December 2023

Computer Vision News

& Medical Imaging News

The Magazine of the Algorithm Community

Maria Koskinopoulou

A publication by RSIP VISION

OpenMask3D: Open-Vocabulary 3D Instance Segmentation

Ayca Takmaz (left) is a second-year PhD student at ETH Zurich, under the supervision of Prof. Bob Sumner and Prof. Siyu Tang. Ayca also closely collaborates with Dr. Francis Engelmann. In her research, she focuses on 3D scene understanding, particularly for augmented reality applications. Elisabetta Fedele (right) is a Master's student at ETH Zurich Computer Science department, specializing in theoretical computer science and machine learning, with a focus on computer vision.

by Ayca Takmaz and Elisabetta Fedele

3D instance segmentation, which is the task of predicting 3D object instance masks along with their object categories, has many crucial applications in fields such as robotics and augmented reality. In recent years, 3D instance segmentation approaches have achieved significant success. However, current methods operate under a closed-set paradigm - they can typically only recognize object categories from a pre-defined closed set of classes that are annotated in the training datasets.

As a result, these methods often have a limited ability to understand a scene beyond the categories seen during training, and cannot directly answer more natural, free-form questions such as "Where can I sit?" "Where is the side table with flowers on it?", "Do you know where my leather bag is?", "Where are my keys, the ones with the blue keychain?" or identify less common objects such as "a toy penguin", "Cetaphil soap", and "an abstract-style wall-art".

In an attempt to address and overcome the limitations of a closed-vocabulary setting, there has been a growing interest in open-vocabulary approaches, which have the capacity to understand categories which are not present at all in the training set, and respond to free-form queries in a zero-shot manner. Recently proposed open-vocabulary 3D scene understanding approaches, however, typically output a heatmap over the points in the scene given a query, which has limited applications, particularly when one needs to identify object instances.

To address these limitations, we introduce the task of open-vocabulary 3D instance segmentation. We propose OpenMask3D, which can perform zero-shot 3D instance segmentation in an open-vocabulary manner (Fig. 1).

Figure 1: Given a 3D scene and free-form user queries, our method OpenMask3D segments object instances described by the open-vocabulary queries.

Let's take a look at how OpenMask3D works: given a set of posed RGB-D frames together with the reconstructed 3D geometry of the scene, OpenMask3D outputs a set of class-agnostic 3D instance masks, and for each of these instances, it computes a feature vector in the CLIP space (Fig. 2).

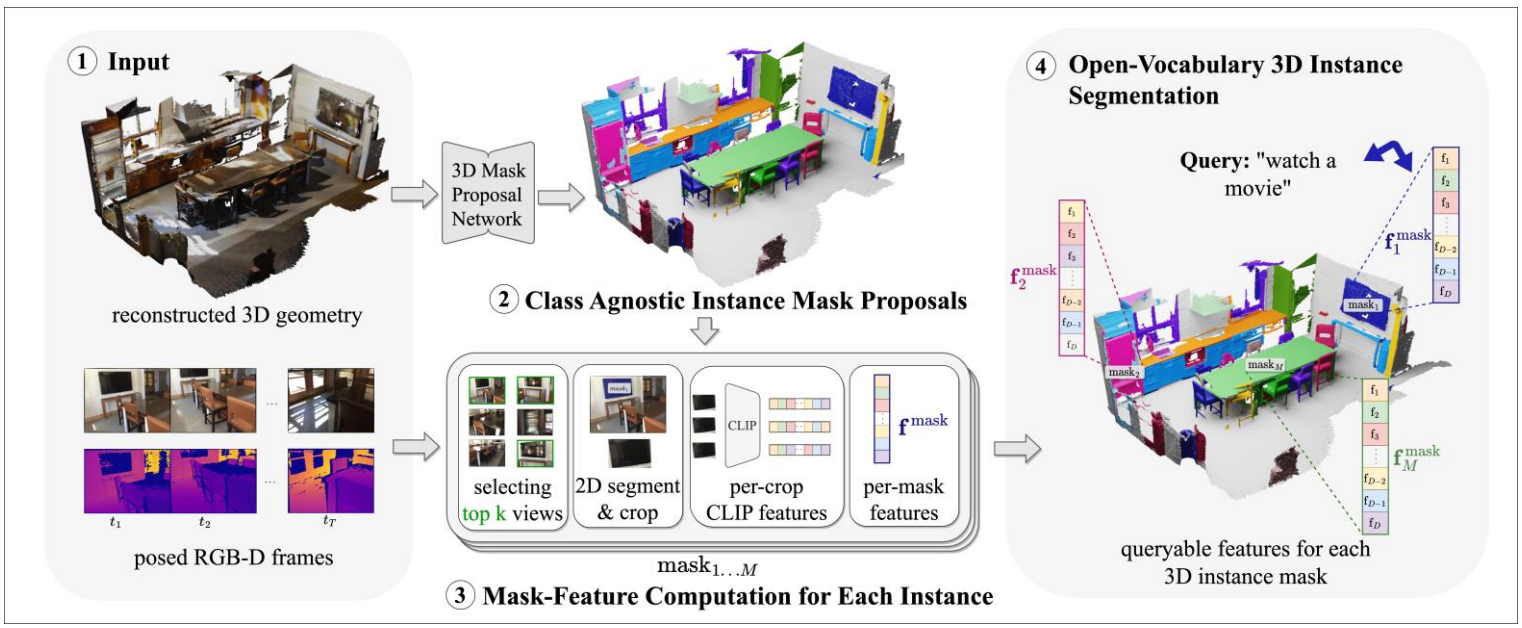
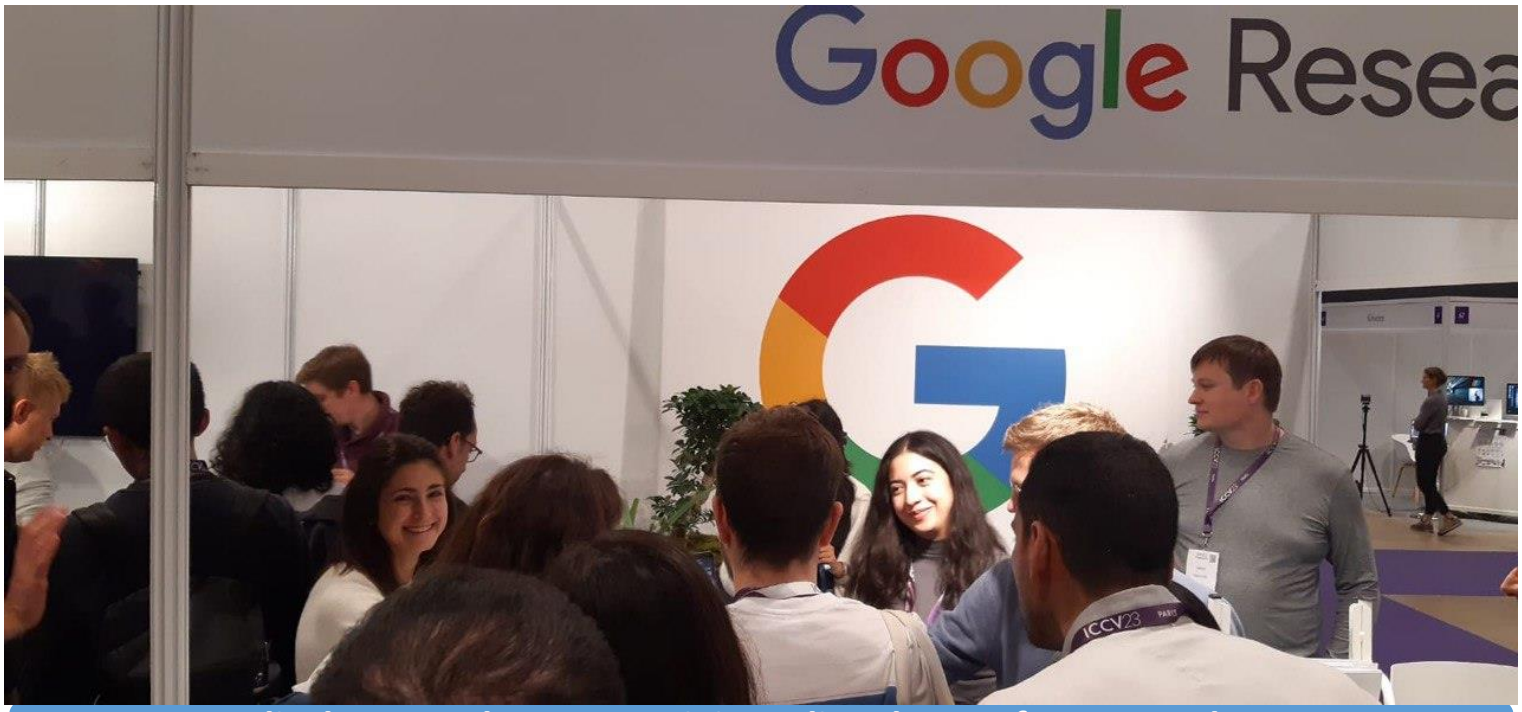


Figure 2: Overview of OpenMask3D

These per-mask features can be used to compute the similarity between each object instance and a given text query embedded in the CLIP space. This enables our model to respond to open-vocabulary queries in an instance-based manner. For example, the query “watch a movie” obtains the highest embedding similarity with the TV object.

Thanks to its zero-shot learning capability, OpenMask3D is able to segment instances of a given query object that might not be present in common segmentation datasets, such as “Pepsi”, “angel” and “dollhouse”. Our approach also preserves information about object properties such as affordance, color, geometry, material, and state, resulting in strong open-vocabulary 3D instance segmentation capabilities.



Elisabetta and Ayca are giving a live demo of OpenMask3D at the Google Research Booth at ICCV 2023 in Paris, France.

This opens up new possibilities for understanding and interacting with 3D scenes in a more comprehensive and flexible manner. We encourage the research community to explore open-vocabulary approaches, where knowledge from different modalities can be seamlessly integrated into a unified and coherent space.

Check out our project website (https://openmask3d.github.io)! Try OpenMask3D on your own scenes and let us know the most interesting object you are able to identify with OpenMask3D!

Do you want to learn more about OpenMask3D? Visit Ayca and Elisabetta during their NeurIPS poster session on Tue 12 Dec between 10:45 am - 12:45 pm CST, Great Hall & Hall B1+B2, poster #906!

Elisabetta and Ayca presenting OpenMask3D at the AI+X Summit in Zurich, Switzerland.

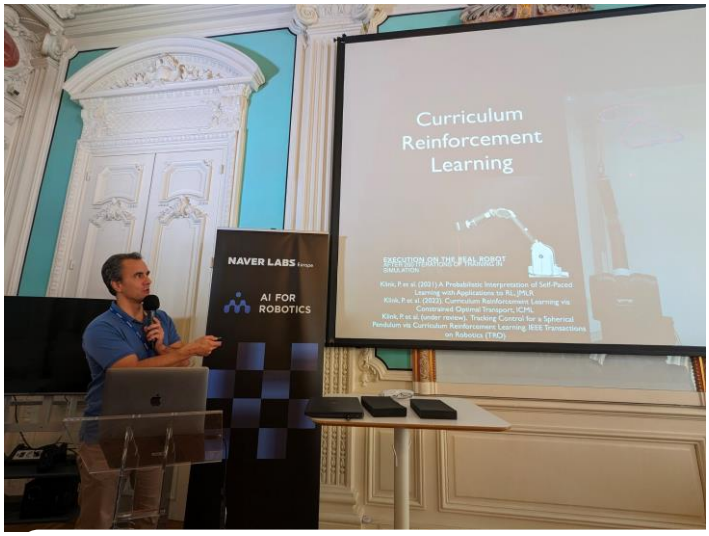
Francis Engelmann, Federico Tombari, Ayca Takmaz, Elisabetta Fedele (left to right), Robert Sumner (not in the photo) and Marc Pollefeys (not in the photo) are the authors of OpenMask3D.

