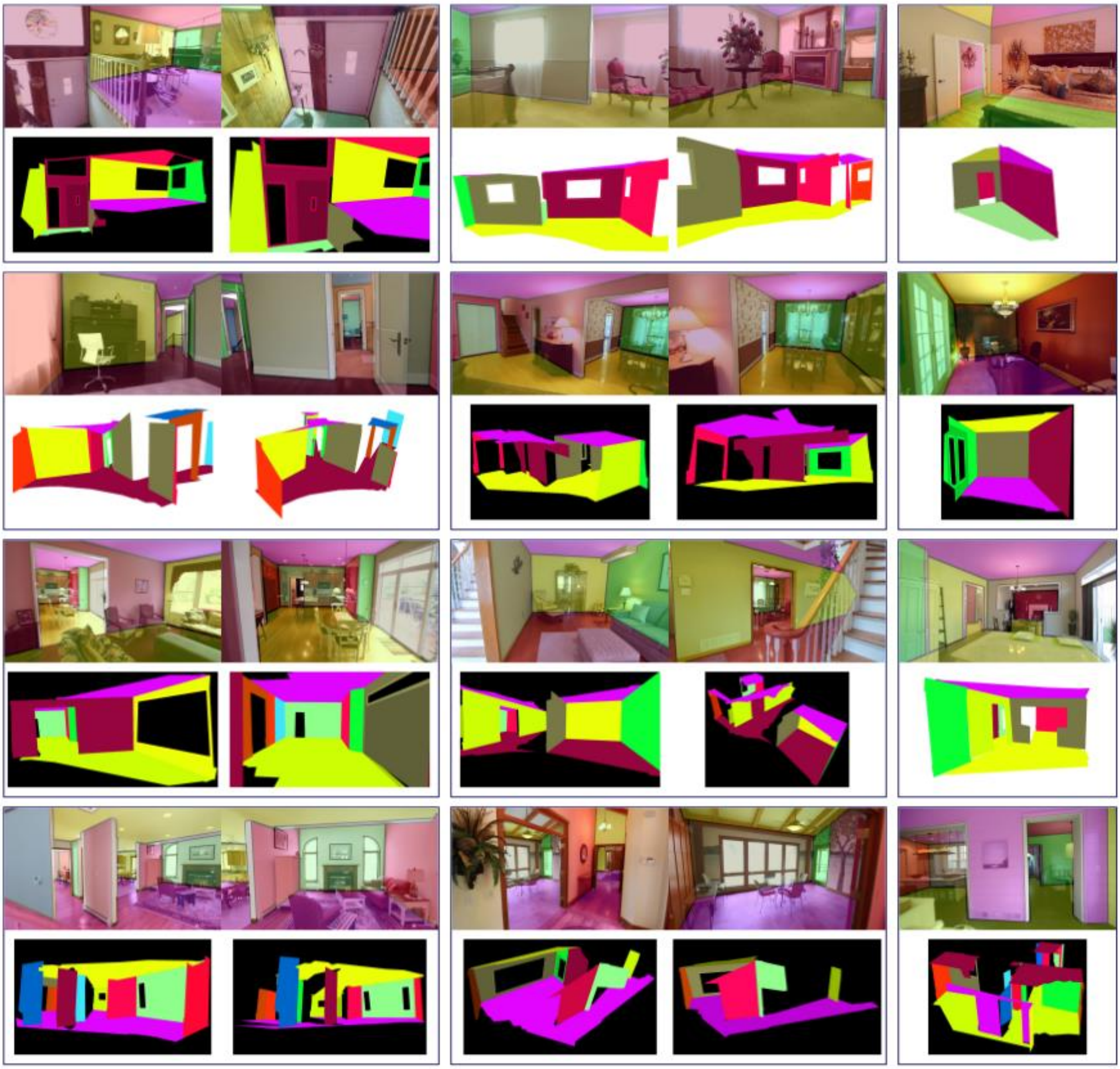
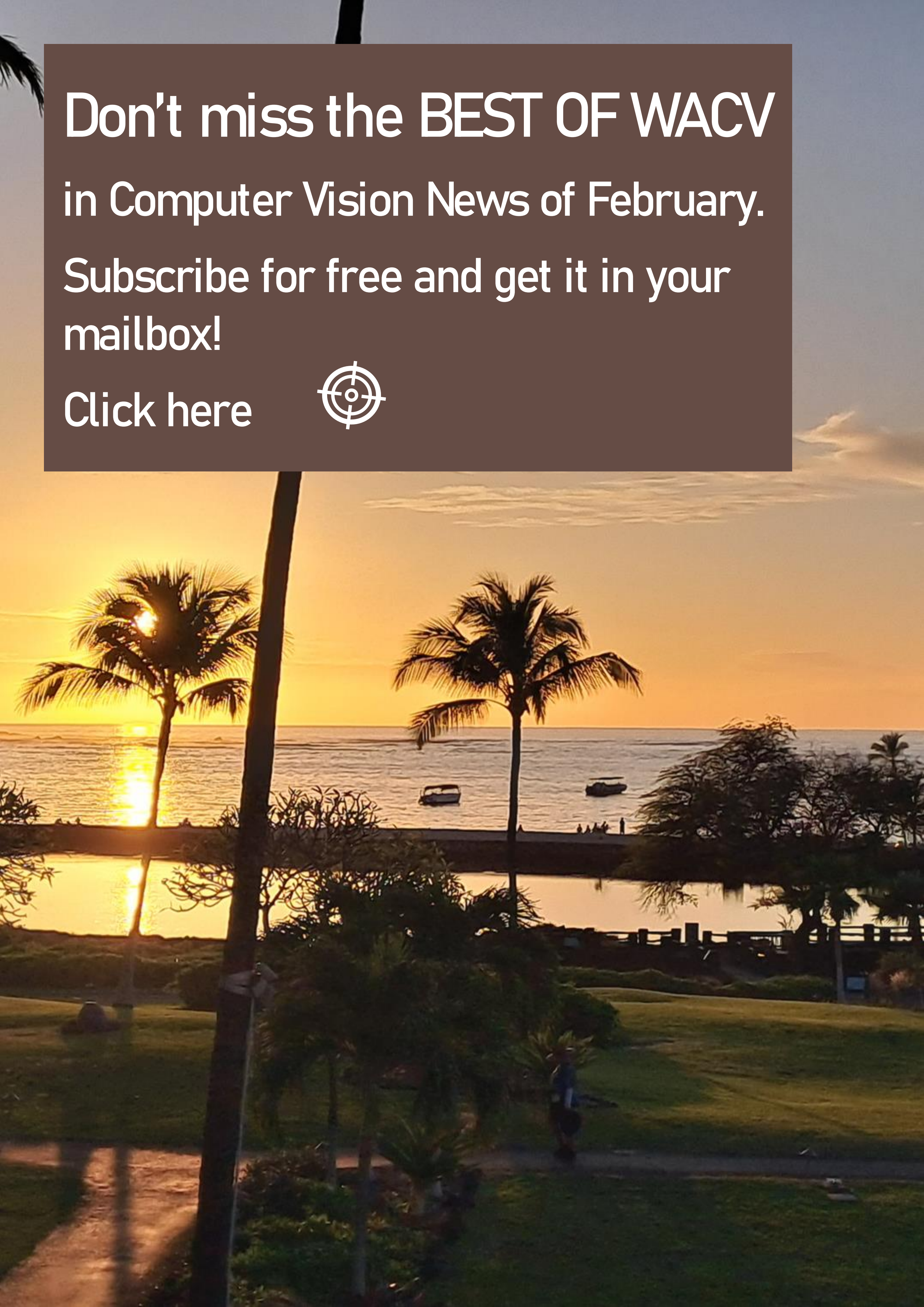


Figure 6: **3D Room layout reconstruction examples.** For each example scene, we show one or two of the input annotations overlaid on the video frame (top), and the final reconstructed layout (bottom).



Grounding Everything: Emerging Localization Properties in Vision-Language Transformers

Walid Bousselham is a PhD student at the University of Bonn under the supervision of Hilde Kuehne.

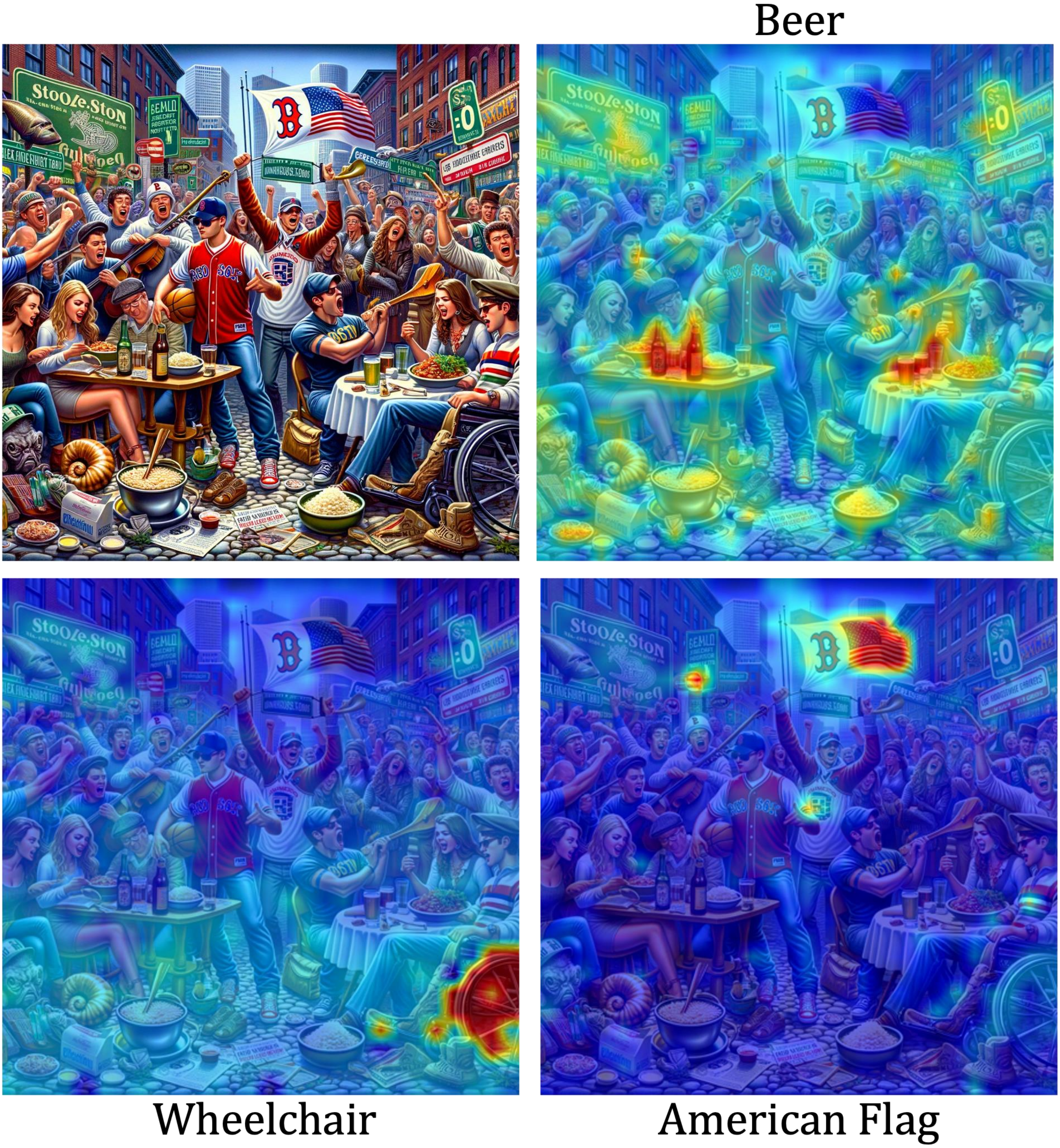
He speaks to us about his new paper proposing a Grounding Everything Module (GEM), which aims to adapt vision-language models to not only identify objects within an image but also localize them.

Vision-language models such as **CLIP** and **BLIP** have proven powerful at **zero-shot classification**, excelling in **recognizing objects within images**. However, when it comes to **localizing** these objects, traditional models fall short.

GEM's primary objective is to extend the capabilities of these models, adapting them for localization without disrupting the extensive vocabulary they have learned through pretraining on millions of images. *"One of the ways to do that is not to retrain the model but just to adapt the forward pass so that we're able to do localization without perturbing the weights,"* Walid explains. *"We introduced a module called self-self attention that essentially enables the clustering of the internal features of CLIP so that we can query the features later with text."*

GEM proves particularly useful for the popular computer vision task of **semantic segmentation**. Unlike classical methods with restricted sets of localizable classes, GEM is based on CLIP and can **segment and localize any object described with text**. Application of the technology could extend to **robotics**, with robots precisely locating specific objects based on textual prompts. This versatility opens new possibilities for industries relying on **object localization** for various tasks.

GEM's development was not without challenges. Given the emphasis on training-free adaptation, the project faced the issue of introducing **out-of-distribution features** when modifying internal features of the layers of the neural network. "*I always say that if it's easy, it's not research!*" Walid laughs. "*We took inspiration from another paper called **CLIP Surgery**, which built an alternative pathway. It has the original vision encoder, and then, in parallel, it builds another pathway that aggregates the information we need*



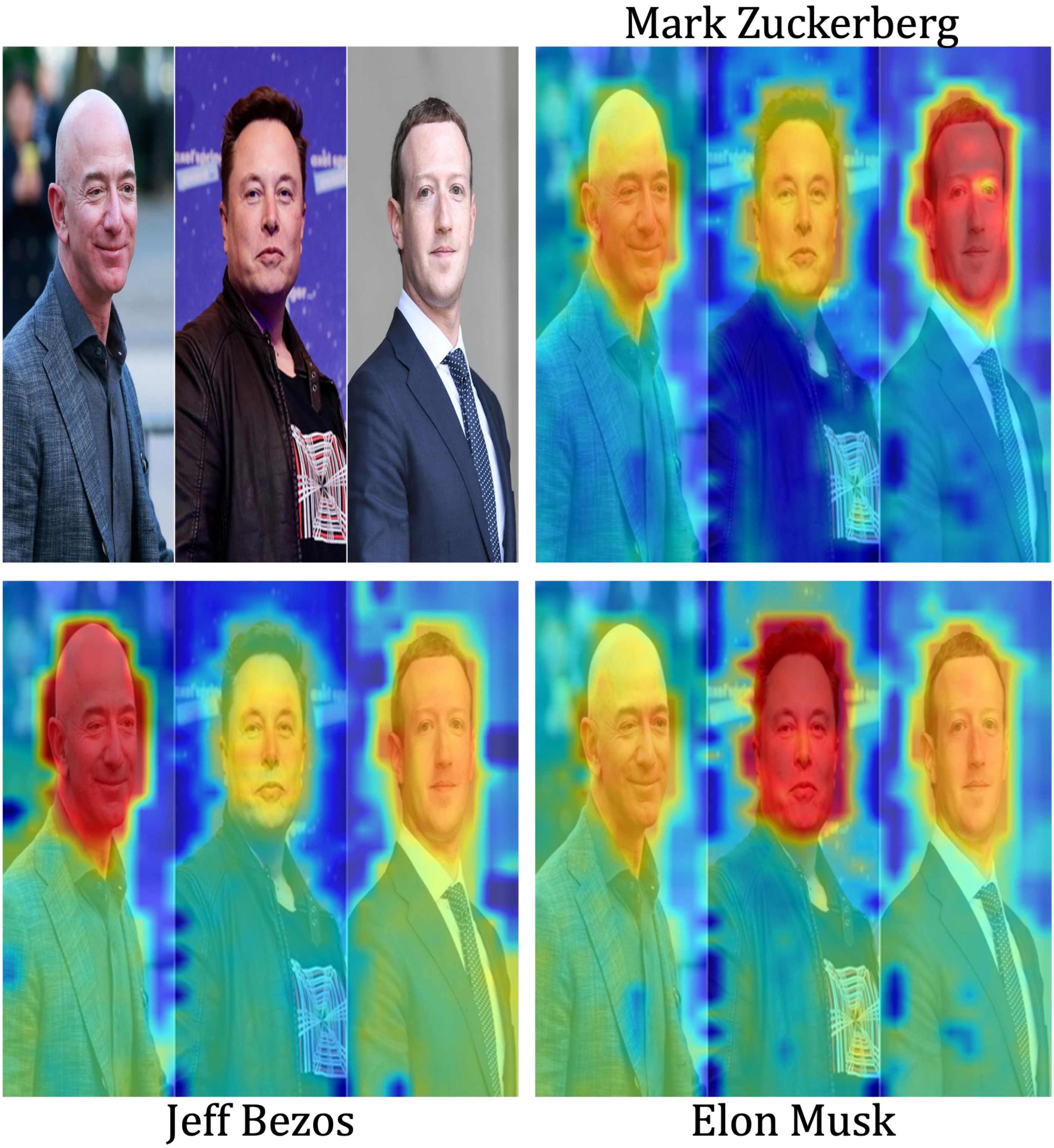
Beer

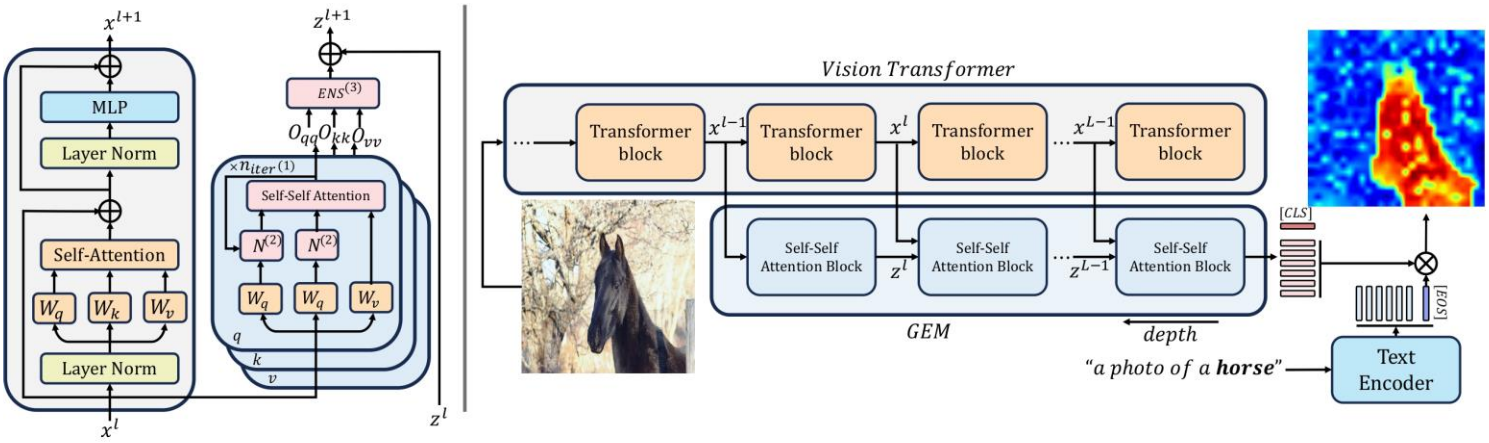
Wheelchair

American Flag

from the vision encoder. It is the same thing but with our self-self attention block. And it worked!"

Looking ahead, Walid plans to extend the scalability of GEM, acknowledging that **the current framework can scale to a large vision transformer model** but has limitations after that. The goal is to broaden the scope of this method, potentially applying it to different types of vision-language models.





He hopes that in the future, the model could be extended even further to include proper reasoning capabilities. However, its non-disruptive and user-friendly approach already makes GEM an **accessible and easy-to-use tool**.

"I released a Hugging Face demo where you can just upload the link of an image, or an image itself, and query it with text to see if ***it's working well for the type of image you uploaded****,"* he adds. *"I've seen on Twitter people are using this model to query on* ***the output of diffusion models or on animes****, and it's working pretty well because it's CLIP and was trained with these kinds of images. Also, there's the code on GitHub, and if you want to install the model, you pip install gem_torch, and then you're good to go!"*

As we wrap up the interview, Walid reveals that while he was born in France, his parents are from **Morocco**. His connection to Morocco was especially poignant during a recent visit when he witnessed the country's resilience in the wake of the severe September earthquake. *"I was there in October, right after the earthquake,"* he tells us. *"When I was there, it was almost a month, and I traveled throughout Morocco. It was doing way better."*

2 NeurIPS PAPERS in this issue!

Find them on page 2 (by Denys Rozumnyi)

and on page 38 (with Nina Montaña Brown)

Oana Ignat is a postdoctoral researcher at the University of Michigan.

She is a co-author on a paper which is presented as a poster at WACV 2024: “*Augment the Pairs: Semantics-Preserving Image-Caption Pair Augmentation for Grounding-Based Vision and Language Models*”.

I finished my PhD there last year in August and I continue my research on the intersection of computer vision and natural language processing. I build AI models and datasets that focus on human action understanding. That was my thesis work. Now, I’m extending this work towards analyzing model performance across demographics. Across different languages, different countries, income levels, and so on.

Does that include Romania?

Oh, yes! [*she laughs*] I always look for including Romania in my dataset. I really want to do more work on that area, for sure.

What is the goal of this research?

We want to see how foundation, state-of-the-art models, for example, CLIP – I have a recent paper that was presented at EMNLP in Singapore a few weeks ago – work across different demographics, because these models are usually made by research laboratories in Western countries and focus mostly on Western data. We want to see how well it performs across the world. In this paper, we look specifically at income. How this model works across different income levels. Images from households from different income levels in different countries. We found that there is a considerable gap in performance. The model performs much better on high-income versus low-income images.

How does this help us?

Well, we draw attention to this. First, we show that this gap in performance exists. This is important because it means that this model will not work well for images from those income levels or from those countries. We show that this performance is not uniform. It's not globally uniformly distributed. We want to show that we have to make sure that we train these models on data from different countries, from different demographics, and also include annotators. People who annotated data should be included from different countries and different income levels because we see that images look different. Even images from very common household items, like a toothbrush or refrigerator, we're used to seeing a certain image of that in Western countries, but they vary depending on demographics.

Do you like being a researcher?

Yeah, for sure. I really enjoy it. I gradually got here. Back in Romania, I worked a bit in industry. Like a software developer, but still focused on research applications. I really enjoyed that because you never know what you're going to get. It's research. It's work in progress. It's for discovering new things. That's what I enjoy. It's not predictable.

How long have you been in the States?

Quite long now. Already six years.

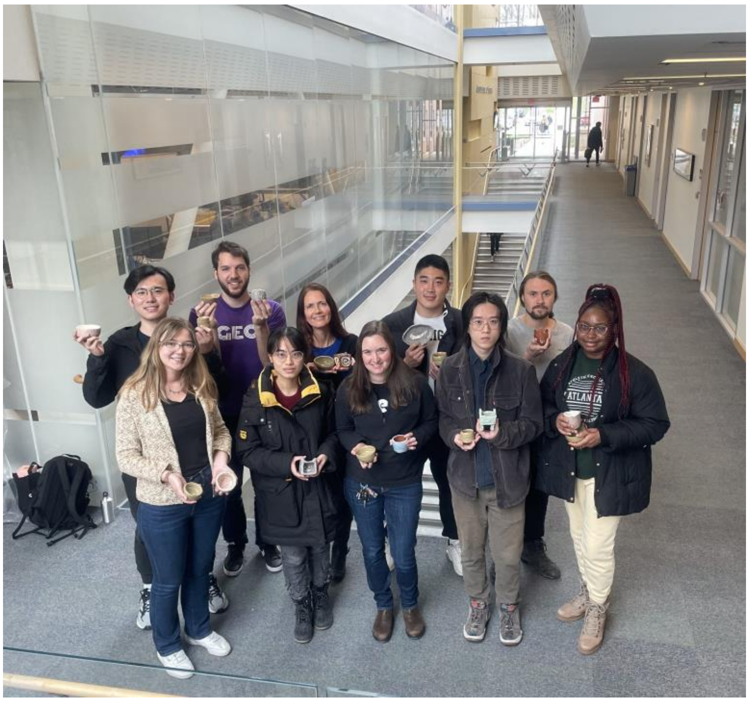
Out of everything you do, what is one thing that you would not be able to do if you were still in Romania?

That's an interesting question. I think there are much more opportunities here in the US. I can see the university is on board with any activity I propose to do and sponsor it. I feel like things are moving much faster here. Like people are really open to investing time and money in projects. Maybe in Romania, it's a bit more difficult to do that because the budget is stricter. There's more paperwork to do. Also, I find the optimism of people here is much more encouraging. When you have an idea, usually, people are open to it. They want to try it and are excited. That excitement is contagious, and I think leads to projects having more success and making more progress.

Is there a chance you will go back to Romania one day to work?

[*Oana laughs*] I'm constantly thinking about that, yes. Actually, when I started the PhD, I started with the thought in mind that I will come back after the PhD and teach, be a professor, and also collaborate with industry because I'm in the middle. I like the application side of industry and the fast progress, and I also like the research side, the academia side, where you're open to think of diverse problems. Now, I'm not sure. I want to stay a bit more in the US to explore more and have more opportunities, but I want to also collaborate with students from Romania, with professors from Romania, so I'm looking for that.

If one day you go back to Romania,



what is the one thing that you will bring back with you from the States?

Oh, I think many things! [*she laughs*] I hope so. Maybe one thing is this attitude that things are possible, that you have to stay optimistic and push hard for things to happen and be determined. I really value that.

I am going to ask you the opposite now – what did you bring to the Americans that they did not have before you came from Romania?

Oh, hmm. [*laughs*] I'm not sure. Let's see. Maybe some food! [*she laughs*] The national food, sarmale, maybe. Soup! I really miss the soup from back home. I wish in the US, there would be more soup places. [*laughs*] It's very healthy!

Where are you from in Romania?

I'm from Botoșani. It's a small city in the northeast of Romania.

In the direction of Ukraine?

Exactly, yes. We're at the border with Ukraine.

Wow. I am very fond of Odessa. Not very far from you guys.

Yes, yes. I hear stories about Odessa all the time from my parents, but I've never been there.

Were they in Odessa?

No, but maybe my father visited or some friends of his. I don't know, but I heard the name before.

With this dichotomy between industry and academia, is there a chance that you might find yourself in a position one day that bridges the two?

Yes, that's what's been on my mind. It took me a long time to decide what I wanted to do after the PhD program. Whether to go into industry or to go as a professor into academia. I had a few internships in industry, and I got to experience it also before the PhD. When I came back as a postdoc, I really enjoyed the mentoring aspect of postdoc. I think that was the last straw that made me decide to go into academia. I really like to mentor students and work with them. In industry, there's not much opportunity for that, or things are much more product-oriented. It would be nice to have maybe some collaboration between those two.

Can you see how that could happen? What would be the setting?

I know, at least in sabbaticals, there are professors who go to industry for a bit, maybe for a year or a semester. They're like advisors or research projects in industry. Yeah, that's an open possibility. I would like to try that to see how it is.

You have spoken to us about your current research. Are there projects coming up that you haven't told us about yet?