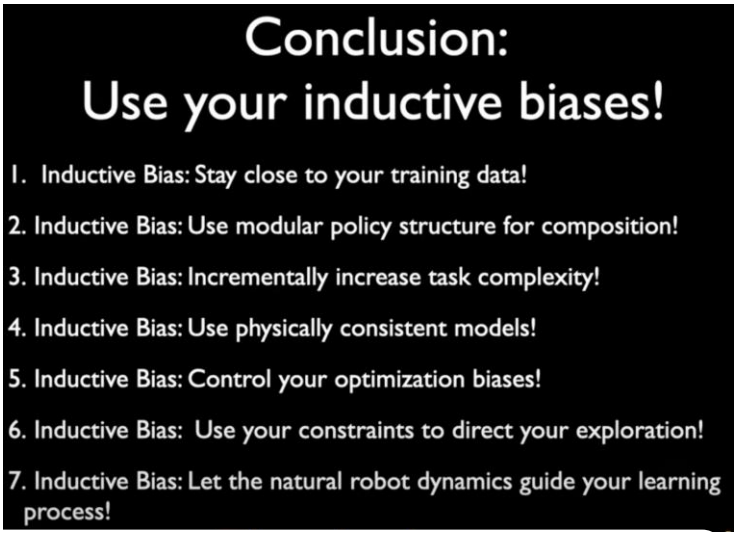
The third edition of this workshop was held on November 15th and 16th at the NAVER LABS Europe site in Meylan, France. The talks given by invited speakers explored how advances in computer vision, machine learning, scene understanding, natural language processing and related fields will make it possible to equip robots with AI components so they can operate and interact with us in the real world and become integrated in our everyday lives. Huge thanks to Christian Wolf and the Naver Labs Europe team for this awesome report!

by the Naver Labs Europe team

Inductive biases and structured representations

Central questions covered during the workshop were the current challenges in robot learning, in particular the generation and usage of large-scale data required for modern learning algorithms, learning in simulation and the corresponding sim2real gap. **Jan Peters**, Professor at **TU Darmstadt**, suggested a whole list of inductive biases which can enable successful learning including the creation of modular policies, incrementally increasing task complexity (curriculum learning), exploiting constraints in exploration and using robot dynamics during learning.



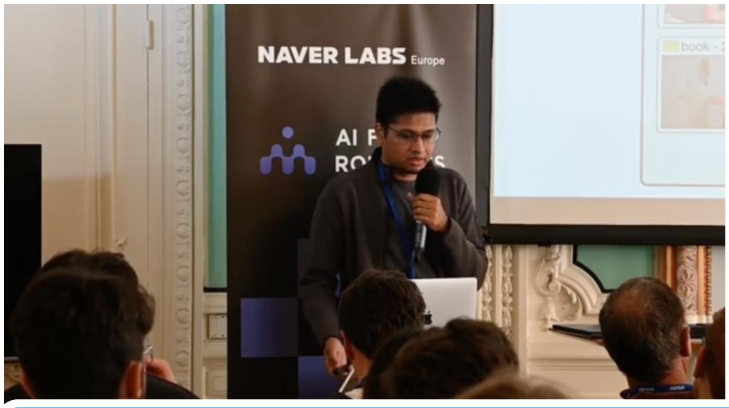


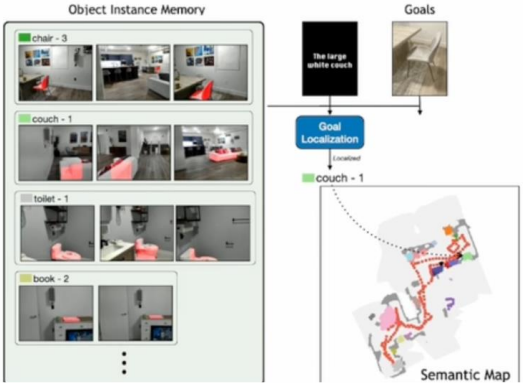
**Jan Peters (Technische Universität Darmstadt)
on how robot learning can benefit from inductive biases.**

In the same vein of exploiting inductive biases to improve the sample efficiency of the learning process, **Sylvain Calignon**, senior researcher at **IDIAP Research Institute**, emphasized the benefits of leveraging the structures of both the data and the geometry when addressing complex tasks in robotics.

In particular, he explained how data factorization techniques using Tensor Networks could be useful for global discrete/continuous optimization in hard planning problems, and how geometric algebra could offer an elegant and efficient framework to acquire skills from a small set of interactions.

Modular approaches were also suggested by **Devendra Chaplot**, researcher at **Mistral AI**, in particular for indoor navigation. He showed that this kind of structured decomposition is capable of being deployed in real environments of various types after being trained in simulation only. He also showed that

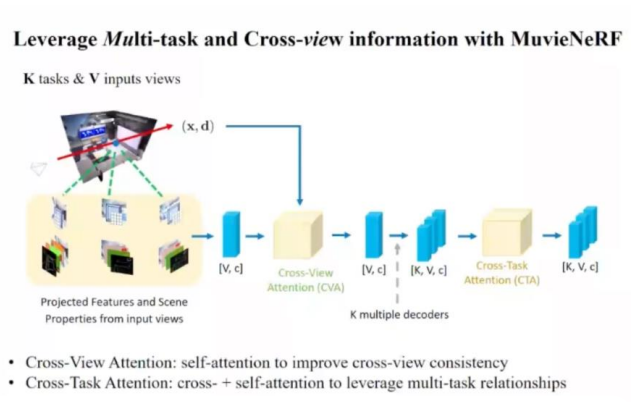




Devendra S. Chaplot (Mistral AI) on modular approaches for navigation, trained in simulation and evaluated in real environments.

the goal specification can be generalized using large language models (LLM) and visual language models (VLM), allowing goal specifications through coordinates, goal images, or language input, and targeting various different embodiments.

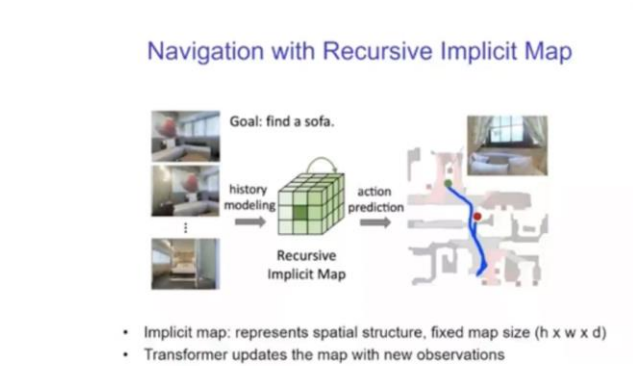
Structured representations were also suggested by **Martial Herbert**, Dean of **Carnegie Mellon University**, in particular for perception as a means to structure video input. He showed how self-supervised learning can be leveraged for learning with minimal supervision to exploit temporal consistencies and multi-view consistencies. This is enabled through slot-attention on the one hand or, on the other hand, extensions of implicit representations (NERFs) and multi-task learning.



Martial Hebert (Carnegie Mellon University) on structured representations for video understanding and robotics

Cordelia Schmid, research director at **Inria** and researcher at **Google**, also argued in favor of structured representations. She presented representations for end-to-end trained agents, used for perception, mapping and decision taking and, application-wise, covering navigation as well as manipulation.

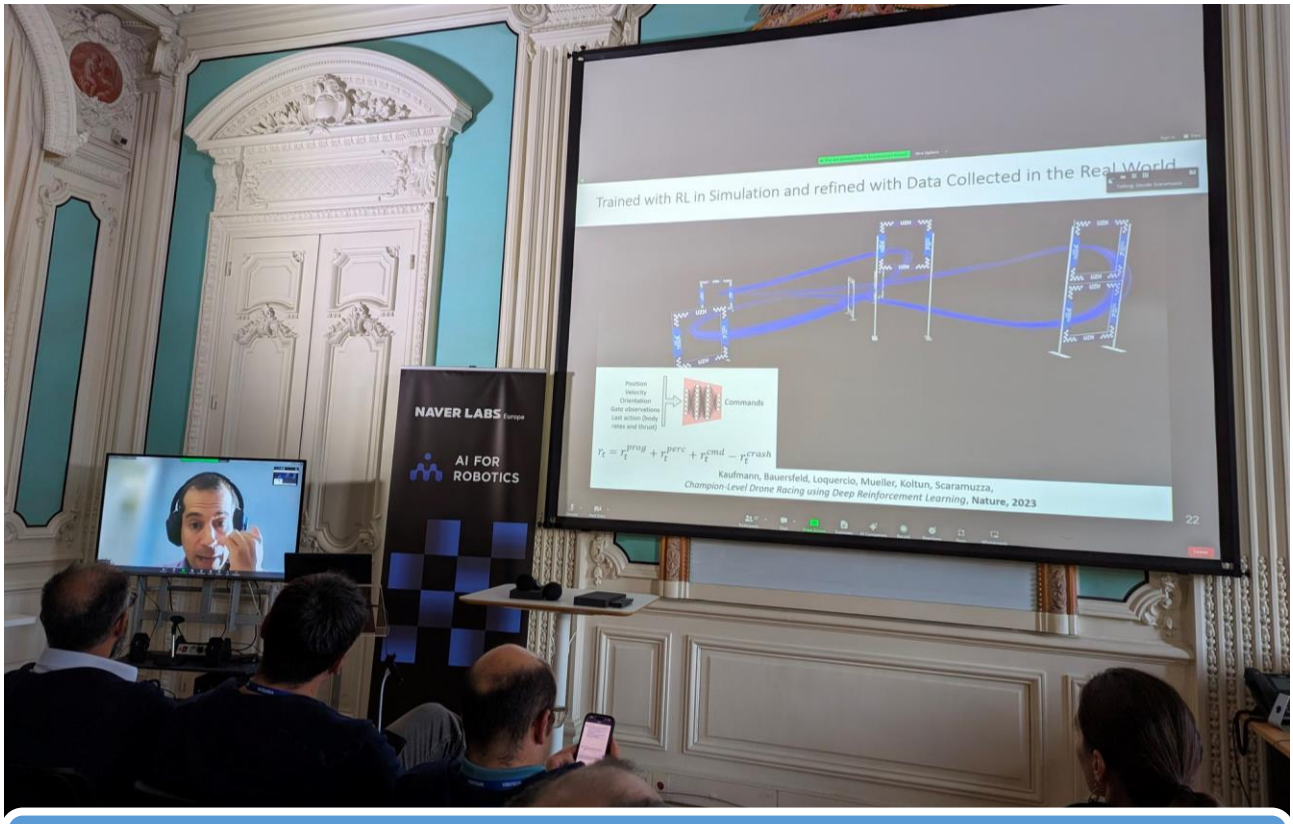
As an example she presented neural implicit maps, which are differentiable latent representations respecting projective 3D geometry and trained with imitation learning. A CLIP encoder ensures that the features are linked to semantic content and can be used with language queries.