[*Oana laughs*] Yeah, there's always projects in the works. Now, I'm working with my mentee. We're working together on a multilingual dataset generated by ChatGPT. Large language models are very common, very popular nowadays, so it's very interesting to see how they work and their output. Especially, I'm very interested across languages. How their output differs. I'm really interested in the analysis of the data generated by GPT.

Would it be a safe bet to guess which languages you chose for your cross-language work?

Do you want to guess the language that I'm working on?

Would it be easy to guess them?

Oh, well, yeah, maybe! [*laughs*] Yes, I think so.

Okay, so English, Romanian.

Yes! [*she laughs*]

It was an easy guess.

But we have 10 languages. We have more than those two.

What is the goal of this research?

We're looking mostly at analyzing the data. We're looking at generating hotel reviews in these different languages. We want to compare the generated data with the real data and see if a model can easily distinguish between what is real and what is generated. I think this is very relevant for reviews because there are more and more automatically generated reviews out there on the internet, and we want to see if we can train the model to distinguish.

Because they are fake?

Yes, exactly. We want to catch the fake information.

Are you going to help the world to prevent scams, frauds, and cheating?

Yes, hopefully that will help. That will contribute to finding those and also analyzing what is different about this data.

Many reviews are written in very bad language actually with typos. People do not ever reread them. Is this noise an obstacle to your research?

That's a good point. We were thinking about this, actually. We were brainstorming what to consider when we generate this data. We noticed that GPT doesn't really generate this kind of data. It's actually the opposite. It's very polished. Yeah, the style is very formal. It doesn't really look like what a person would write. Maybe we should try to include some noise in the generation process.

I am jumping to a different subject. Is there an eminent Romanian scientist from the past that you admire?

I immediately think of my advisor, Rada Mihalcea. I think she can be in this position of one of the greatest researchers.

What did you learn from her that you would take into your own teaching?

I think this attitude of never giving up. She learns a lot from rejections. She never hides away from trying new things and submitting to conferences. Even if we get back rejections, she always says the more rejections she has, the more she learns or the more success she has. That's very inspiring to me because I didn't take rejections in a good way. It would discourage me very fast. But also, as a researcher, you have to deal with that on a regular basis, and it's very good to adopt this mentality that rejection is not a step back. It's a step forward. You learn from it, and you continue stronger in the future.

Read 100 FASCINATING interviews with Women in Computer Vision!

CVPR 2022
DAILY
Computer Vision and Pattern Recognition
Thursday
Tuesday
Wednesday
Friday
zebra
DNN
Memory (Weights)
✓: correctly classified
✗: misclassified
"cat"
"dog"
Stealthy T-BFA
Stealthy TA-LBF
Workshop: Computational Cameras and Displays
Presenting Work by: Ruslan Partsey Ozan Özdenizci Dima Damen & team
Exclusive Interview with: Mathieu Salzmann
Women in Computer Vision: Anna Rohrbach
Today's Picks by: Anh N. Thai
Workshop: Medical Computer Vision
Presenting Work by: Bowen Cheng Fabio Cermelli and Ping Hu Dario Fontanel
Editorial with: Program Chairs
Women in Computer Vision: Shuran Song
Today's Picks by: Maria Dobko
False Positive
In cooperation with
Computer Vision News
The Magazine of The Algorithm community
A publication by
RSIP VISION
Important CVPR community communication on page 27
Cigdem Beyan

MEDICAL
IMAGING
NEWS

JANUARY 2024

Data is key in all AI projects. Training and optimization of neural networks rely on large quantities of data. This is all the more critical for medical AI.

It is also true that medical AI projects often lack a sufficient amount of data. The deficiency might be in quantity and/or in quality. When developing a next-generation medical device, data may even be totally absent.

Let's discover some mitigation tips by RSIP Vision (part 1)

Ilya Kovler, CTO at RSIP Vision

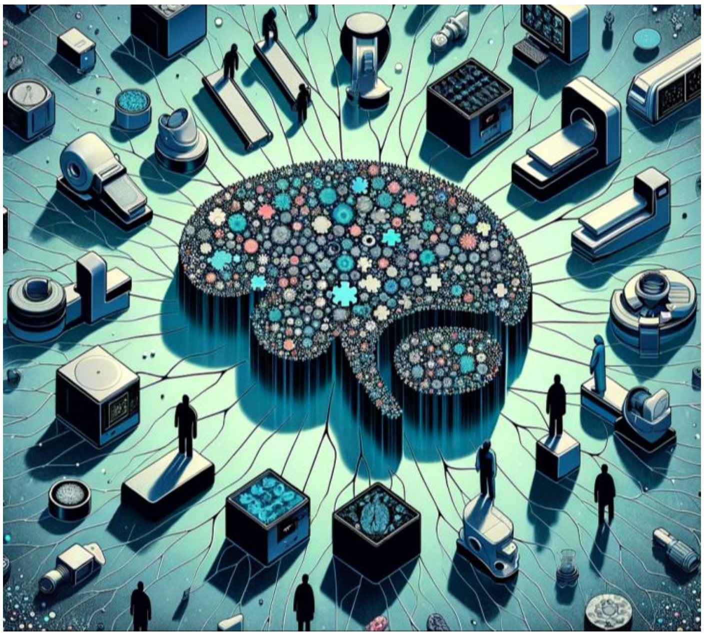
Lacking sufficient data in medical AI is critical because of the vast number of acquisition devices: even simple things like CT segmentation of the anatomy (for instance, airways segmentation or bone segmentation), not to mention more complex tasks such as pathology detection, requires significant variability.

There are many vendors of CT devices, and each one offers many different models. Each one generates scans of different quality, responding to different criteria like price, type, or form factor: **the same anatomy (or the same patient) will be presented with different values, region of interest or quality** by two machines that performed the scan, whether they belong to the same generation of devices or to a different one. In addition, **specific requirements** for a screening protocol could be different in various hospitals based on internal rules.

Data is largely available in hospitals, but in every hospital, there is not much of a variety of acquisition devices. Equipment may have been used in a given for many years, and the quality of CT performed by a 20 years old machine is not the same as that offered by a new device.

On the other hand, it is easier to get data from a restricted number of hospitals, but in that case the AI model will be trained on very specific data, **lacking the required variety** to cover a whole spectrum of devices. Ideally, a neural network would be trained on data from many tens of different sources, with different manufacturers and different models. However, the acquisition of data from each hospital is a long process, that requires significant resources. More often than not, **the acquired data is limited** and this is very challenging for medical AI development. AI models trained on a limited dataset may perform well on similarly acquired scans, but not on different scans performed or by those using different protocols in other hospitals.

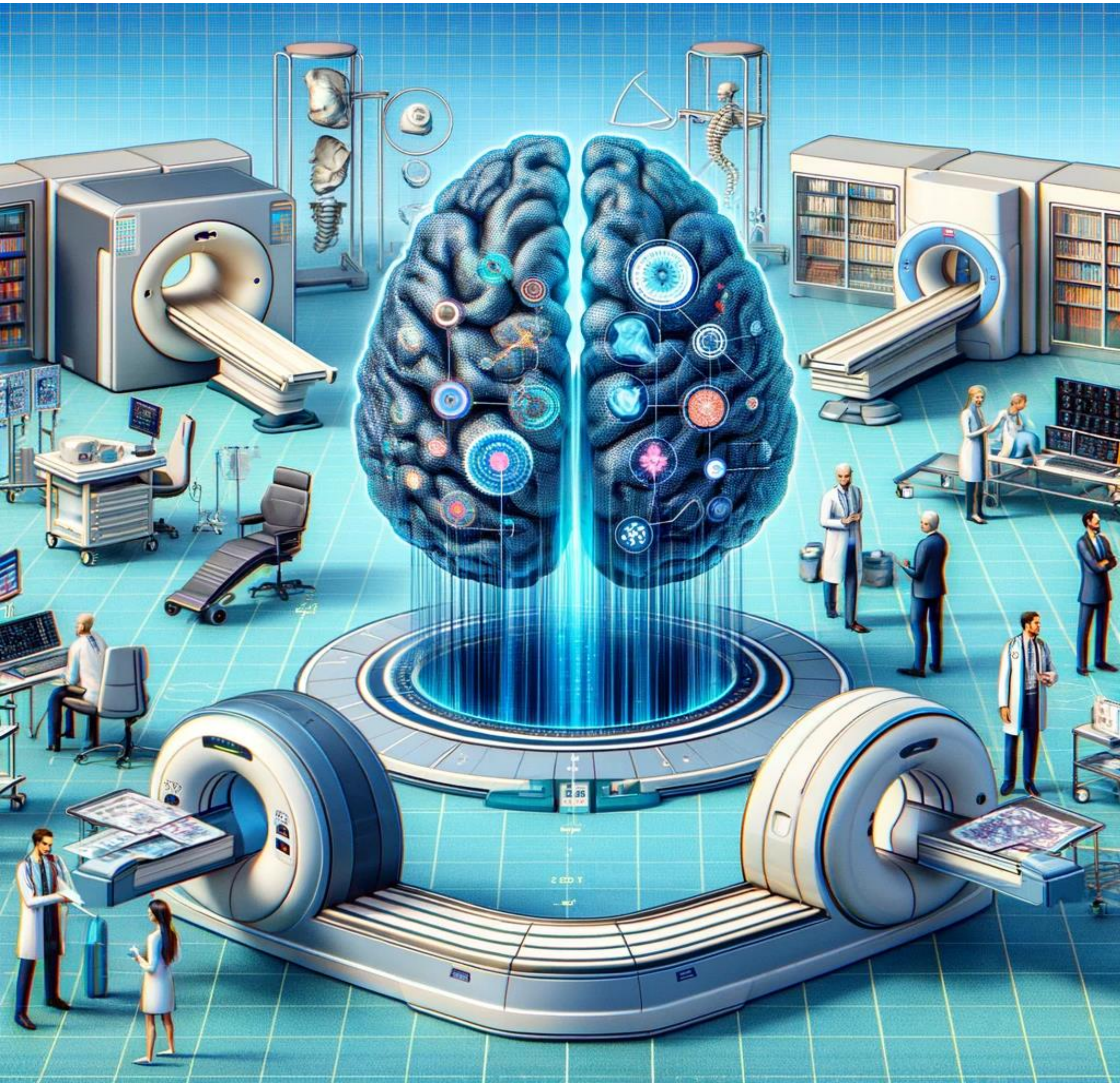
Working with the expertise of **RSIP Vision** enables us to mitigate this challenge in many ways. Our R&D team has put in place modality and application-specific data augmentation tasks. Data augmentation is a frequently performed task, though in this case, we do it in a different way: we define **a specific set of augmentation for this task**, so that the resulting dataset will still make sense from a clinical point of view. "*We work closely with experts like radiologists, orthopedists, and ultrasound specialists to find **the proper range and the parameters for the augmentations**,*" said Ilya Kovler, CTO at RSIP Vision. "*We also use transfer learning from the pre-trained models.*"

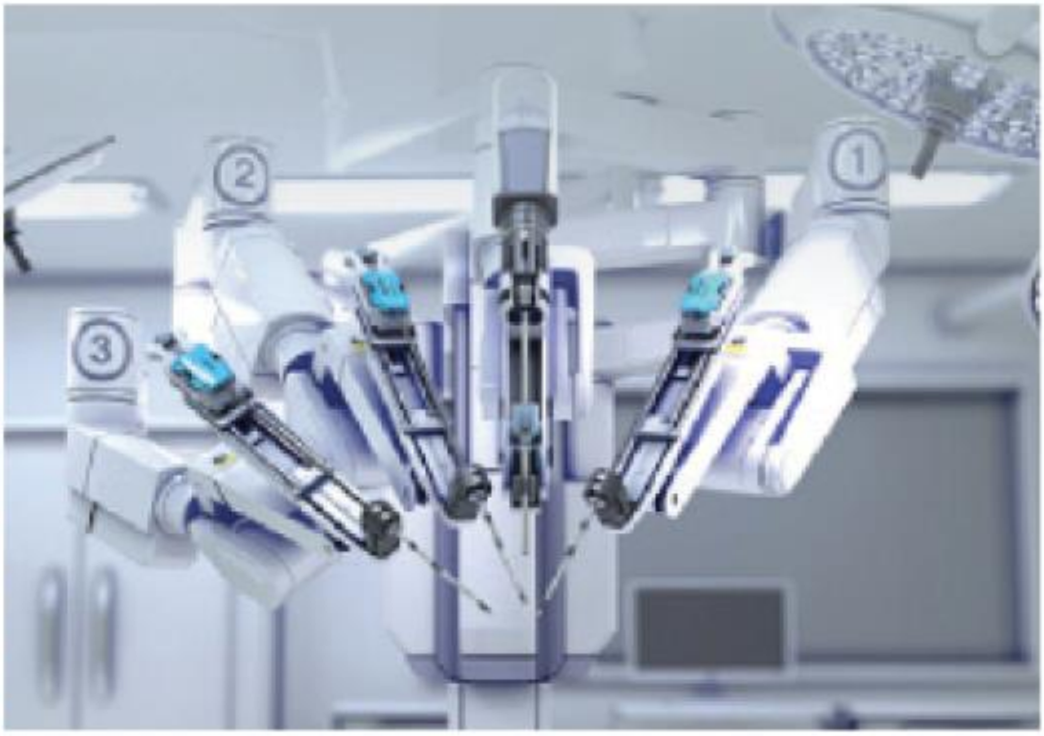


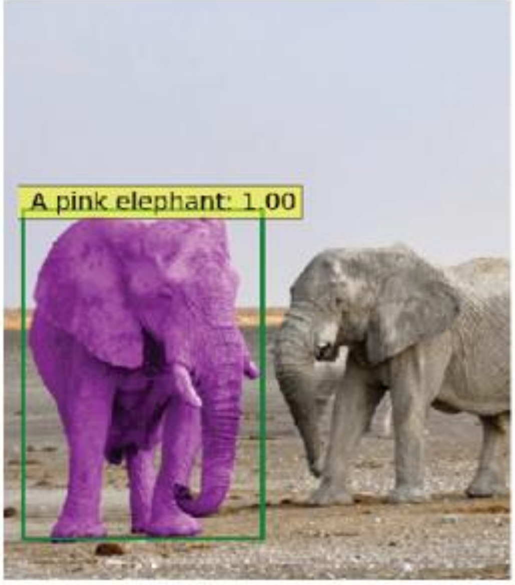
We pre-train neural networks to perform the same task, like segmentation, classification, or detection on a different anatomy. Or pre-train only a part of the neural network layers by training to perform a different task. Also, unsupervised learning can be used very efficiently for pre-training.

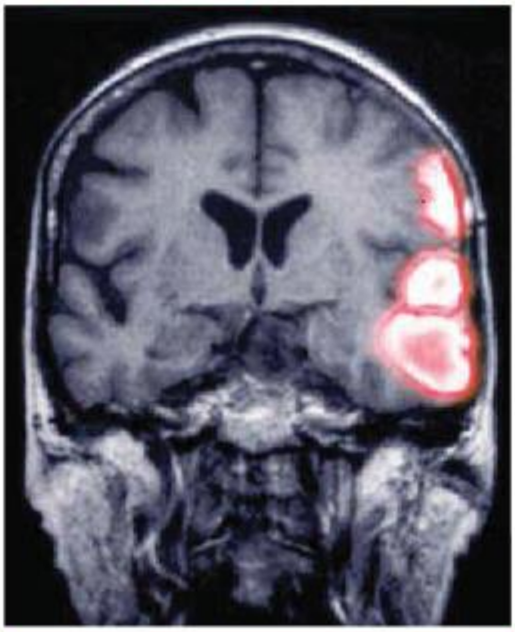
This is just a part of the options that allow us to **significantly increase the AI modules' accuracy, generalization and robustness while working on limited datasets**.

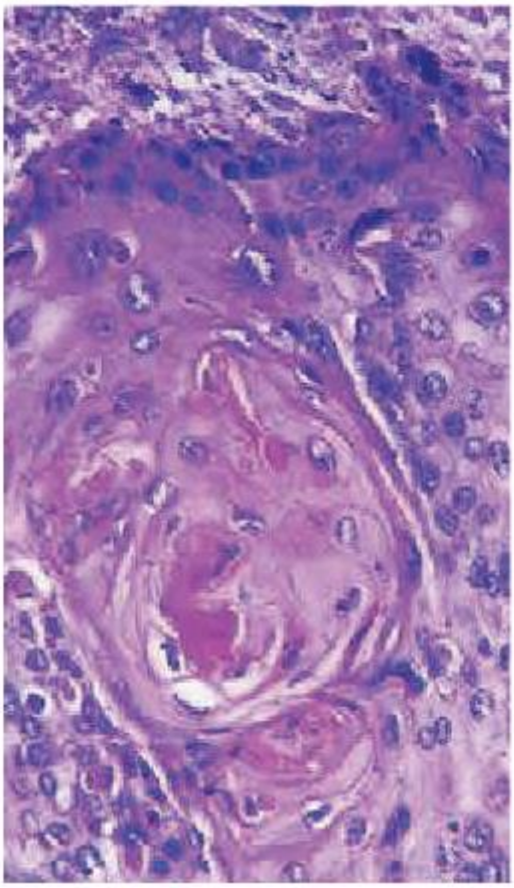
Next month, we will learn about more mitigation techniques by RSIP Vision for the challenges of data quantity and availability.






A pink elephant: 1.00





Knowledge distillation for multi-modal retinal images

"OUR RESEARCH AND INNOVATION MUST ULTIMATELY BENEFIT OUR PATIEN
"CONTINUING MEDICAL EDUCATION IS ESSENTIAL FOR SUSTAINING HIGH STANDARDS OF MEDICA PRACTICE IN SINGAPO

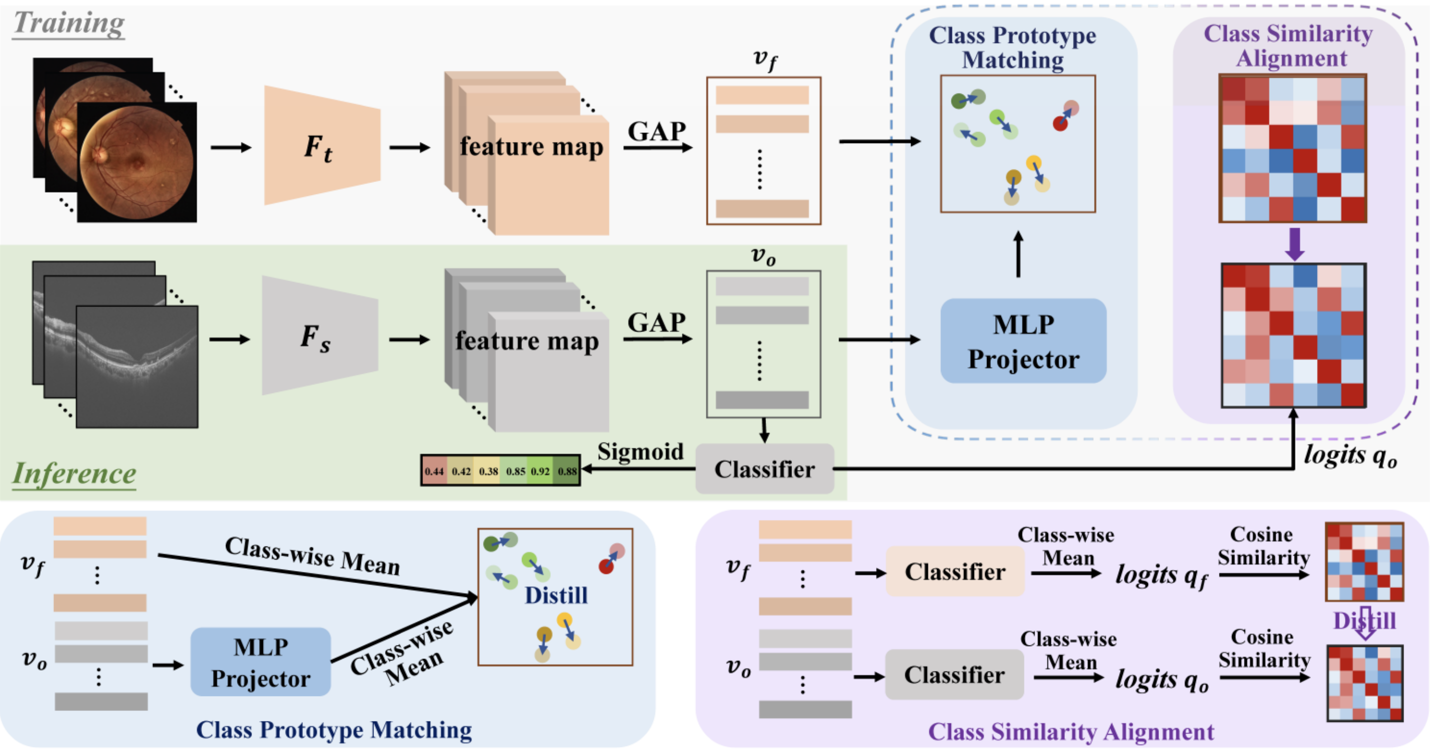
by Christina Bornberg
@datascEYEnce

Hello everyone! I am Christina and as every month, I am happy to share another datascEYEnce story with you! I hope you all had a good start to the new year and are ready to learn about Lehan's work on multi-modal image classification using fundus images to increase the accuracy of OCT classification! This interview is very special to me because it marks the first in-person interview I did for the column!

featuring Lehan Wang

Lehan completed her undergraduate degree in computer science before becoming an intern and later starting her PhD under the supervision of Xiaomeng Li at the Hong Kong University of Science and Technology (HKUST). She was able to present her work *"Fundus-Enhanced Disease-Aware Distillation Model for Retinal Disease Classification from OCT Images"* at MICCAI 2023. This is also the subject of the interview with her.

Now, let's delve into Lehan's journey in deep learning for ophthalmology, which actually didn't start in the vision field, but in NLP. She has been working alongside a brilliant PhD student during her undergrad who taught her more about deep learning and research in general, as well as helped with suggestions for career planning. She ended up choosing HKUST where she had two possible topic options: multi-modal image analysis and image registration. Due to her interest in image fusion and multi-modality, she decided to go with the first option.



Switching from NLP to computer vision was relatively seamless thanks to uni courses and the overlap in basic knowledge of NLP and CV research. The bigger gap she had to fill was the understanding of the data itself. While mostly working with text data that can easily be interpreted she then needed to be able to understand different disease patterns in the two retinal image modalities: optical coherence tomography (OCT) and colour fundus photography (CFP). She also learned, that both modalities give complementary information on a disease, which makes it useful to combine the forces and get a model that learns from both modalities.

After mastering the data, Lehan focused on addressing the multi-modality problem.

The first possible approach involved concatenating features for dataset fusion, but the available research didn't align with her needs - hence it was difficult to find a working baseline. Multi-modal image fusion turned out to perform similar or even worse than single modality with her available data.

Lehan needed to find a way to solve the problem from another angle. She explored two more options for fusing knowledge, one of which was contrastive learning, and the other one knowledge distillation.

Opting for knowledge distillation, she used a teacher model trained on fundus images and a student model for OCT images. She found, that the representation is even for different modalities, and can be seen by the proximity in the feature space. In the next step, methods to pull the two representation spaces of the two modalities closer, Lehan used two methods, class prototype matching and class similarity alignment.

The key contribution was the ability to bridge the gap between different modalities without paired data. The distillation approach involved training the OCT student model with the knowledge from the fundus teacher model and finally allowed single-modality use during inference.

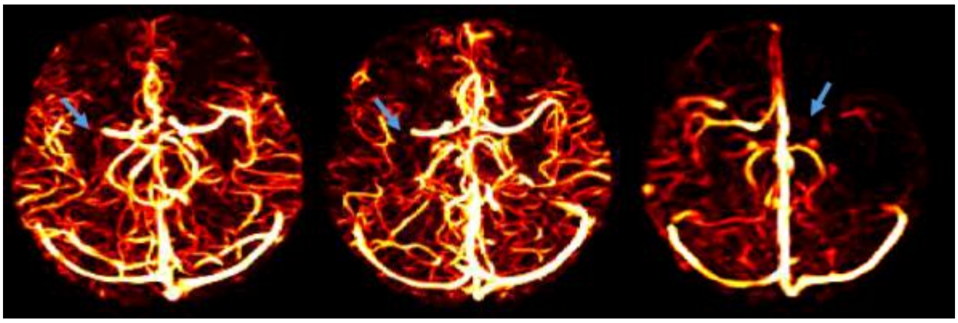
Last but definitely not least, in order to submit (and get accepted) to MICCAI, writing up the entire process proved a bit trickier than anticipated. Finding the right phrases to explain the pipeline needed some more input from her supervisor, who stayed up late together with Lehan before the deadline. This also showed the importance of finding a supervisor that matches one's energy.

I wish Lehan the best for the upcoming years of her PhD, and am looking forward to more fascinating publications from her!

Mumu Aktar defended successfully her Ph.D. dissertation in Computer Science at Concordia University. Congrats Doctor Mumu!

Her research, supervised by Dr. Marta Kersten-Oertel and Dr. Hassan Rivaz, focused on developing machine learning methods to improve treatment decision-making in stroke.

Stroke, a major global health concern, often requires endovascular thrombectomy for effective treatment. The success of this therapy, however, depends on factors like collateral circulation, a radiologic surrogate predicting revascularization response (Figure below). Currently, collateral circulation assessment relies on visual inspection by a radiologist which suffers from inter and intra-rater variability leading to inefficient and time-consuming results.

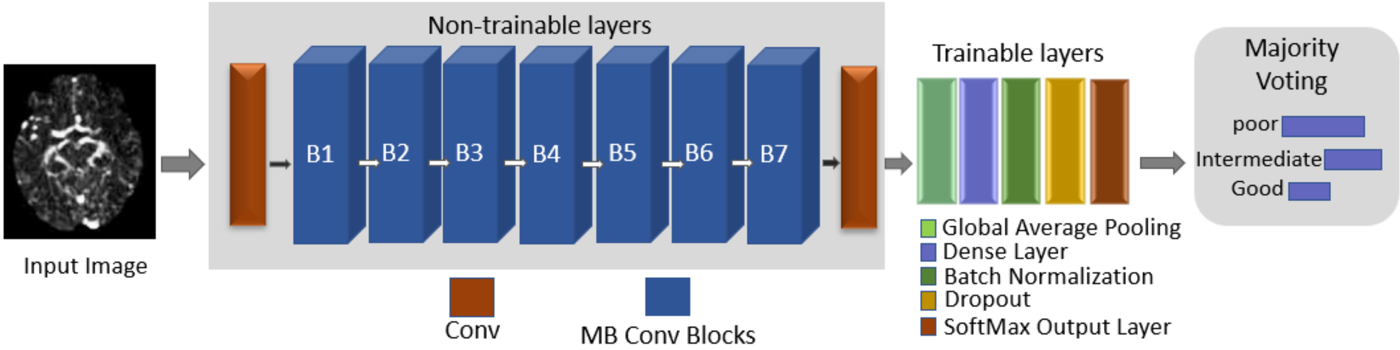


From left to right, an example of good, intermediate, and poor collaterals on contrast-enhanced CTA. The blue arrow indicates the occlusion on the MCA.