Cordelia Schmid (Inria, Google) on structured representations for perception and decision taking.

Learning to take decisions

The question of whether learning should be done from demonstrations through imitation learning, or from reward with reinforcement learning (RL), was hotly debated with several speakers suggesting that RL should be avoided whenever possible due to its difficulty. While there seemed to be a consensus that imitation learning seems to be an easier learning problem, enabling learning for a large variety of problems, RL was also shown to be the method of choice in several successful cases. **Davide Scaramuzza**, Professor at the **University of Zürich (ETH)**, presented a spectacular success case, where RL was used to beat world champions in drone racing from first person views. Trained in simulation combined with recorded trajectories from real drones, this work, combining seven years of research of his group also showed that RL can beat optimal control for this task: whereas optimal control decomposes the problem into planning and control, limiting the range of behaviors, RL can directly optimize a task level objective and also cope with model uncertainty.

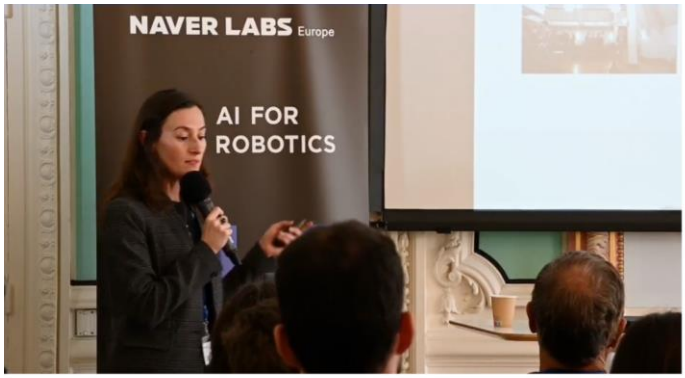


Davide Scaramuzza (ETH Zürich) on how RL beats optimal control for drone racing from first person views.

RL was also the method of choice presented by **Olivier Sigaud**, Professor at **Sorbonne University**, who insisted on social aspects in particular in situations where artificial agents are either teachers or students, in which case they require pedagogical teaching capabilities or pragmatic interpretation capabilities, respectively.

Oya Celiktutan, Senior Lecturer at King's College London has also been studying social robotics for several years working on robots assisting humans on the physical, social and cognitive levels. She specifically introduced a method for identifying clusters of individuals engaged in conversation, enabling the robot to navigate between these groups rather than within the same group of people (social navigation).

Laure Soulier, associate professor at Sorbonne University/ISIR lab, conducted a comprehensive review of the literature regarding the application of LLMs in robotics, with a particular focus on harnessing their planning capabilities to generate sequences of actions.



Oya Celiktutan (King's College, London) on social robotics and social navigation

Simulation as an infinite source of data

Gül Varol of the **IMAGINE research team at Ecole des Ponts ParisTech**, summarized recent work on bridging the gap between natural language and 3D human motions, with several examples on human motion control and synthesis from textual descriptions. The possible applications range from the control of humanoid robots to human-robot interactions. Motion capture is limited and expensive, and generative models are a solution to provide unlimited data. In this domain she stated that "*Leveraging physics would be the cherry on top of the cake, it would bring more realism but would not help the lack of data in terms of semantics*."

Generating large-scale data to compensate for the relatively small datasets available in robotics (typically millions of times smaller than the ones used to pre-train foundation models in NLP and vision) was also the focus of the presentation done by **Markus Wulfmeier**, Research Scientist at **Google DeepMind**. He illustrated how to semi-automatically generate diverse and relevant training data for learning navigation or visuo-motor control policies on two use cases, namely training a real robot in soccer and recommending the best routes in Google Maps.

Hao Su, Associate Professor of Computer Science and Director of the Embodied AI lab at the **University of California, San Diego (UCSD)** presented his work on modeling the 3D physical world for embodied intelligence. His team has spent many years building a full-stack system for robot learning, called SAPIEN, which provides physical simulation for robots, rigid body and articulated objects. He strongly believes that robot simulation is an important ingredient for the community as it allows scalability, replicability and fast prototyping and sim2real should be solved by making simulators closer to the real world.

The AI for Robotics international workshop is a biennial event organized by NAVER LABS Europe. The videos of the latest edition and the two previous workshops (2021, 2019) are available online. Again, huge thanks to the Naver Labs Europe team for this awesome report!

A Flexible Nadaraya-Watson Head Can Offer Explainable and Calibrated Classification

We are starting today a new section of Computer Vision News. How many papers that have been published are underrated? Did any of your past papers deserve better fortunes? This section will give a second life to neglected papers that are worth having a look at.

Mert Sabuncu, Professor at Cornell University, was kind enough to be the first professor to play the game. Mert thinks that this paper is underrated and has asked first author Alan Wang to tell us about this work. But first, Mert tells us why he thinks we need to have a second look!

Let's point out that Alan, PhD candidate at Cornell, is on the academic job market and he's a great catch!

Alan Wang

by Mert Sabuncu

In the fast-moving field of machine learning and computer vision, as we tackle new challenges (such as robustness, fairness, interpretability, calibration, and human-in-the-loop AI), we often suffer from a recency bias – our approaches are constrained by recently popular techniques and modeling choices. In this work, led by **Alan Wang**, we revisited a classic paradigm, known as the **Nadaraya-Watson (NW) estimator** and married it with more recent neural network architectures. To be sure, this has been done by others and thus is not a new approach. What we show here is that this so-called NW head offers unique advantages for **interpretability and calibration** and further promises a new approach for other important challenges we face, such as **robustness and fairness**. This is why I advised Ralph to start this new series of "underrated papers" with Alan's work, as it effectively demonstrates the potential of reevaluating traditional methods from a fresh perspective.

The paper was first published in Transactions on Machine Learning Research.

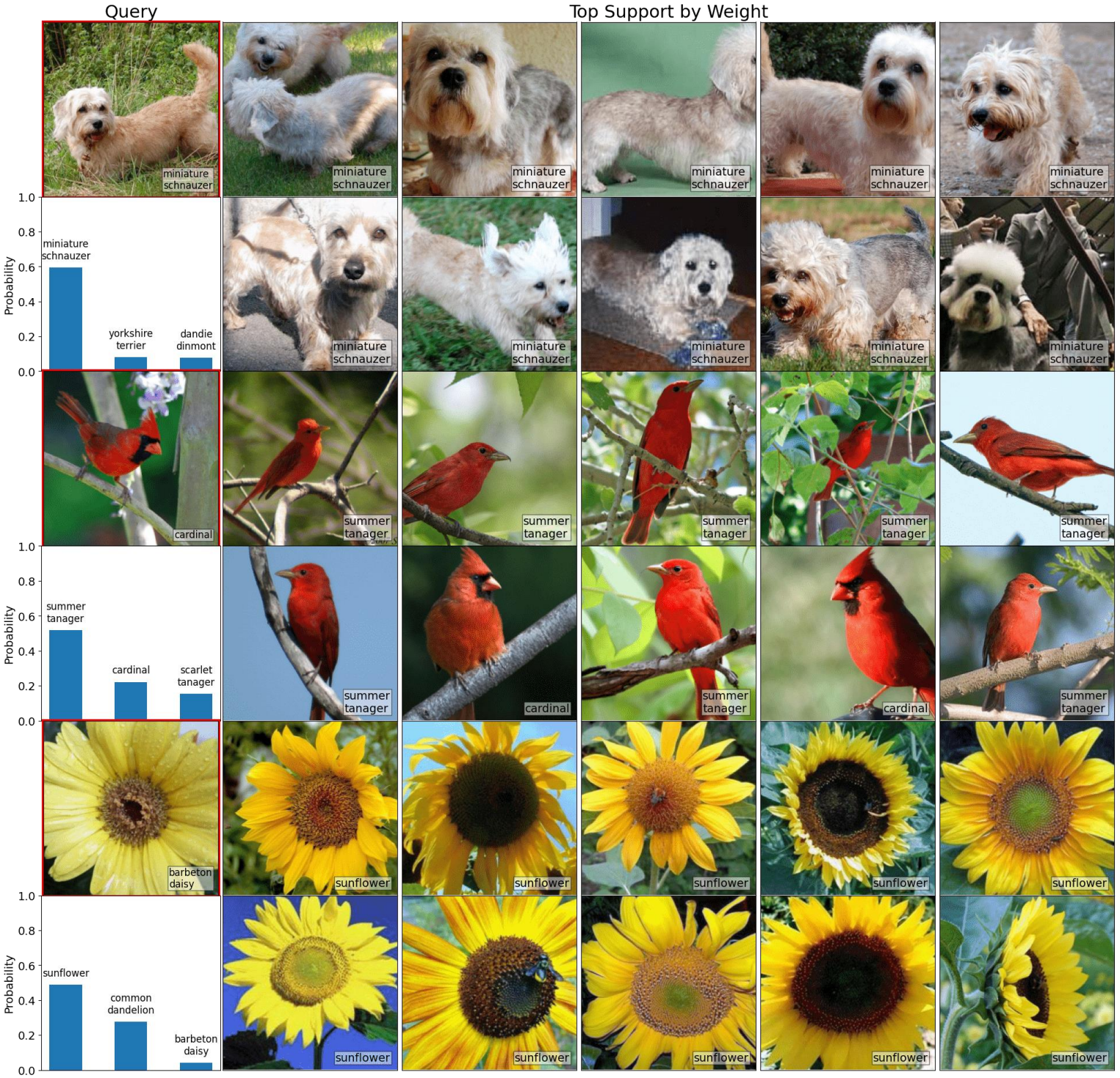
by Alan Wang

The dominant approach to image classification in computer vision is to feed the image through a neural network which extracts features of the image (e.g. ResNet, ViT), and then passes those features through a fully-connected (FC) head. This architecture is a black box and is non-interpretable, since a human cannot understand how the model arrived at its decision. In addition, the predictions are often poorly calibrated, meaning that the model can be overconfident about its correctness. This can mislead humans as to how trustworthy the model really is.

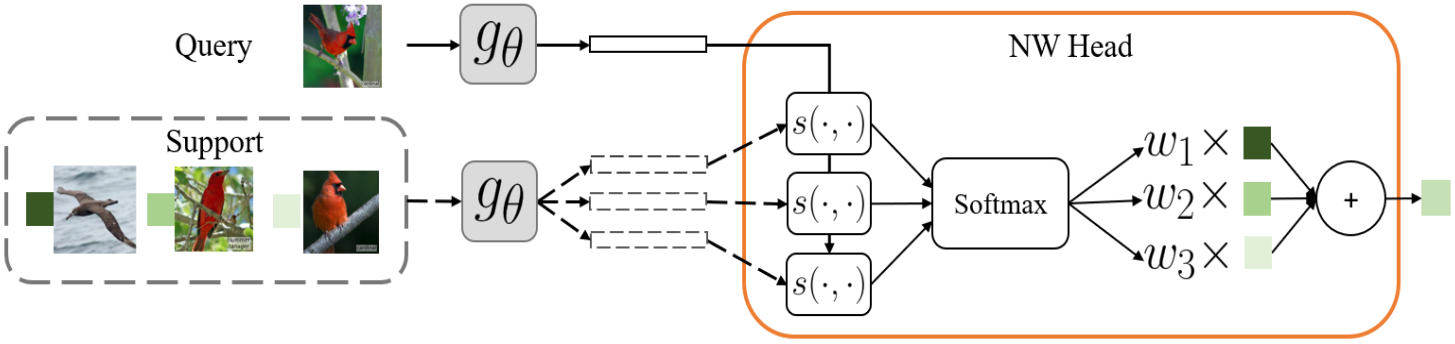
As an alternative, we propose the Nadaraya-Watson (NW) head, which we show provides better calibrated predictions and better interpretability and explainability while performing comparably to the FC head. Essentially, the NW head can be viewed as a soft variant of the nearest-neighbor classifier. For each query image to be classified, we assume we have access to a "support set" of real samples and their labels from the training dataset. To produce a prediction, the NW head passes both the query image and all support images through the feature extractor, and subsequently computes a similarity between the query feature and each support feature. These similarities are normalized to weights, which are in turn used to compute a weighted average of the labels in the support set used as the final prediction (see image next page).

How is the support set constructed? During training, one can sample the query and support set randomly from the training set at each training step. During inference, the user has a high degree of flexibility in how to choose the support set. In our experiments, we try randomly sampling from the training set, using the entire training set, and performing within-class clustering to construct the support set. Each of these inference "modes" has its own advantages and trade-offs, providing flexibility to the user.

Experimentally, we find that the NW head exhibits comparable to superior performance to the FC head while providing better calibrated predictions. These characteristics are essential in real-world deployment scenarios.

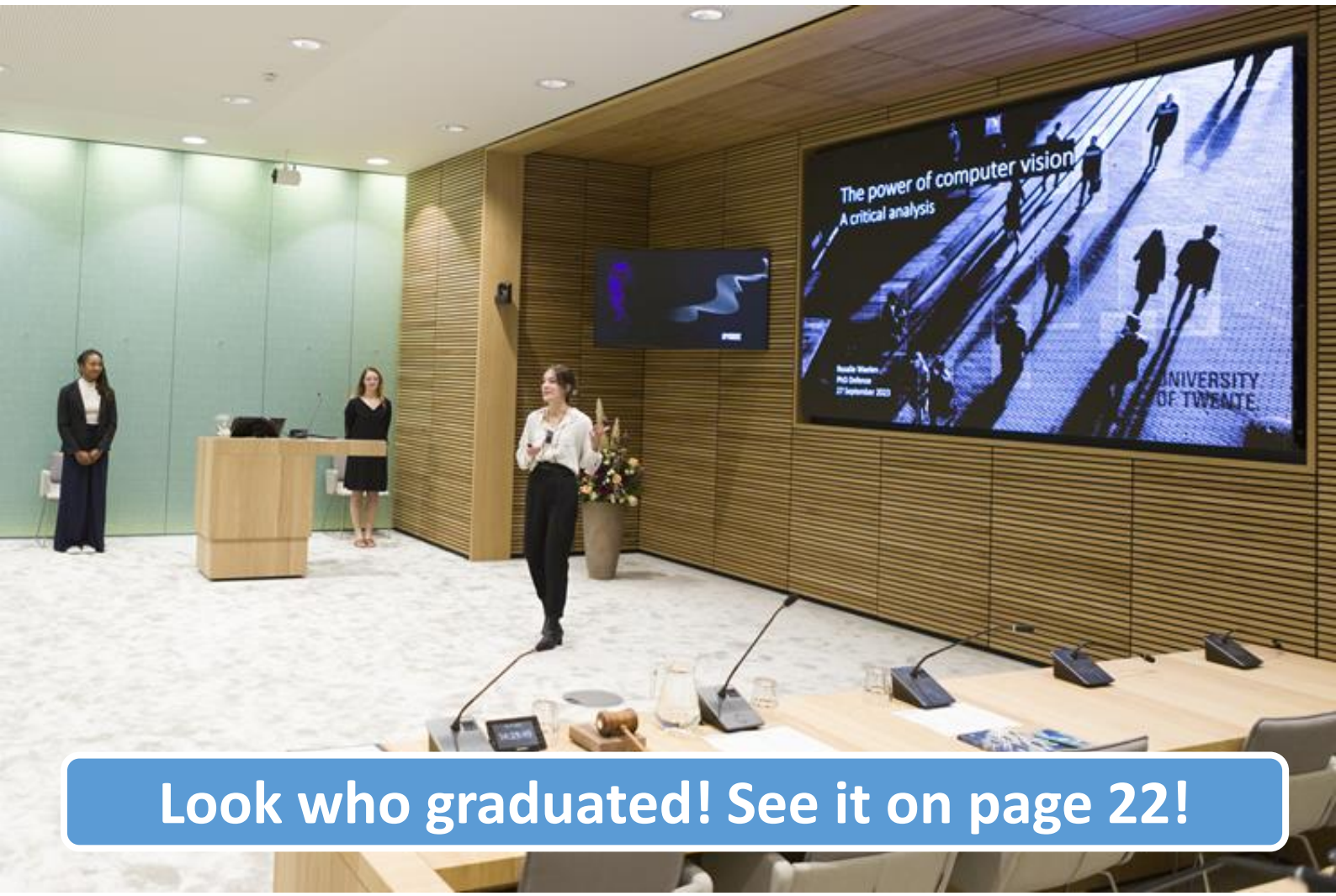


In addition, the NW head allows for two means of interpretability/explainability not possible with the FC head. The first is interpreting the support set weights, which directly correspond to the degree of contribution a specific training datapoint has on the prediction. By interrogating these nearest neighbors, a human can see which other real images are driving the model's decision making (see image next page).



The second is our novel concept of “support influence", which can highlight both helpful and harmful support examples and can be used as a diagnostic tool to understand and explain model behavior. Computing the support influence is exact and computationally efficient, in contrast to computing influence for the FC head which requires approximations and assumptions on convexity.

Are you a computer vision / AI / medical imaging professor? Do you know of any paper that would have deserved better fortunes? Tell us about it!



Look who graduated! See it on page 22!

Maria Koskinopoulou is an Assistant Professor in Robotics and Computer Vision at Heriot-Watt University in Edinburgh, Scotland.

Maria, tell us about your work.

I'm mostly doing robotic manipulation. I'm working with arms with robotic manipulators, trying to solve control strategies and path planning for all the arms in order to perform pick and place movements and novel tasks in industrial environments. For example,

Read 100 FASCINATING interviews with Women in Computer Vision

Check out the video!

for assembly or in medical robotics as well as other applications. Combining also computer vision.

Why did you choose this field? Was it your specific ambition?

Hmm, very interesting question. I did a master's in neuroscience when in Greece, in Crete. From this master's, I studied different topics related to computational neuroscience. At that point, there was an opportunity to participate in a national project funded by the Government of Greece about how to develop a computational model of how to teach robots the latent behaviors and how to coordinate their actions with humans. This project was particularly focused on manipulation tasks and how to teach a robot to act in an environment in a novel way. I started working on this project and started to feel like I liked this field of robotics, and I focused mostly on the manipulation part.

Do you find it more difficult to teach robots or to teach humans?

[*she laughs*] Very interesting question! I think it's particularly difficult to teach humans because they interact in a different way. We're not at such a level I think in our human-robot interactions. Robots can somehow have limited interaction towards a human, so you will not be asked very challenging questions, and you will not have this physical and seamless interaction and communication. I think you feel kind of safe that you have the primary role, and you can lead the communication and collaboration in your own way. However, with humans, this is not the case all the time!

Maybe we should transform all our students into robots!

Yes, this is the safest way! [*Maria laughs*] Also, for example, in learning from demonstration, like imitation learning that I was doing for my PhD, this is the topic that studies these kinds of relationships. They're inspired by the way we're teaching humans, especially children, how to learn and how to interact with their environment – how to learn to walk, how to learn to answer, and how to imitate our parents and relatives while we're growing. These are the kinds of skills and abilities we're trying to transfer to the robotic platforms.

With respect to manipulation, are there times that you are able to teach something and times that you try to, and it does not work?

Definitely! Yes, yes, yes. Many times, I'd say that they're not working! [*laughs*] These are everyday challenges that we face. I'm trying to be patient, and it's very interesting that now, I have the same feeling from my students that every day, they're kind of disappointed. They're facing these kinds of situations where one day everything is working, and you have a very good

"They inspired me and gave me the first mentoring and all the assets that made me what I am now!"

setup, you're going to do experiments, and everything is working perfectly, and I don't know why, the next day you can't run any experiments. The cameras, the sensors, and the robots aren't working. You have to calibrate the system from scratch. The students say, *"Why? Yesterday, everything was functional, and why is it like this today?"* I'm trying to be patient and say that these things can always happen, and tomorrow will be another day, and you will solve all the problems and debug everything.

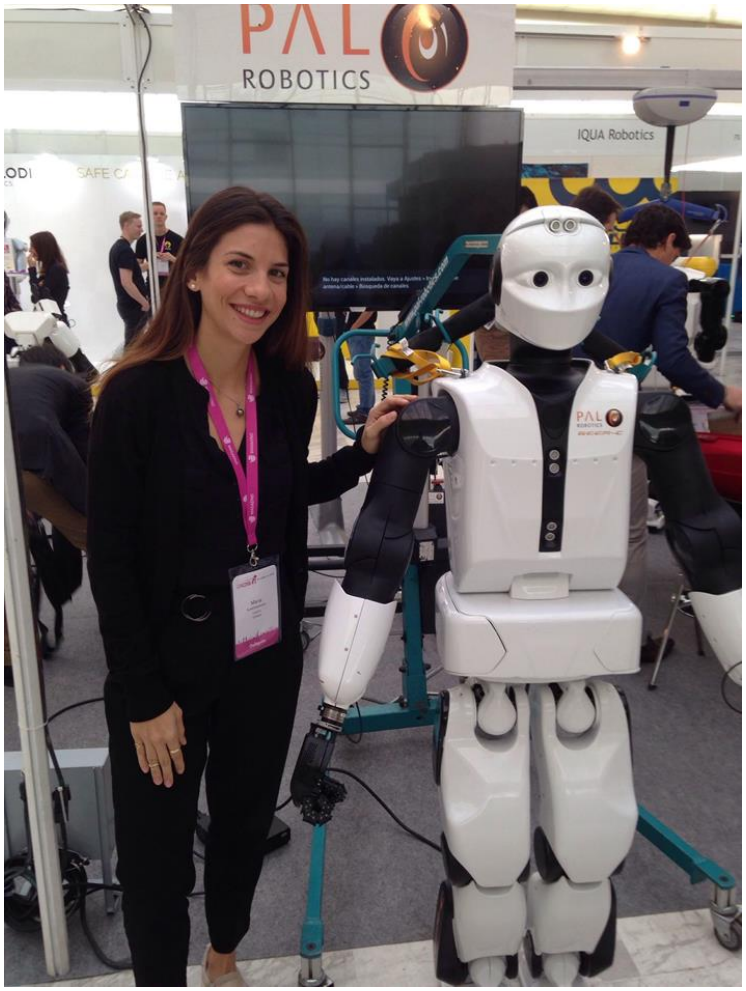
I am sure you sometimes use other words, like "Stupid robot!"