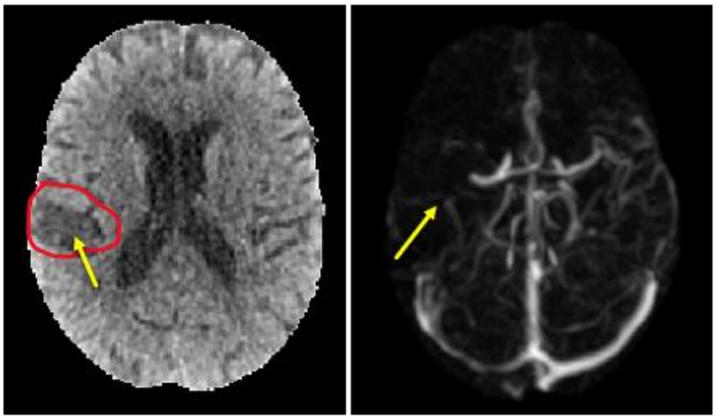
Mumu's research involved developing computer-aided methods to enhance collateral circulation assessment in ischemic stroke. Using deep learning aligned with radiologists' criteria, the approaches offer a more robust evaluation than traditional methods. Overcoming the challenge of small and imbalanced datasets in ischemic stroke, the research represents an advancement in improving the accuracy and reliability of collateral scoring to contribute to better treatment decision-making for stroke patients.

Over the course of the Ph.D., Mumu developed several computer-aided decision support algorithms for collateral evaluation using a 4D CTA dataset, considering two key phases: (1) 2D images from 3D MIPs of the 4D CTA and (2) NCCT extracted from the 4D CTA before the contrast agent.

First, she developed a quantitative collateral scoring method employing low-rank and sparse decomposition, focusing on radiologically assessing filled versus unfilled vessels over time using 4D CTA. Further, recognizing the potential of deep learning over the traditional methods, in automating feature extraction, she implemented a deep learning-driven automatic evaluation system using 4D CTA. This approach leverages knowledge transfer from a pre-trained EfficientNet B0 network (Figure below for the architecture used in this approach) to alleviate the substantial manual engineering typically associated with classical machine learning and quantitative methods. This method utilized focal loss and 2D MIPs from 4D CTA to alleviate the imbalanced and small dataset issues.



Given that radiologists compare an ischemic patient's affected and unaffected sides to determine collateral scores, Mumu further introduced an approach following this criteria that enhances the efficacy of automated evaluation through machine learning. This method focuses on the radiomic features extracted from ischemic damage using NCCT images of both sides of the brain to evaluate collaterals (figure below).

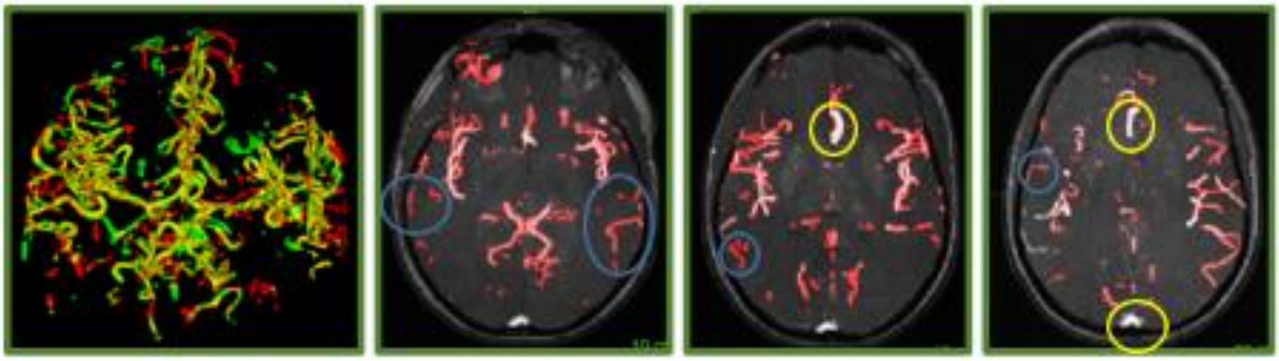


NCCT with ischemic damage

CTA with poor collaterals

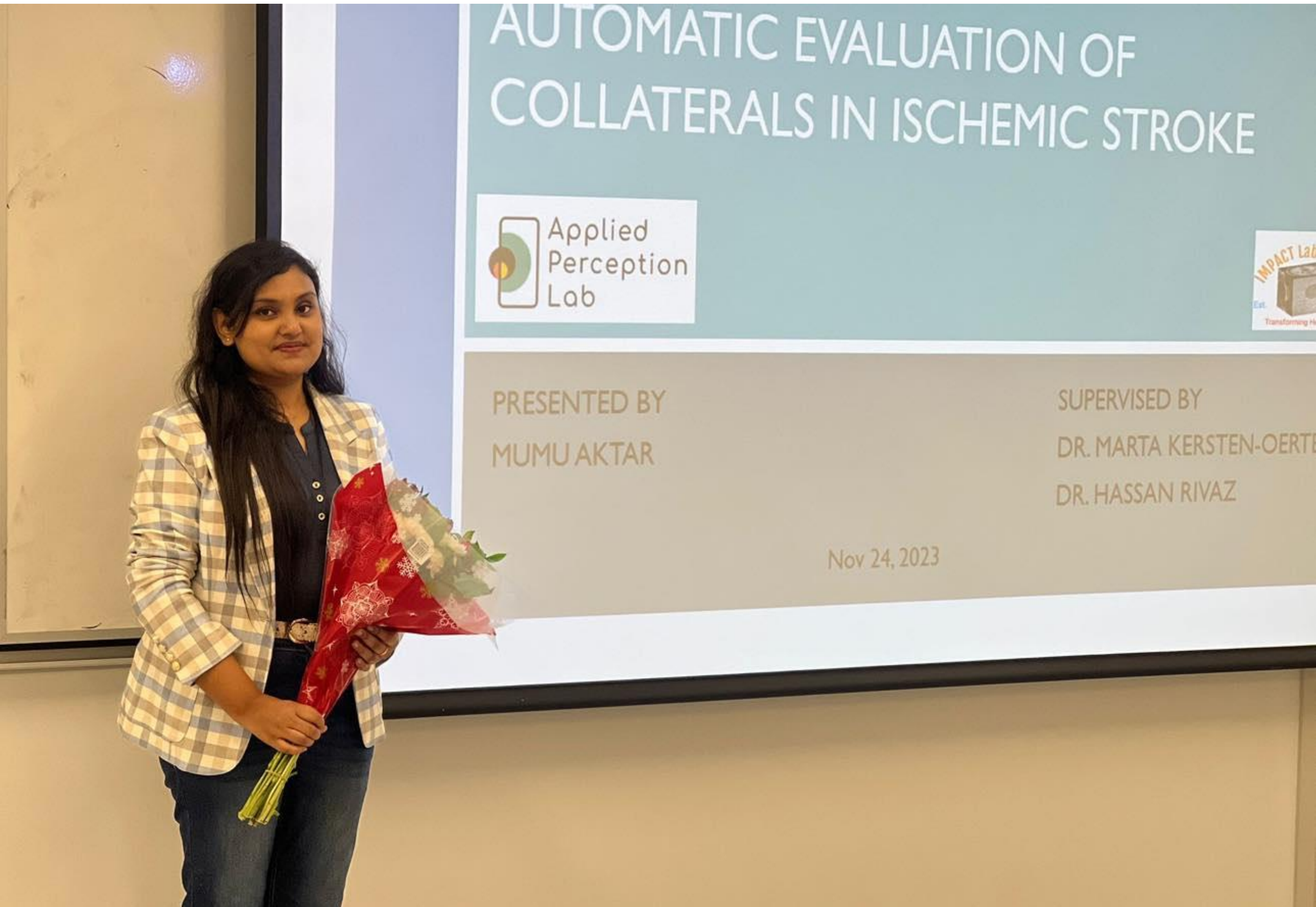
Further, to validate the previous method, a technique employing Siamese networks is proposed which addresses the challenges of small and imbalanced NCCT datasets in collateral evaluation. This approach enhances adaptability to data-scarce medical tasks.

To enhance collateral evaluation, Mumu finally introduced a few-shot learning-based 3D blood vessel segmentation approach, minimizing the need for extensive label annotation in time-consuming 3D scenarios and overcoming limited pre-trained weights for deep learning models in the medical domain (figure below). It can be utilized as a pre-processing step to enhance collateral evaluation.



The results of the research led to novel automated collateral evaluation methods, effectively addressing rater variability as well as small and imbalanced datasets. Unlike existing approaches, Mumu's innovative techniques identify collateral scores from ischemic damage in NCCT or 4D CTA, demonstrating robustness with limited data through transfer learning. Siamese networks expand the methodology to similarity-based problem-solving with minimal data requirements.

Congrats, doctor Mumu!



AUTOMATIC EVALUATION OF COLLATERALS IN ISCHEMIC STROKE

PRESENTED BY
MUMU AKTAR

SUPERVISED BY
DR. MARTA KERSTEN-OERTE
DR. HASSAN RIVAZ

Nov 24, 2023

Image Registration Challenge

Learn2Reg is an international open science initiative in medical image registration which aimes to establish a standardized benchmark for transparent and reproducible comparisons of image registration algorithms.

The freely accessible database is continuously being expanded and currently comprises eight tasks covering a wide range of clinically relevant tasks. Recognized for its achievements, Learn2Reg was awarded with the University of Lübeck's Open-Science-Award 2023.

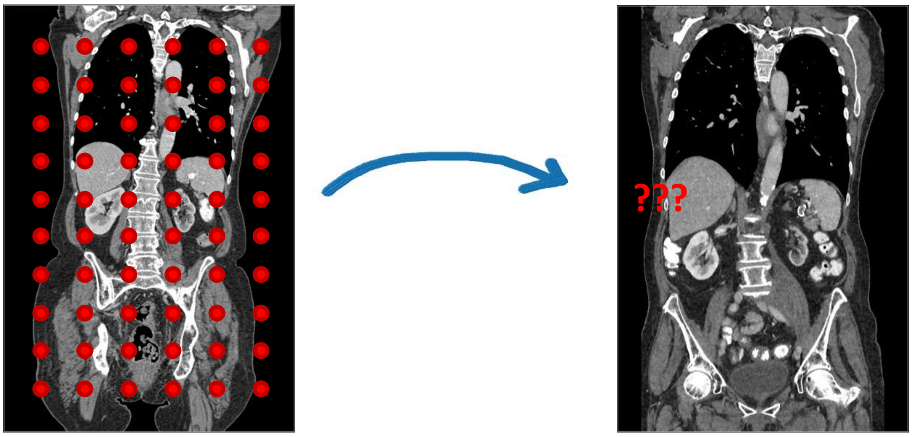
by Alessa Hering

Launched in 2020, the Learn2Reg challenge tackles the gap of transparently evaluating image registration methods. The primary issue in developing and evaluating image registration is the absence of 'ground-truth' deformation fields.

A 'ground-truth' in image registration

requires establishing a 2D or 3D displacement vector for each point in the fixed image, equating to two or three continuous labels per pixel or voxel. This vector bridges the fixed and moving images, meaning the solution exists not within either image, but in the space between them. Unlike for segmentation or classification, medical experts cannot annotate the 'ground-truth' for image registration.

Aim of image registration is to establish spatial correspondences between two or multiple images



Given a fixed image $\mathcal{F}$ and a moving image $\mathcal{M}$. Find a plausible deformation ϕ, such that $\mathcal{F}(x)$ and $\mathcal{M}(\phi(x))$ are similar.

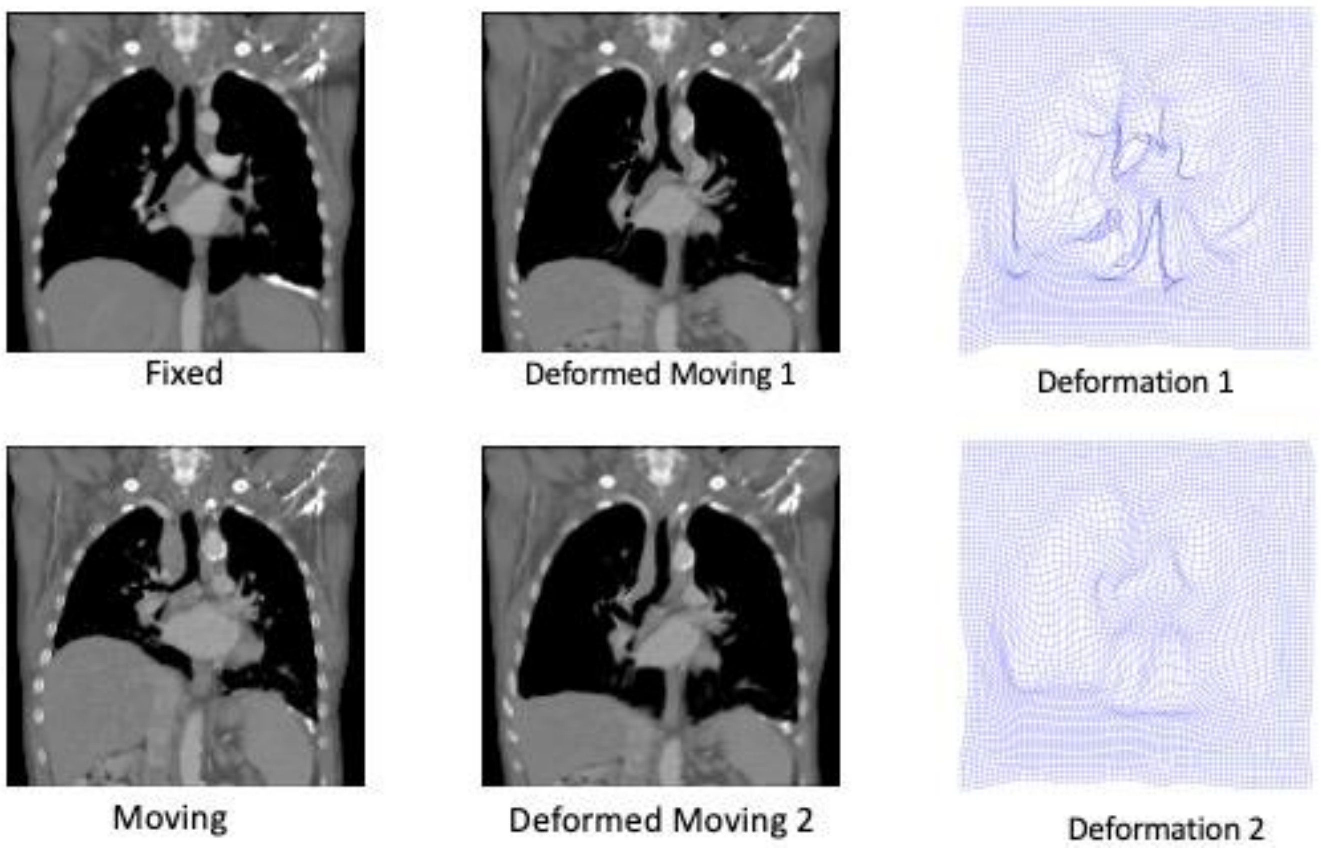
Despite this challenge, we have compiled a diverse dataset encompassing inter- and intra-patient, and single and multi-modal registration tasks. A substantial part of our training dataset includes annotations like segmentations or keypoints for supervised training. We also provide unannotated data for unsupervised or weakly-supervised training using cost functions from conventional registration methods.

To evaluate the performance of registration methods, we use various auxiliary metrics. These metrics measure similarity through segmentation and keypoints, deformation plausibility, robustness, and runtime. We ensure that methods are only ranked higher if their results show a statistically significant improvement, accounting for random noise effects.

Over time, more advanced deep-learning-based image registration methods have been developed.

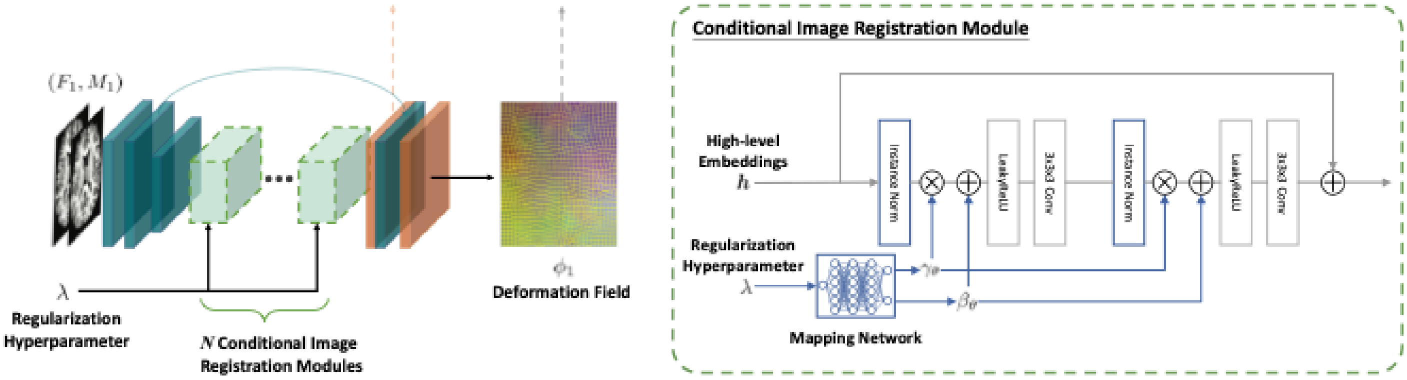
"LapIRN utilizes the advantages of a multi-resolution strategy while maintaining the non-linearity of the feature maps throughout the coarse-to-fine optimization scheme," said Tony Mok about his winning LapIRN method. *"The notorious hyperparameter tuning problem is addressed by introducing a conditional image registration learning paradigm [see Figure]. The framework of LapIRN is flexible and can be easily transferred to various medical image registration applications with minimal effort."*

Yet, Learn2Reg still attracts submissions of conventional methods yielding impressive results. Interestingly, when executed on GPUs, these conventional methods match the speed of their deep-learning counterparts.



The two deformation fields achieve similar accuracy metrics, yet Deformation 2 is smoother, making it more plausible. This shows the importance of evaluating the deformation field's plausibility as well.

Looking forward, Learn2Reg will concentrate even more on clinically relevant tasks. We have just launched a new virtual challenge, **OncoReg** [https://learn2reg.grand-challenge.org/oncoreg/], to improve registration techniques in cancer treatment and screening, with a submission deadline for docker solutions at the end of January. Additionally, new algorithms for all other tasks can be evaluated anytime via grand-challenge or locally by us. In our effort to make image registration more accessible and user-friendly, we are collaborating with **NVIDIA** to develop a **MONAI image registration framework**.



Conditional image registration module of LapIRN which reduces hyperparameter tuning.

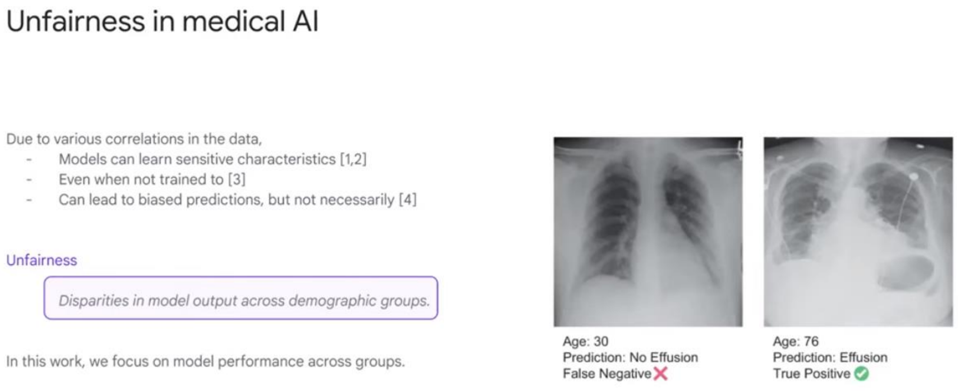
Thank you Alessa

"Datasets through the Looking-Glass" is a webinar series focused on introspecting into the data-related facets of Machine Learning (ML) methods. Our goal is to build a community of enthusiastic researchers interested who care about understanding the impact that data and ML methods could have in our society.

The webinar is part of "Making MetaDataCount" project and is organized by Veronika Cheplygina (left in the picture) and Amelia Jiménez-Sánchez (right in the picture) at IT University of Copenhagen. We had four successful editions so far (in February, June, September, and December 2023) with 14 speakers in total, the videos are available on our YouTube playlist.

In our last webinar, we covered several topics about spurious correlations or shortcuts, fairness, out-of-distribution data and augmenting annotations of publicly available medical image datasets.

Jessica Schrouff is a research scientist at **Google DeepMind**, working on **responsible AI through a causal perspective**. Jessica discussed in her talk how a model could learn sensitive characteristics due to various correlations in a dataset. They hypothesize that when a model learns a shortcut, increasing the encoding of the sensitive attribute affects the (un)fairness metrics. She uses the term **unfairness** as disparities in model output across demographic groups. Their method correctly detects shortcut learning when a spurious correlation is engineered, and they could vary the



Unfairness in medical AI

Due to various correlations in the data,
- Models can learn sensitive characteristics [1,2]
- Even when not trained to [3]
- Can lead to biased predictions, but not necessarily [4]

Unfairness

Disparities in model output across demographic groups.

In this work, we focus on model performance across groups.

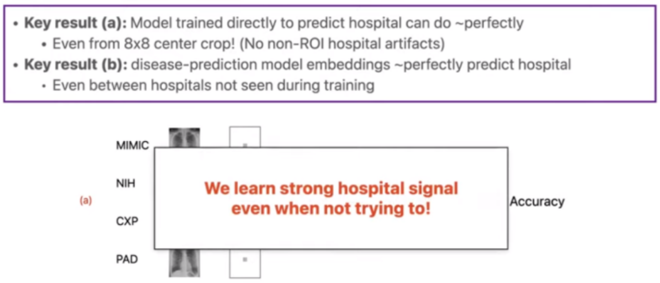
Age: 30
Prediction: No Effusion
False Negative ✗

Age: 76
Prediction: Effusion
True Positive ✓

Check out the video!

amount of sensitive encoding (e.g. age) with an intervention. She discussed that the encoding of a sensitive attribute is not predictive of a model's fairness characteristics. She showed that subsampling can be an effective mitigation tool. Finally, she showed that with their method, when shortcutting is not the (main) source of unfairness, fairer models could be selected.

The speakers of the second talk were **Rhys Compton** and **Lily Zhang**. Rhys Compton holds a Master's in Computer Science from New York University (NYU), and Lily Zhang is a PhD candidate at NYU in the Center for Data Science. Rhys Compton along with Lily Zhang talked about their work "**When More is Less**". They discussed how introducing additional datasets can hurt performance by introducing spurious correlations. They found that **hospital signal is strongly embedded in chest X-ray**, models were able to predict source hospital of an image with 98% accuracy! They discussed how balancing can help but does not always improve classification performance. Balancing seemed to be most beneficial when datasets had significantly different disease prevalence.

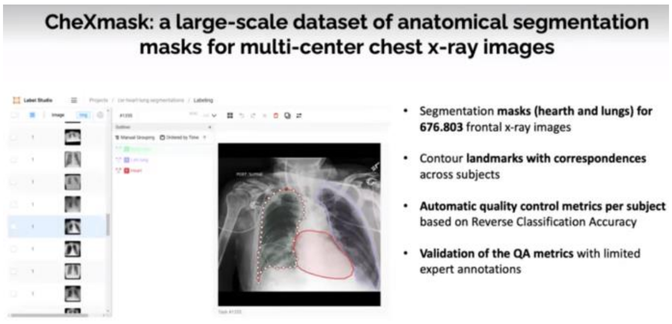


- **Key result (a):** Model trained directly to predict hospital can do ~perfectly
  - Even from 8x8 center crop! (No non-ROI hospital artifacts)
- **Key result (b):** disease-prediction model embeddings ~perfectly predict hospital
  - Even between hospitals not seen during training

MIMIC
NIH
(a)
CXP
PAD
We learn strong hospital signal even when not trying to!
Accuracy

Check out the videos!

Enzo Ferrante is a faculty researcher at Argentina's National Research Council where he leads the Machine Learning for Biomedical Image Computing lab. Enzo presented their journey when creating **a large-scale dataset of segmentations (CheXMask) for several publicly available chest X-ray datasets**. He discussed some of the challenges they encountered for the creation of a derived dataset. He highlighted the importance of providing quality assessment (QA) metrics when leveraging automatic annotations, and emphasized the importance of expert validation. They also showed that QA metrics can be used as surrogates to audit for fairness in new populations without ground-truth annotations.

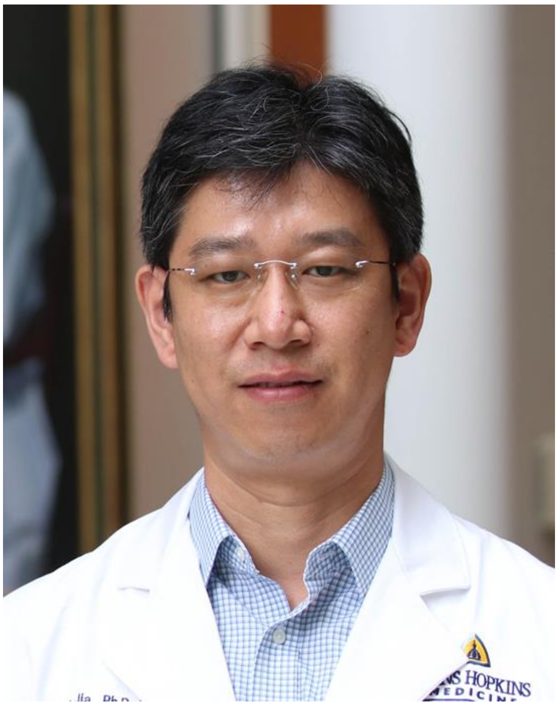


CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images

- **Segmentation masks (hearth and lungs)** for **676.803** frontal x-ray images
- **Contour landmarks with correspondences** across subjects
- **Automatic quality control metrics per subject** based on Reverse Classification Accuracy
- **Validation of the QA metrics** with limited expert annotations

We are planning to hold our next webinar in March 2024, sign up for our newsletter if you want to stay updated!