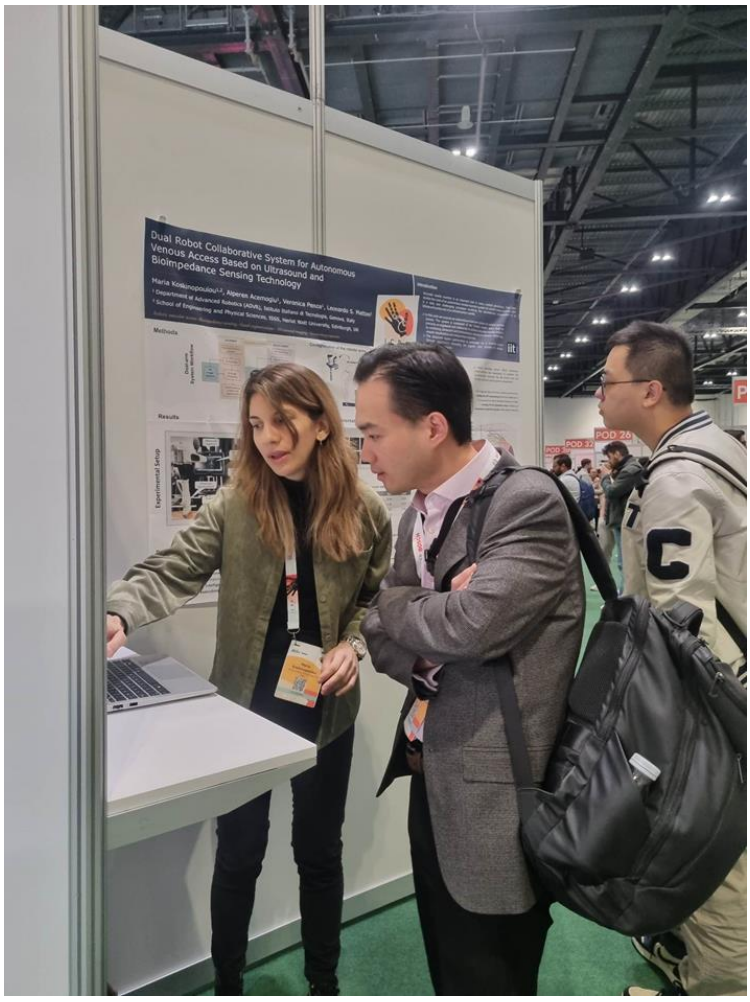
[*Maria laughs*] No – actually, I'm trying to think. Robots are very intelligent. We can transfer our intelligence to the robots, which is a very useful thought, and it helps me to be fair about the robot. I'm trying to think that, okay, it's like an equal agent. We're interacting in a very equal way. I'm trying to be polite and think very positively!

You sometimes try to teach the robot again and again, but it still does not work. If I could give you a magic wand, what kind of manipulation or action would you choose to be successful?

Okay, so for the last almost three years, I was working in a lab of IIT in Italy, particularly focused on medical robotic applications. I saw from the inside how important this field is. I think this is kind of the future. It's very important to have intelligent robotic systems applied to surgical applications and be able to perform surgery in a safe and precise way and not be particularly dependent on human interventions.

For example, I was trying to develop smart medical devices for intravenous access. This is an intelligent device with which you can perform this very common but also difficult intervention for blood examination and delivering drugs into veins. It has quite a high percentage of failures, especially in kids, and you can have a very bad experience with it, especially in children or elderly people who may have hard veins, and then you need to do it again and again. In this





situation, we're particularly dependent on the clinician's skills and the skin characteristics of the patient. It's important to have a robotic device that can do this with just one shot, without failures, by helping the clinician and the nurse to visualize our veins and be very precise during this operation. I would say the next thing I'd like to do is have more experiments towards this with animals and even with humans, trying to make this happen and bring it to the market.

I have interviewed many Greek women in this series of Women in Computer Vision. That is not by chance – there are very many successful Greek women in science. Why do you think that is?

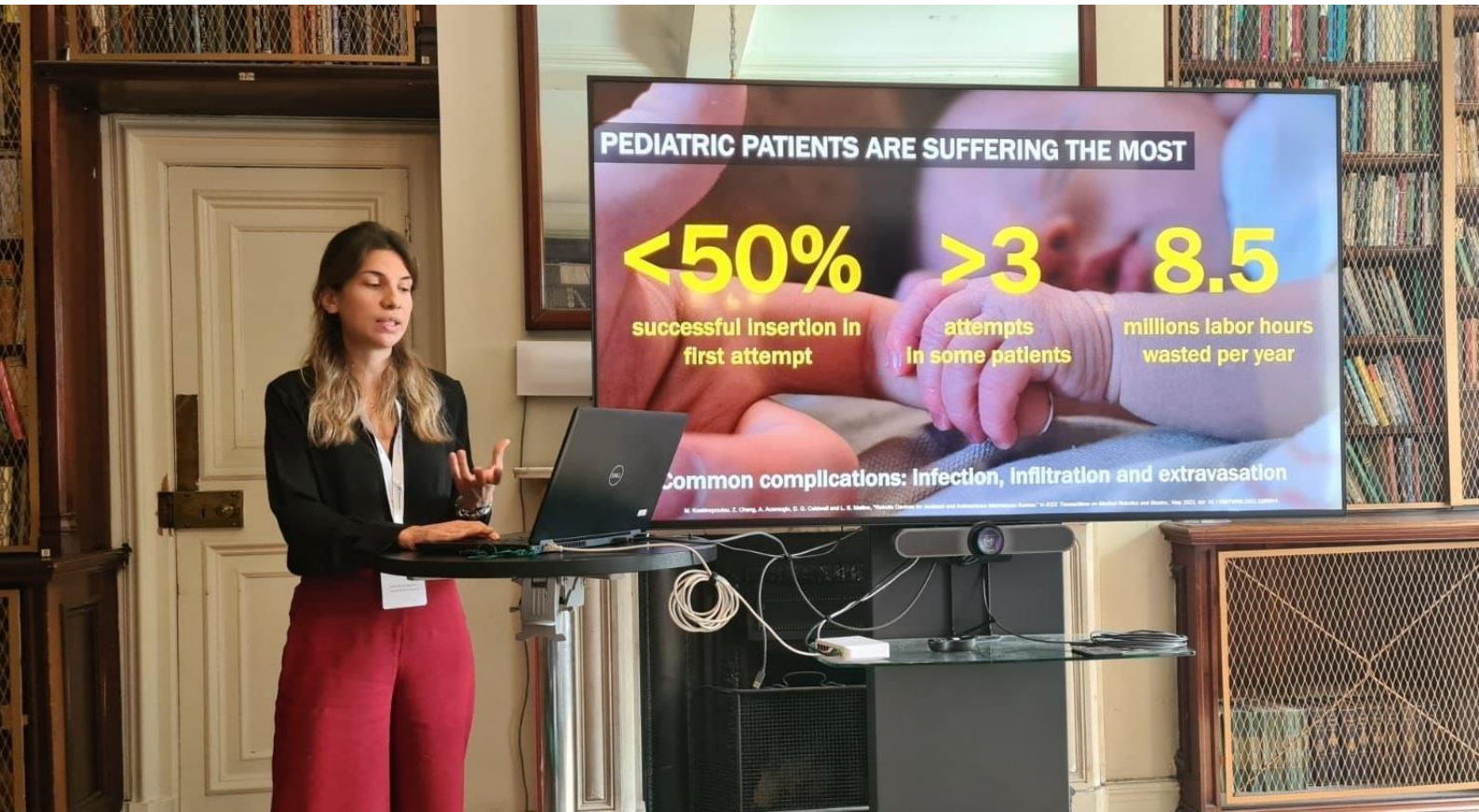
Ah, this is very nice and a good thing to notice. I'm very happy about this actually and very proud. I don't know! [*laughs*] I can't say why this happened, but I think in Greece, we have a very high level of education and most probably research, and we're kind of curious about research and progress. We have different interests. We're trying to succeed in different fields. In engineering, there are many good researchers, especially women.

Is there one Greek scientist from the past that you admire in particular?

There are many Greek scientists I admire. First of all, my first professors, my supervisors in Crete in Heraklion, where I pursued my PhD and master's, Prof. Panos Trahanias and Prof. Antonis Argyros. They inspired me and gave me the first mentoring and all the assets that made me what I am now. I'm particularly proud of being part of this group at that time and that I have the chance to work with them and continue my career up to now.

We have spoken about the past and we have spoken about your present work, but we are missing something about the future. Where are you going, Maria?

[*she laughs*] Kind of the future is now for me, because I've just started here. I started in this group in May, so I'm feeling quite new. I'm starting my own lab, hiring new people, participating in master's projects, and hiring PhD students. I'm very happy to be here initiating a new lab and putting the first objectives on how to see myself in some months and years. For now, I would like to see my future here – my short future, at least [*laughs*] – trying to bring new researchers here and have fruitful collaborations and groups that can take forward our research in robotic manipulation and perception.



Rosalie Waelen is a philosopher and applied ethicist, who recently completed her doctoral studies on the ethics of computer vision, at the University of Twente in the Netherlands. The title of her thesis was *"The power of computer vision. A critical analysis"*.

Rosalie will continue doing research on the ethical and societal implications of AI, as a postdoctoral researcher at the Sustainable AI Lab of the University of Bonn in Germany. Congrats, Doctor Rosalie!

Following the rise of deep learning, the potential ethical and societal implications of AI development and use have also become a hot research topic and a pressing issue on the agendas of policy makers. Since computer vision is a subfield of AI, the ethics of computer vision can also be treated as a subfield of AI ethics. The goal of Rosalie's Ph.D. research was to explore the ethical and societal impact of computer vision.

To address the ethics of computer vision, Rosalie developed a new approach to AI ethics, which she called "a critical approach to AI ethics". This critical approach entailed analyzing the power dynamics involved in AI development and use. The main goal of this approach, and of Rosalie's thesis, was to determine the variety of ways in which computer vision (potentially) affects human autonomy and emancipatory progress in society. After all, technological progress is not real progress if it does not empower people and support their autonomy.

In her thesis, Rosalie offers a general overview of the different ways in which computer vision tools could harm human autonomy and emancipatory progress. However, she also discusses some specific topics in-depth. One such topic is impact of cameras on people's behavior, norms, and identity

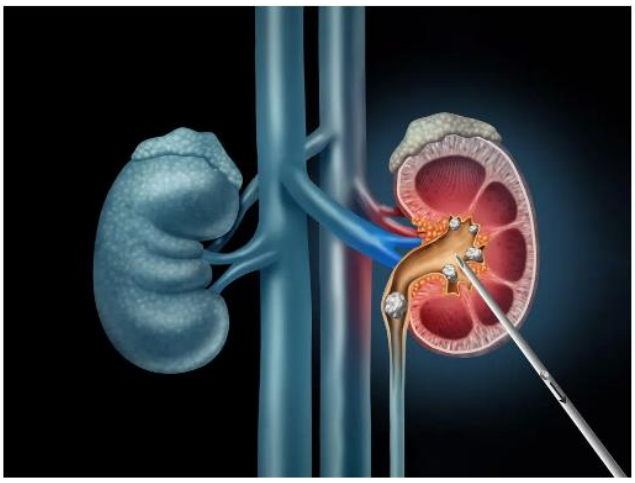
formation. Throughout history, cameras have enabled big governmental and corporate institutions to exercise control over people's behavior and self-development. The company Kodak, for instance, actively taught people to associate picture taking with happy moments. It is through the advertisement of Kodak, and later through the influence of social media, that many of us now understand certain moments as something we should take a picture off and feel like the event was less real or valuable if we did not capture it on camera. On the basis of this historical analysis, Rosalie warns us to be aware of the potential for social control in the context of emerging smart camera applications.

Another topic that Rosalie researched in detail is the idea that facial recognition tools cannot only misrecognize people in a technical, factual sense, but also in a social and legal sense. When facial recognition tools misrecognize someone's identity or characteristics, this can simultaneously be a failure to recognize the person's social and legal worth. This is a moral problem, not only because it is discriminatory, but also because it can harm people's sense of self-worth.