Quantitative Intravoxel Incoherent Motion (IVIM) Diffusion MRI (dMRI) Reconstruction Challenge



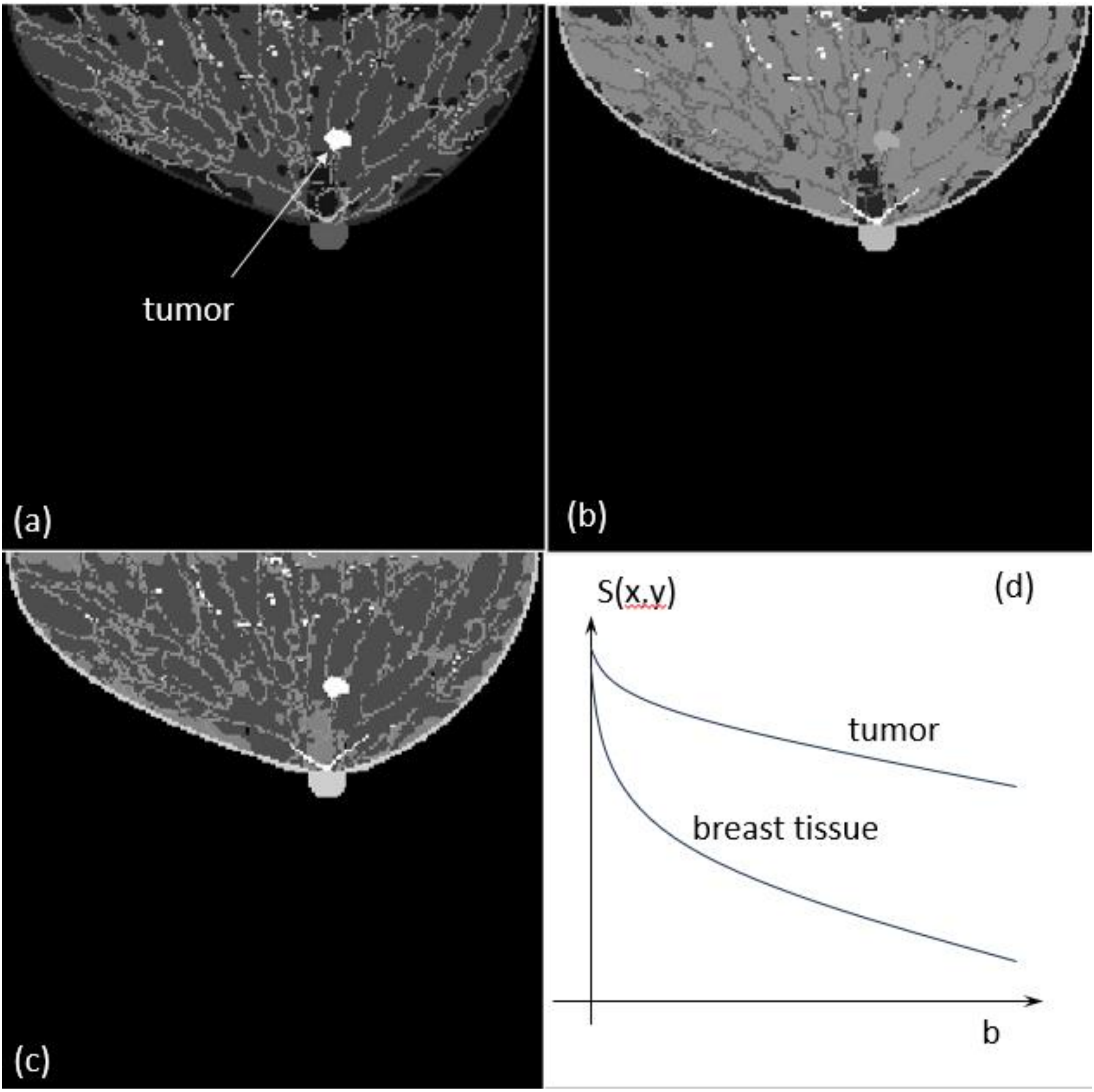
Xun Jia (left) is Chief of the Medical Physics Division at the Radiation Oncology Department of Johns Hopkins University. He is also the Lead Organiser of the IVIM-dMRI Reconstruction Challenge, one of two grand challenges sponsored by the American Association of Physicists in Medicine (AAPM) this year. Xun is here to tell us more, with Emily Townley, AAPM's Staff Program Manager for the Working Group on Grand Challenges, and Karen Drukker (right) from the University of Chicago, the Working Group Chair. Thank you so much Emily for all that you did to make this feature possible!

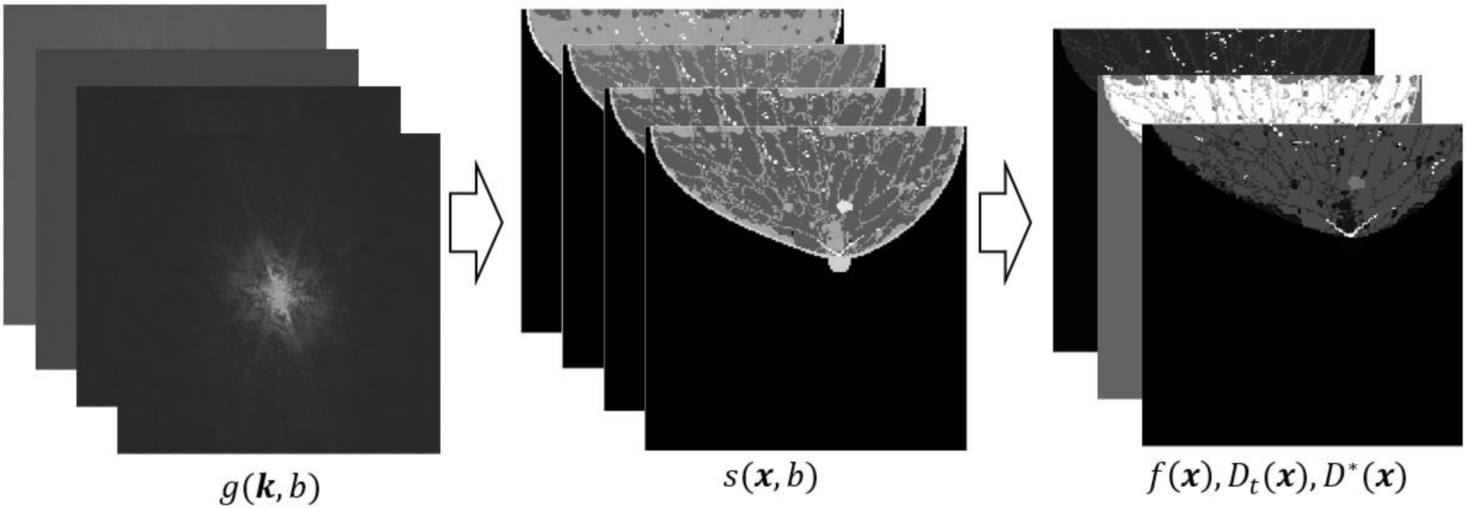
The **IVIM-dMRI Reconstruction Challenge** tackles the problem of **parameter reconstruction in intravoxel incoherent motion (IVIM) diffusion MRI (dMRI)**. The technology, crucial for disease diagnosis and other clinical applications, involves reconstructing three biologically significant parameters.

"The practical challenge is that the reconstruction is usually unreliable," Xun reveals. *"It has a lot of uncertainties because of the mathematical challenges of this reconstruction problem. That gives the parameter mapping considerable uncertainties that affect downstream applications. We set up this challenge to have participants compete on who can achieve* ***the most reliable parameter reconstruction in this IVIM-dMRI context****."*

While similar challenges exist in other medical imaging research fields, such as **MICCAI**, this is the first challenge of its kind in the **medical physics domain**. Xun attributes its uniqueness to the fact that MRI has not traditionally been a major study topic in the domain, with organizations like the **International Society for Magnetic Resonance in Medicine (ISMRM)** taking the lead in MRI research. However, with MRI gaining traction in the field, Xun hopes more people will want to dive into this problem.

"From a scientific perspective, as a researcher, this is a common ground





for people to study a challenging problem with a common dataset," he points out. *"It's an opportunity for people to establish the performance of their own algorithms. That's very important if they want to develop novel algorithms."*

The challenge seeks to explore improvements in parameter reconstruction beyond the current clinical standard by **leveraging advancements in deep learning and classical reconstruction algorithms**. Participants will be provided with a pre-prepared sample code to understand the baseline level upon which to improve.

For those participants who are not MRI researchers or familiar with MRI principles, Xun says understanding the problem at hand is vital for optimal performance. *"In reconstruction, we talk about the forward problem going from the solution to the measurement,"* he points out. *"The actual challenge is the inverse problem, going from the measurement to the solution. Participants have to understand the forward problem carefully."*

Karen agrees: *"For people not from the medical physics field, it's good to know the actual clinical question and its importance. We get a lot of participants for challenges that are in no way, shape, or form from medical physics or even medical imaging. They come from computer vision or completely unrelated fields. It's a way to bring more expertise into the field."*

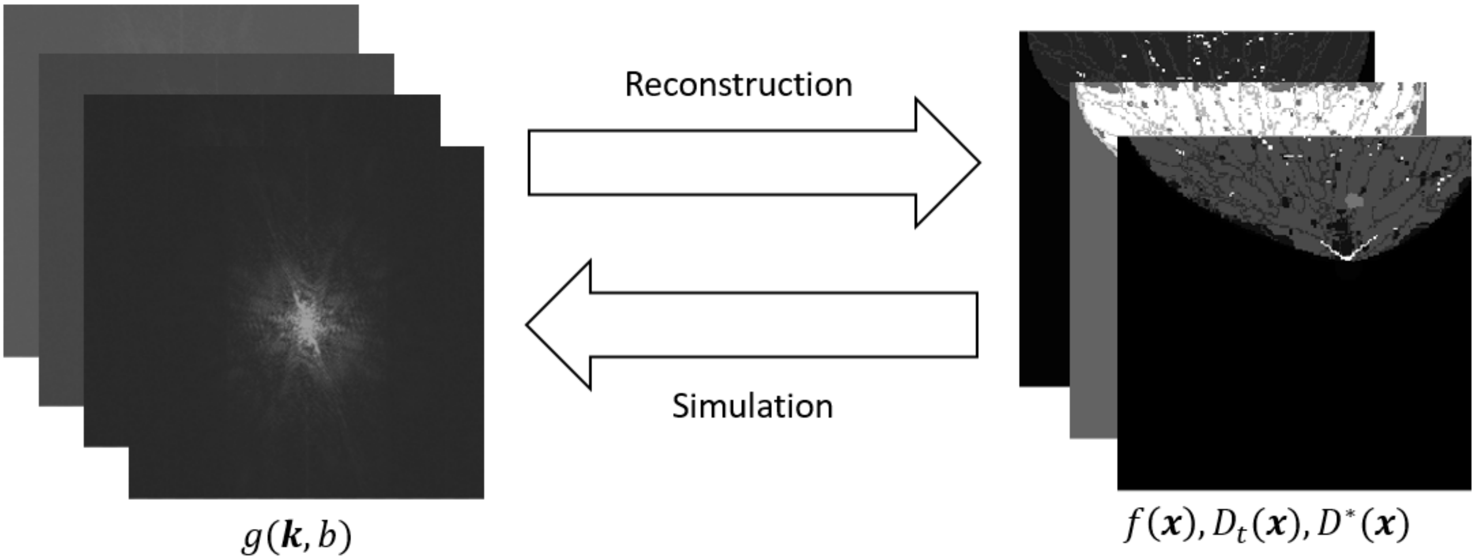
"It's an opportunity for people to establish the performance of their own algorithms. That's very important if they want to develop novel algorithms!"

By uniting people from various fields, the challenge aims to **drive innovation, increase accuracy, and reduce uncertainty in parameter reconstruction**. While it serves an important and serious purpose, Karen wants to highlight that participating in these events should be fun, and taking part is not without its perks. "*It does take quite a bit of time and effort, but it's a friendly competition,*" she reminds us.

"*The winners will get to go to the AAPM annual meeting with their registration comped by AAPM. The meeting is in Los Angeles this year, so it's a pretty good destination!*"

Emily adds: "*There will also be recognition in potential future publications, as well as at the dedicated Grand Challenge symposium at AAPM annual.*"

Registration opens on January 15. See the Grand Challenge website for more information, and the organizers are on hand to answer any questions that people may have.


Reconstruction
Simulation
$g(\boldsymbol{k}, b)$
$f(\boldsymbol{x}), D_t(\boldsymbol{x}), D^*(\boldsymbol{x})$

SARAMIS: Simulation Assets for Robotic Assisted and Minimally Invasive Surgery

Nina Montaña Brown is a final-year PhD student at UCL jointly at the Centre for Medical Image Computing (CMIC) and the Welcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS).

She speaks to us about her NeurIPS paper proposing the first large-scale dataset of 3D rendering assets designed to simulate computer vision data for synthetic tasks about surgery.

The ever-increasing need for data has driven advancements in predictive models to perform specific tasks, particularly within the realm of computer vision. We have seen incredible technological advances in self-driving cars, for example, which can learn to navigate in an open-ended environment based on imaging and depth information.

"Increasingly, we're seeing algorithms being trained on ***synthetic data****,"* Nina begins. *"It's quite similar to when you play a video game. You can simulate a camera and its position in a synthetic environment that's made up of 3D objects and their textures and the lighting conditions, and that gives us an output of an image of that environment."*

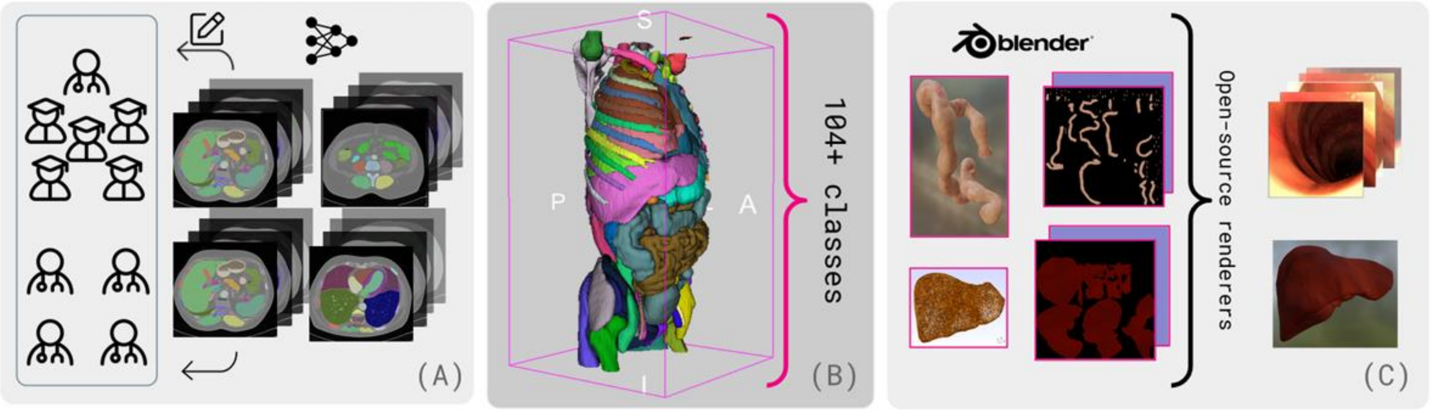


Figure 1: Summary of SARAMIS pipeline. We annotate a large-scale dataset of open-source CT scans (A), remesh and convert them into simulation files (triangular and tetrahedral meshes, normal maps, diffuse maps) (B) that can be used with a open-source renderers (C) to produce synthetic training data for MIS applications.

Datasets like **Cityscapes** collect real data with sensors attached to cameras to train the algorithms. However, **collecting real data for medical scenarios is challenging due to the high annotation cost and logistical difficulties in obtaining data during surgeries**. There is not a large volume of data available that can be used to make realistic scenes.

SARAMIS seeks to overcome these hurdles by providing **the first set of environments and data that can be used to generate a wealth of synthetic data**, enabling the training of computer vision algorithms for medical applications. It comprises 3D meshes, textures, and tetrahedral volumes covering various anatomical targets. It can create environments for **training reinforcement learning (RL) agents** to perform specific actions or generate realistic images with paired labels, such as segmentation maps, depth maps, and optical flow.

"SARAMIS was acquired from a relatively large set of open-source CT scans, which were previously available," Nina tells us. *"The dataset is made up of over 2,500 individual patient cases where all these anatomical targets were segmented, reviewed, and then converted into assets. We have 2,500 now, but realistically, if another large set of CT scans was openly available, we could apply the algorithms and methods to more scans to obtain an even larger set of these assets."*

Currently, the dataset encompasses **106 anatomical targets**, including two previously unlabeled in the literature. The team used an open-source

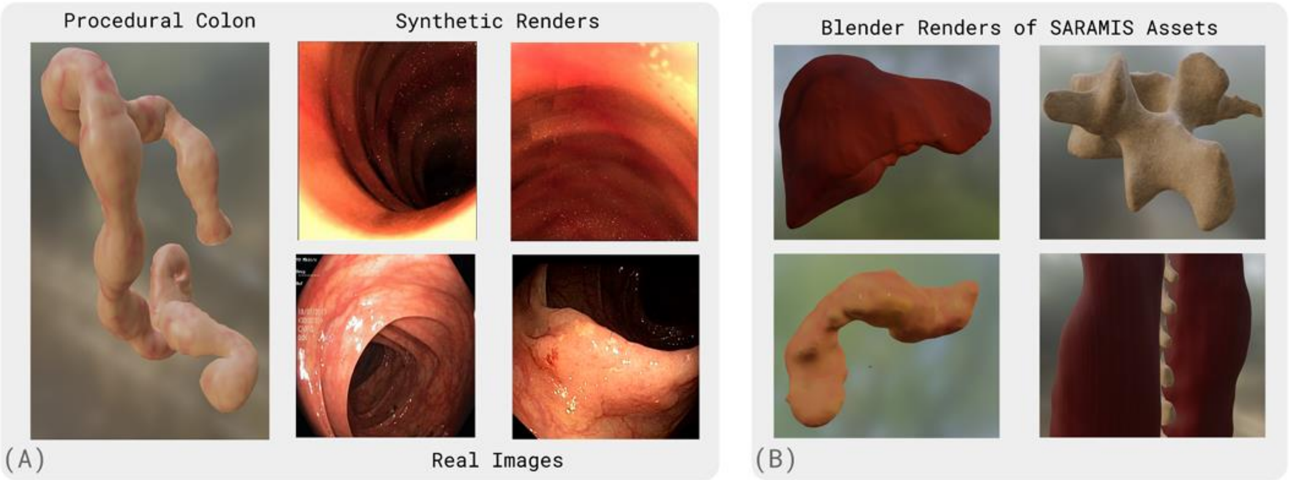


Figure 2: Textured and shaded assets from SARAMIS. In A), we render a procedurally generated colon, with two examples of synthetic renders of the colon, as well as reference real images from the HyperKvasir dataset [6]. We showcase other assets from SARAMIS in B), namely the liver (top left), vertebrae (top right), pancreas (bottom left), and muscle (bottom right).

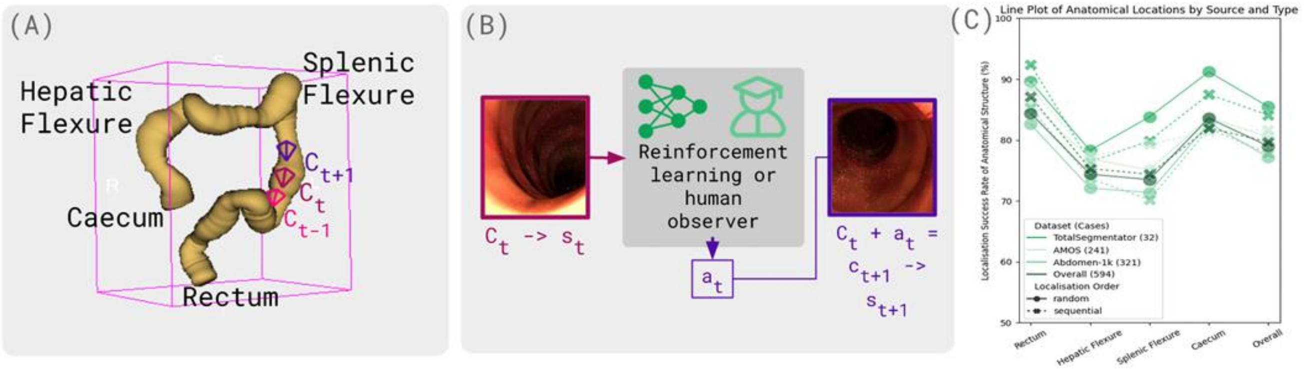


Figure 4: Summary of autonomous navigation experiment. Given a patient-derived mesh model of the colon, and defined navigation targets (A), the camera pose in the environment is used to render the synthetic view inside the colon. Using the rendering as the state s_t, a human observer or an RL agent may sample action a_t with the aim of reaching navigation structures (B). We report the success rates by sub-dataset on the hold-out test set in (C), showing good generalisation accross unseen test sets.

model, **TotalSegmentator**, to generate the first set of labels, including the cardiovascular system, abdominal organs, muscles, bones, and vertebrae. In doing so, they found two new vertebrae that hadn't been annotated previously. "*We know that model has now been updated,*" she adds. "*There are more labels, so we could apply this new model and get newer labels to expand the dataset further.*"

Furthermore, enhancing soft tissue resolution is a key consideration for future iterations of this work. As a base imaging modality, CT scans do not have the best soft tissue resolution. Ligaments and fascia are visible in MRI but not CT, but they would be helpful for the simulation and environment aspect for surgery. The same or updated algorithms could be applied to different data types to expand SARAMIS further.

Presenting SARAMIS at **NeurIPS** was a strategic decision as the conference featured a dedicated **Datasets and Benchmarks Track**, providing an ideal platform to showcase **a data-centric paper**. It also meant it would reach medical imaging practitioners and a broader audience of computer vision researchers who could leverage the dataset for their algorithms.

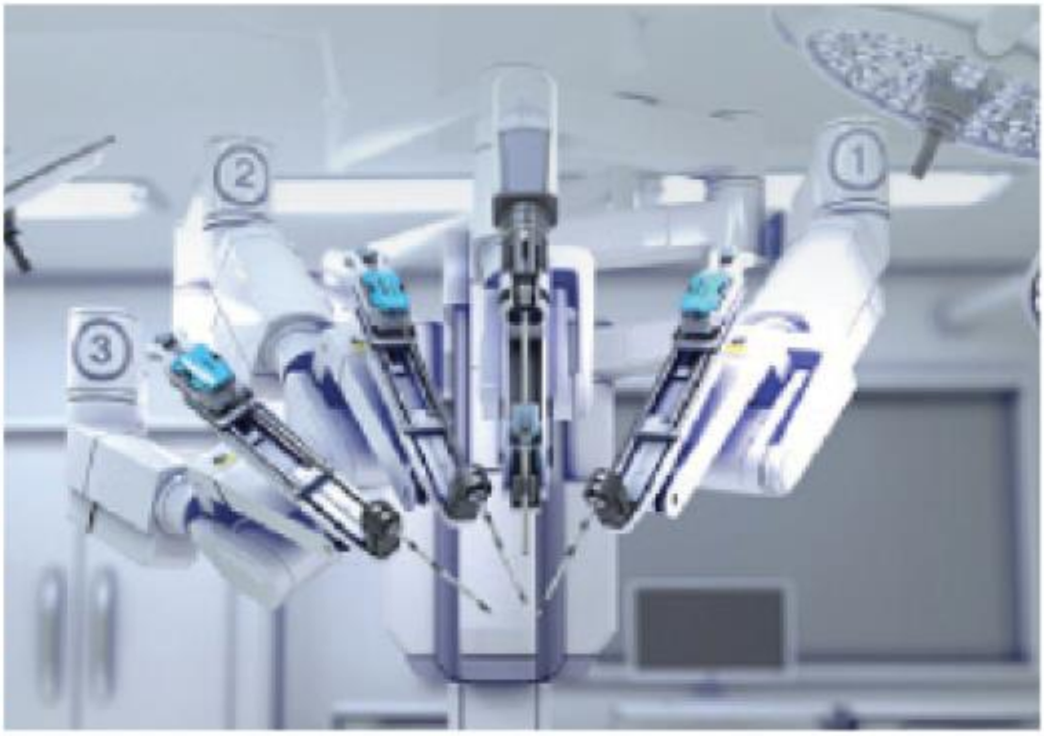
While SARAMIS already represents a significant achievement, Nina acknowledges areas for improvement. Manual texturing, for instance, could be replaced with a data-driven approach for **more realistic rendering**. "*The best way to do that is to create a scattering function, which you can physically simulate how the light interacts with,*" she explains. "*You get* *realistic rendering and speckles and changes in appearance, which are based specifically on the physical characteristics of the materials. Under the supervision of a surgical resident, we used* ***Blender*** *to design these scattering functions visually to match the appearance of those organs during surgery by using some open-source images. With such a large quantity of data and the availability of open source 2D video of surgery, it would have been nice to learn in a data-driven way what those scattering functions look like, so you can get as close to data realism as possible."*

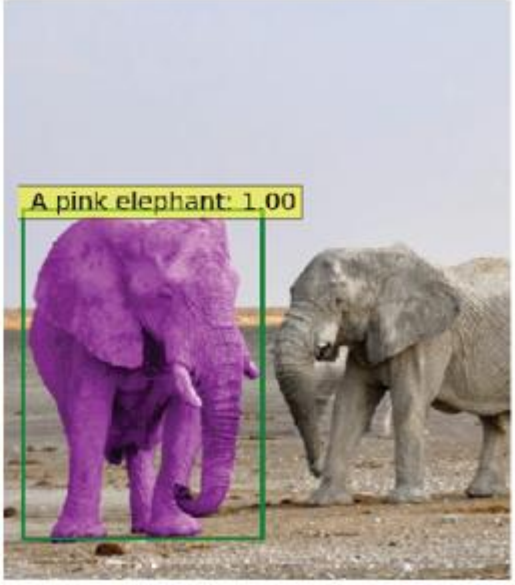
SARAMIS is a collaborative effort between **WEISS** and **CMIC** teams involving clinicians, labelers, and algorithm developers. Nina is based primarily in WEISS, but the team comprises different subgroups, including a clinician subgroup of practicing radiologists from UCLH hospitals and a labeling subgroup of mainly PhD students in medical imaging, with a lot of expertise in what anatomy looks like. The latter did the first pass of the labeling before sending the data to clinicians. The algorithms and dataset were primarily developed at WEISS, directed and organized by Nina.

"I'd like to warmly invite the wider computer vision community to look at SARAMIS," she says. *"I think it's quite interesting in terms of* ***opening new avenues for improving vision algorithms in surgery and potentially opening the door to pure computer vision practitioners to the surgical setting****. We'd really appreciate the feedback. Just drop me a line if you want to chat about it."*

Beyond SARAMIS, Nina, originally from Spain and now based in London, is pursuing her PhD focused on **developing automatic registration methods for trackerless image guidance in laparoscopic liver surgery**. She is also an ML engineer at **Tortus**, developing an **AI co-pilot for doctors**, which helps with automated documentation and computer control of EHR systems.

"I started my PhD just before the pandemic," she reveals. *"I was going to be in surgery collecting lots of data and doing supervised learning to enable these algorithms and methods. However, the pandemic happened, which threw a bit of a spanner in the works! We couldn't go into surgery because all the hospitals got turned into Covid wards, so the data collection aspect was thrown out of the window. Instead, I pivoted my work to look at* ***synthetic data generation****, learning regimes that use synthetic data, and the potential of* ***using algorithms in a synth-to-real approach****."*




A pink elephant: 1.00