Rosalie's PhD research resulted in various publications as well as a blog series on the ethics of computer vision. Reach out to Rosalie for more information.

PCNL – Planning and real-time navigation

Urolithiasis, or kidney stones, is a common pathology affecting nearly 10% of the population in the USA. **Percutaneous Nephrolithotomy (PCNL)** is a minimally invasive urology procedure intended to **remove stones from the kidney and proximal ureter** which are less amiable to other endourological modalities, specifically large stones, or in the presence of abnormal anatomy. This procedure requires real-time ultrasound and/or X-ray guidance and is usually preceded with a CT scan for diagnosis.



Gaining access is the first and most challenging part of the procedure. The tract needs to be created in the exact direction and depth of the kidney's collecting system and in close proximity to the stone. This becomes more challenging in cases where multiple stones need to be targeted, and it is preferred to avoid the need for a second tract. Hence, obtaining a precise tract is essential for a successful and safe procedure.

Larger stones in the kidney and proximal ureter need to be treated with PCNL (Percutaneous Nephrolithotomy), a minimally invasive urology procedure intended to remove them. It is today a very effective procedure, as it can be operated via a relatively small opening with a reduced time of hospitalization. There is also a version of it called mini-PCNL, a tubeless procedure with an even smaller port of entry, while maintaining a similar rate of efficacy.

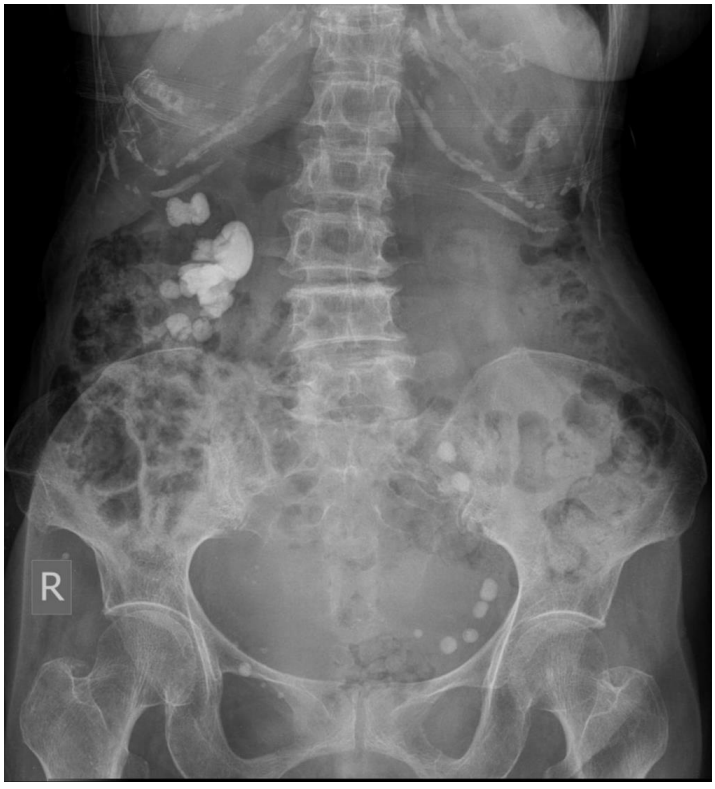
More surgeons would be able to perform PCNL if the access part were made simpler. That is one of the most challenging parts: it starts at the moment of the initial puncture until a path is created from the skin to the kidney itself in the correct location. That would make the surgeon's learning curve improve faster and at the same time more post-op complications could be avoided. As things stand today, the surgeon takes all decisions regarding the access out of his experience and judgement, aided by fluoroscopy or ultrasound.

The intervention can be delivered when the patient is prone (lying on the stomach), which is quite challenging for anesthesia; or supine, when the patient is face up,

which is simpler. When the needle is inserted in the kidney, the surgeon may be assisted by fluoroscopic guidance to perform the best possible needle insertion - the right angle and the tracking of the needle during the insertion; when ultrasound guidance is preferred to avoid radiations, some more dexterity is required from the surgeon, since less of the anatomy is visible. At the end of the procedure, the surgeon might choose a tubeless or fully tubeless solution, which avoid the insertion of a catheter to serve as a drain (nephrostome) for a few days.

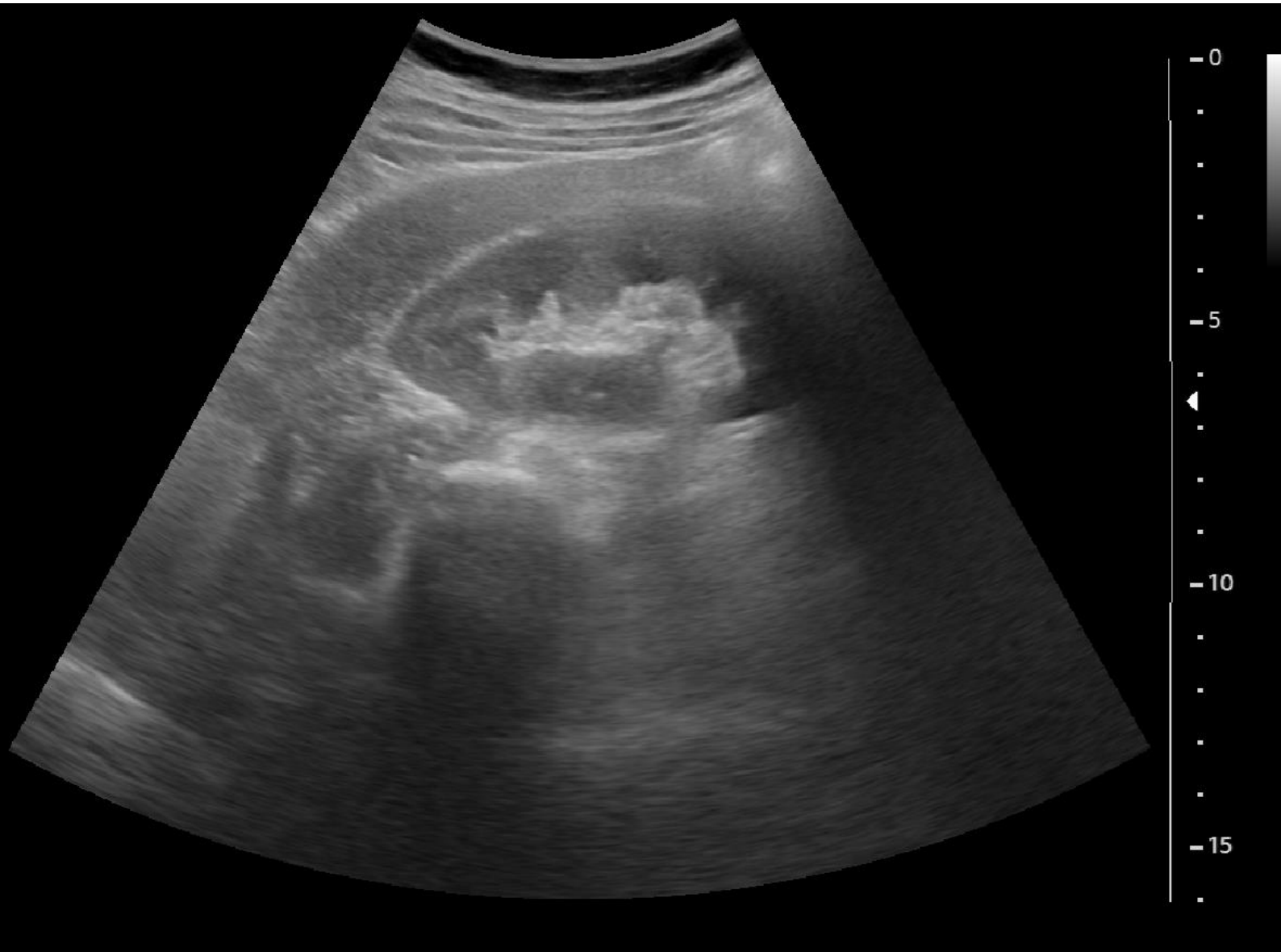
This procedure is quite common and successful. It is possible though to improve it. In particular, we suggest to improve the access part of PCNL. Access is one of the most challenging parts of the procedure, and we believe that dramatic improvements can be made by integrating artificial intelligence solutions combined with hardware. This will procure better outcomes and easier procedures even for less experienced surgeons. Less punctures will be needed since the number of incorrect accesses, involving removing the needle and puncturing again, will be much lower. Less punctures lead to less complications in the way.

The surgeon must choose very carefully the access route for the needle. How can we ask computer vision and AI to improve this choice? The first improvement is in



the planning part: the surgeon can look at the CT and select – based on his/her experience – the best way to puncture, from which part of the body and to which calix. The choice lays with the expert, which AI can assist but not replace, of course. AI algorithms can also help during the procedure itself. Using special tools, it can show where the needle is, where the calices are, and whether the trajectory is correct according to the plan. It is crucial that the needle is inserted into the right calix. When the surgeon uses fluoroscopy, computer vision analysis sees the needle from different angles and combines the views into one and conveniently communicates this information to the surgeon, assisting him/her to make sure that the access is done correctly. When the surgeon uses ultrasound, the image is not straightforward to interpret. In this case too, AI assists the surgeon to make informed choices.

How AI does this? There are a few tools that can be built using AI. The first one is segmentation of the region, of the needle, of the kidney

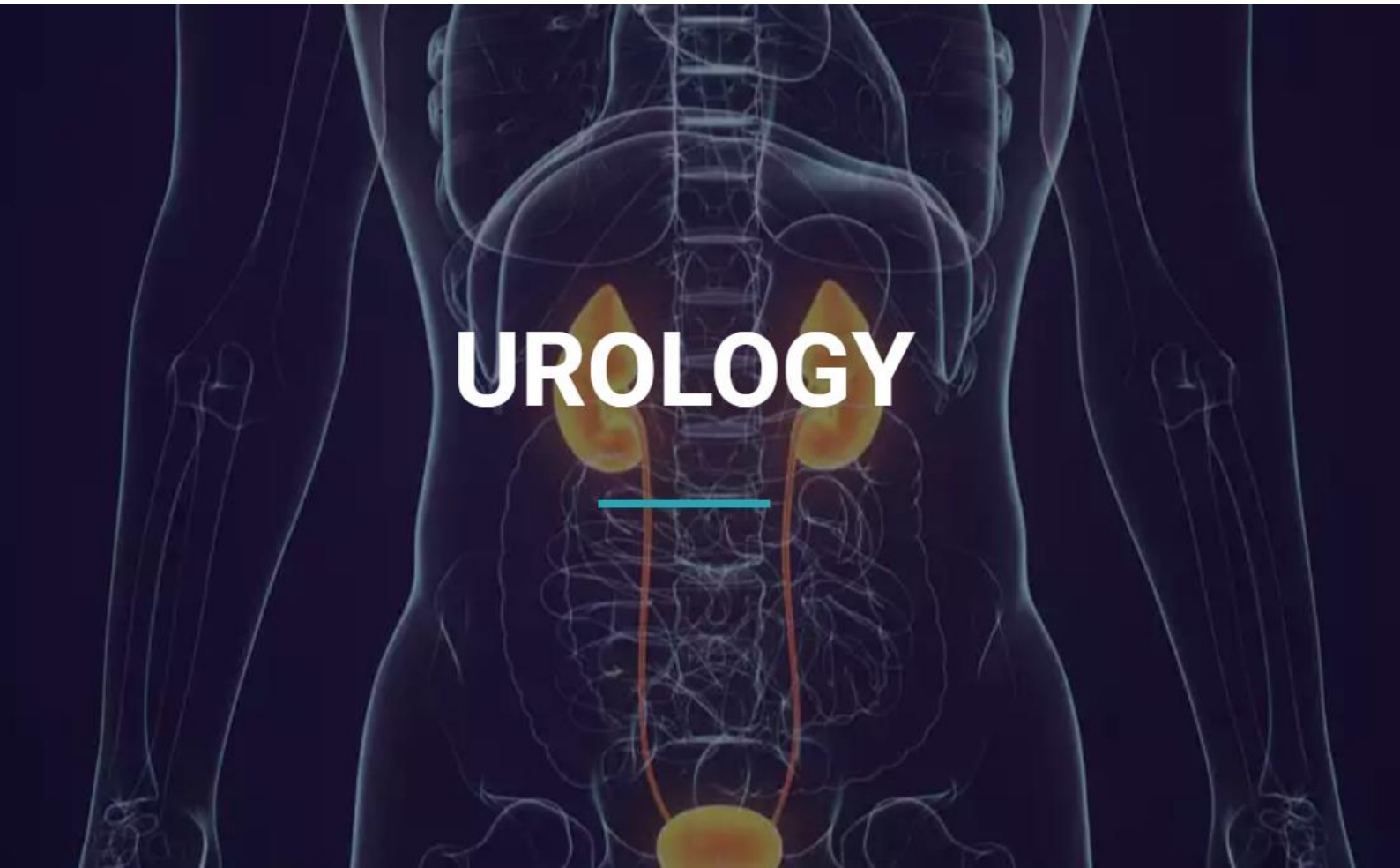


itself. Additionally, computer vision can merge two images taken with fluoroscopy and provide 3D information: what would be complicate for the mind is made much easier by mathematical formulas. When we have additional external hardware **like electromagnetic trackers (EMT)** or optical trackers, we can add this information and compute the position of the needle, even when it is not clearly visible in the image itself. We can also fuse all this information to provide additional assistance to the surgeon. If the needle is beneath the ultrasound plane, this info will guide the surgeon to point the ultrasound probe in a better direction.

AI algorithm should be the work of expert engineers: providing correct information to the surgeon is a major concern and any misleading info about the exact location of the needle is potentially dangerous. AI and the surgeon do not replace one another: they should complement each other and work together. The key contribution of AI is in the detection of relevant anatomical structure and the needle, in the mathematical formulas for 3D fusion and in the combination of 3D data coming from external sources, like an EMT or Multiview fluoroscopy images.

More articles about AI for Urology.

Contact **RSIP Vision** to learn more.



From Research to Hospital

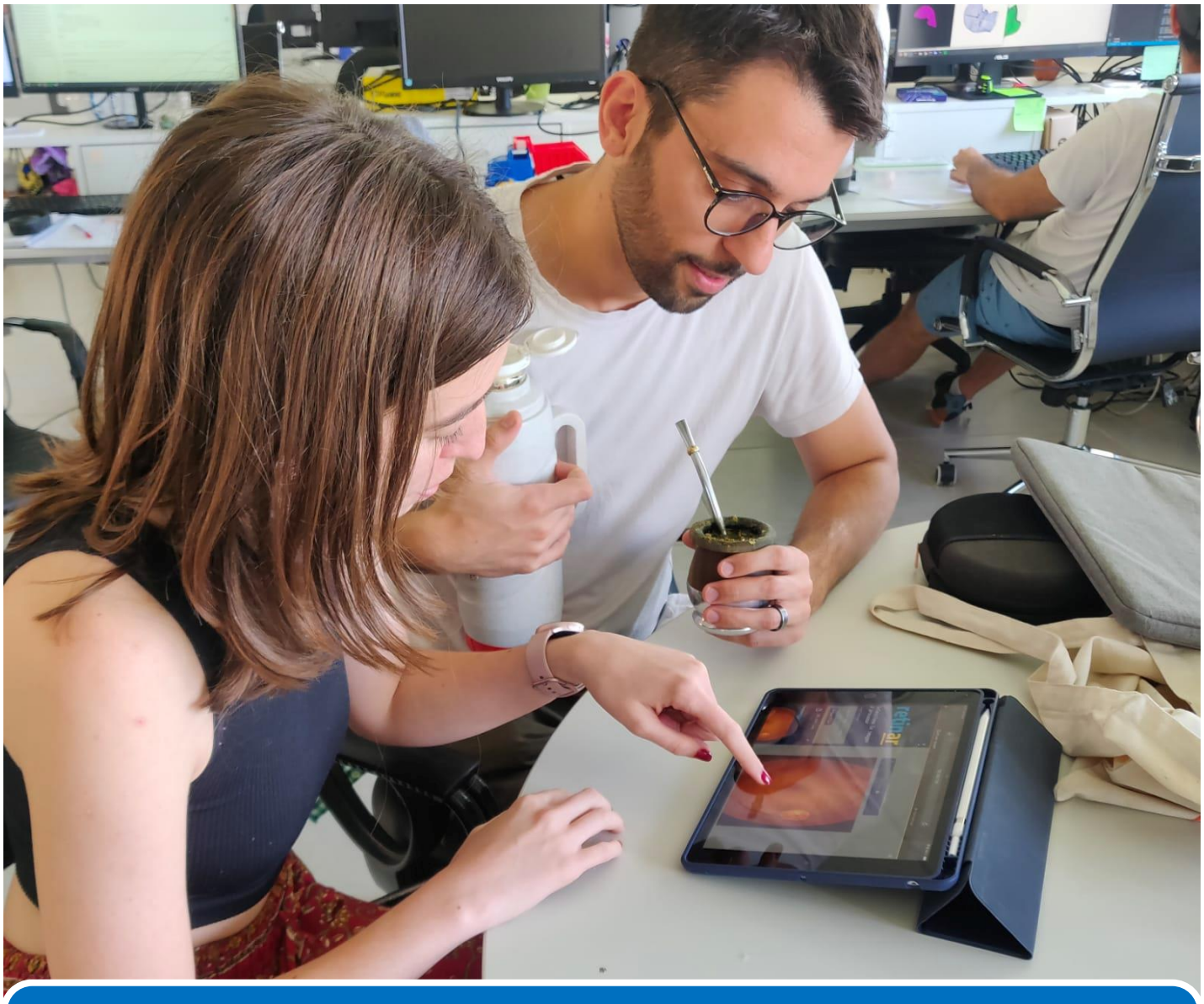
by Christina Bornberg
@datascEYEnce

Hello everyone! I am **Christina** and I interview different researchers in the field of deep learning in ophthalmology as part of the **datascEYEnce** column! This time I had the pleasure to talk to **Nacho**, who has a very important message for the community! No matter whether you are a first year PhD student or a senior researcher - make sure to go the extra mile for creating a product out of the research. And if you currently don't have the capacity for it, keep it at least in your mind that our work is not just about playing fancy maths with medical data, but we need to make those algorithms used by society.

featuring José Ignacio "Nacho" Orlando

Nacho's research journey started with a PhD in Computational and Industrial Mathematics at **UNICEN in Tandil (Argentina)**, with internships at **Inria in France** and **KU Leuven in Belgium** where **Matthew Blaschko** introduced him to machine learning for ophthalmology. His postdoc at the **OPTIMA Lab (Medical University of Vienna, Austria)** also influenced his current work - it was fascinating for him to see how the results of papers transitioned into tools that end up in the hands of clinicians.

Fast forward, in 2019 he came back to Argentina and currently is an Associate Researcher at **CONICET**, he's part of the **Yatiris Lab at Pladema-UNICEN**, and director of **AI Labs** at the US company **Arionkoder**. Now it's his moment to turn research into a product by leading the **retinar (short for Retina Argentina) project**. And what would be a better use case for showing the world that you can make anything happen if you set your mind to it, than a project conducted in a country with limited research grants and infrastructure and asymmetries in access to public healthcare!



Nacho with one of his PhD students, Eugenia Moris. She's showing to him her latest results on optic disc and cup segmentation in fundus pictures using semi-supervised learning.

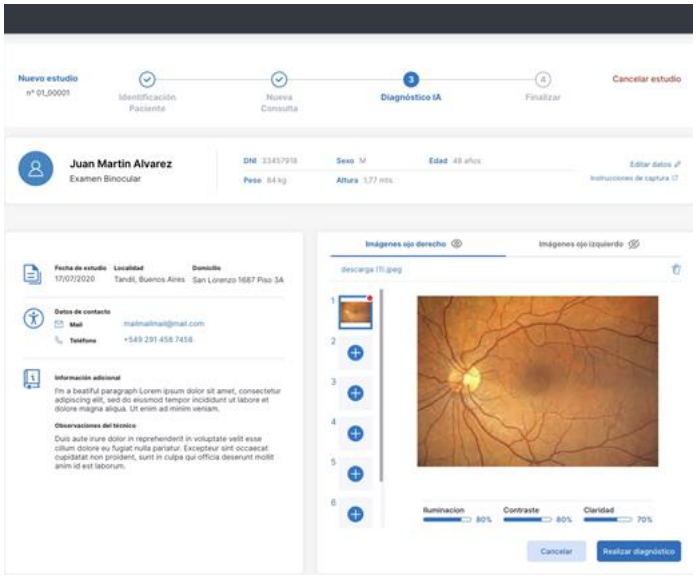
Firstly, in order to go the extra mile from research to product, it is important to find a clinical partner. In Nacho's case it was the Ophthalmology Department of the Hospital de Alta Complejidad El Cruce, in Florencio Varela. Clinicians often already have an idea about what would make their life easier. Involving them is a very important part since we don't want to build something they don't need or don't know how to use.

Together, they decided to create a screening and telemedicine platform for patients and clinicians. The goal is to not have a single center to inform remote cases but making it possible for every clinician to register to their platform as an expert. Doctors then get connected with patients without necessarily being on-site, for making diagnosis and suggesting treatments.

Furthermore, in case treatment or

further imaging is required, the closest facility that fulfills the requirements gets suggested.

And the best part about the platform is that **all the amazing deep learning research we all work on daily comes into action.** Different algorithms are combined to analyze both image and text data to give additional support to clinicians. The result is a PDF-summary report with medical images, highlighting regions of interest alongside other clinical information.



So what kind of algorithms are in use or currently in progress? The list is actually quite long and ranges from classification to segmentation up to the use of LLMs. Supervised, self-supervised and semi-supervised learning strategies allow the use of labeled, unlabelled and weakly labeled images.

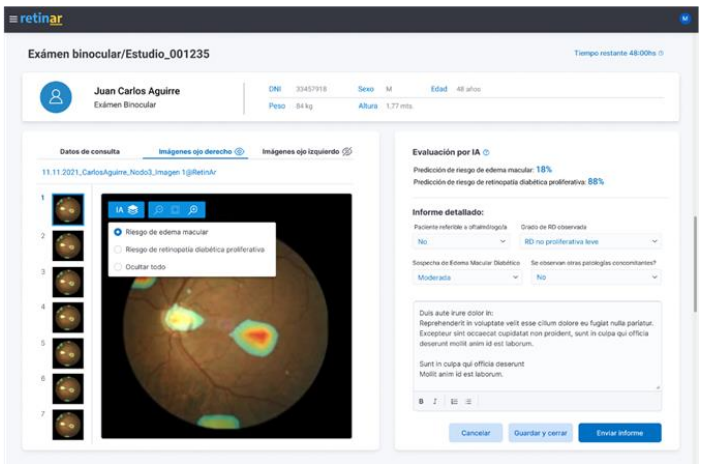
Let's start with **classification**. Grading of currently five main eye diseases in fundus images is (or will become soon) available in the tool: diabetic retinopathy, age-related macular degeneration, glaucoma, hypertension and cataracts. The approach used is quite simple: a ResNet-18 with smart data augmentation and with standard class activation maps. The main goal was to make it as robust as possible, specially to changes in the acquisition devices, which are expected to differ in the social context where the platform will be deployed. Hence, the team had to make sure to get as much variety as possible from different databases.

Next, **image quality assessment** is of great importance for clinicians since "ungradable images" are … well … ungradable. Here, Nacho experienced that the current research doesn't align with what clinicians want. Usually, research focuses on binary classification or in the best case classifying the images into usable, acceptable and rejectable. But technicians don't like that. Instead, giving information on contrast, brightness and clarity of an image is better accepted, as it might help them to correct the acquisition protocol right away.

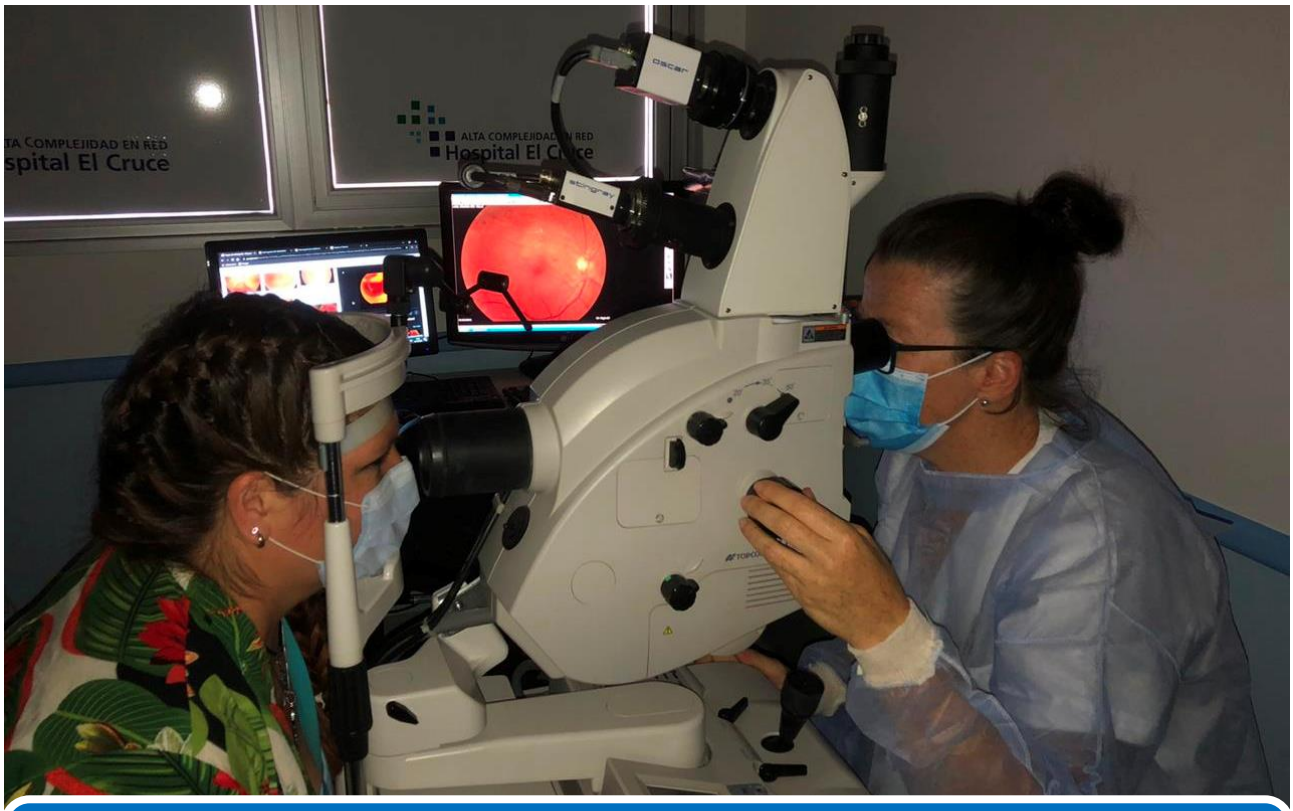
Moving on to **segmentation**, the current focus is optic disc and cup segmentation, which plays an important role in the diagnosis of glaucoma. Nevertheless, the algorithms that Nacho and his team are creating will soon enough be extended to a variety of lesions, giving additional information on eye health.

Last but not least, Nacho mentioned the development of a **multi-modal foundation model**. With the help of LLMs, the retinar team uses a concept of crawling through the internet and getting all sorts of retinal images with accompanied text to characterize the images. In the next step, the foundation model will be trained in a self-supervised setup and later on fine-tuned with a restricted dataset to better adapt to the target clinical scenarios.

After hearing so much about various data processing on the platform, an obviously important topic is patient privacy. Nacho explained that the private clinical report stays private until shared by the patient with a specialist. Furthermore, a patient can give consent for each image to be used for research purposes, enabling future investigations and improvements.



As a final remark, I want to mention quite a nice phrase from Nacho that we should all keep in mind - we are working on data of human beings with feelings, family, friends and life goals. Being able to increase their quality of life and making an impact on treatment outcomes is a great motivation for him and hopefully for all of us!



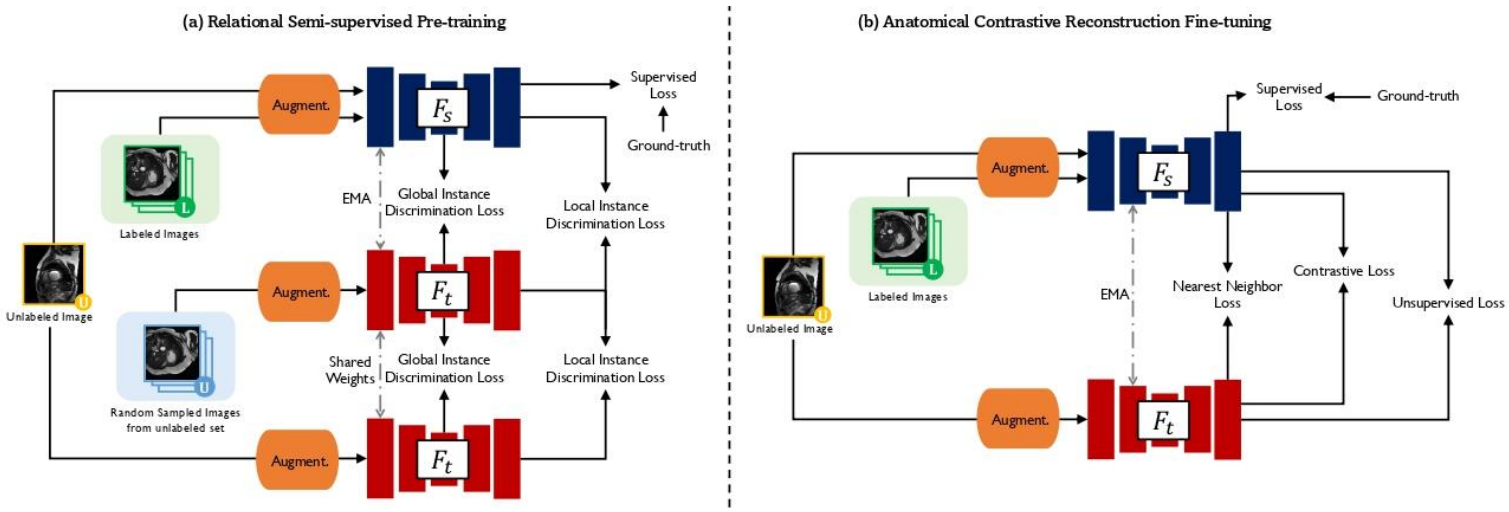
Mercedes Leguía, leading ophthalmologist of retinar, acquiring fundus pictures for the project at Hospital El Cruce (Argentina).

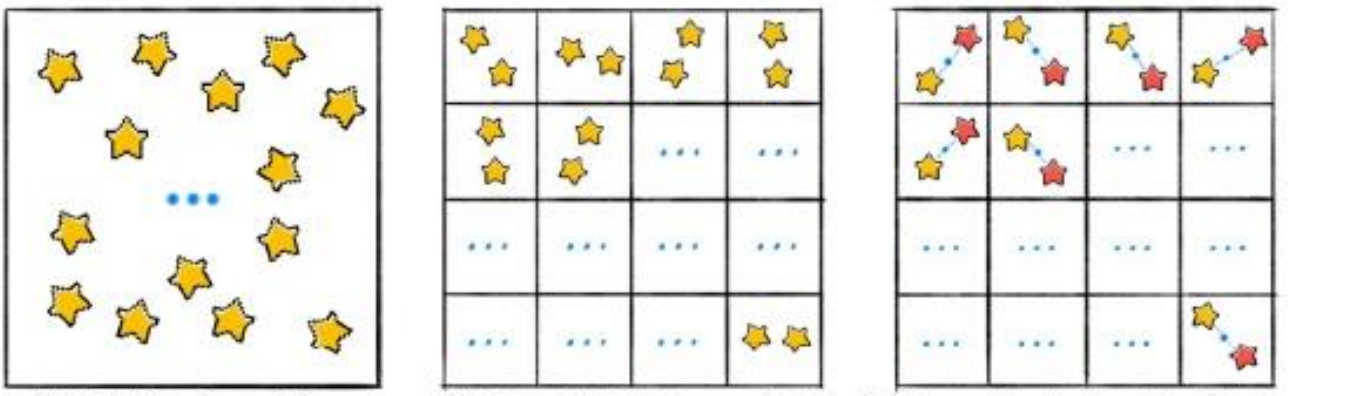
Rethinking Semi-Supervised Medical Image Segmentation: A Variance-Reduction Perspective

Chenyu You is a final year PhD candidate at Yale University working on medical image analysis and machine learning, supervised by James Duncan.

Chenyu's work has been accepted at NeurIPS2023.

This paper introduces ARCO, a semi-supervised contrastive learning (CL) framework with stratified group theory for medical and semantic segmentation.





(1) Naïve Sampling (NS) (2) Stratified Group (SG) Sampling (3) Stratified-Antithetic Group (SAG) Sampling

In this paper, Chenyu proposes and evaluates a novel approach to semi-supervised medical image segmentation with limited labels. Recognizing the shortcomings of existing machine learning models, which sometimes lack theoretical grounding in safety-critical scenarios such as medical imaging, this work provides a new theoretical perspective driven by the pursuit of enhancing robustness and efficiency.

"*Suppose we have a large amount of medical data, and most of the data is unlabeled,*" Chenyu poses. "*Contrastive self-supervised learning is the predominant approach to train models on a large amount of unlabeled data, but if we directly train models using a contrastive learning framework, they will suffer some model collapse issues.*"

Preliminary results revealed the model's tendency to misclassify minority labels, primarily due to class imbalance issues. In medical imaging, different organs account for different numbers of pixels, leading to this imbalance. Ultimately, the contrastive learning framework can go into model collapse, where it misclassifies different pixels from the high-frequency boundaries around the organs.

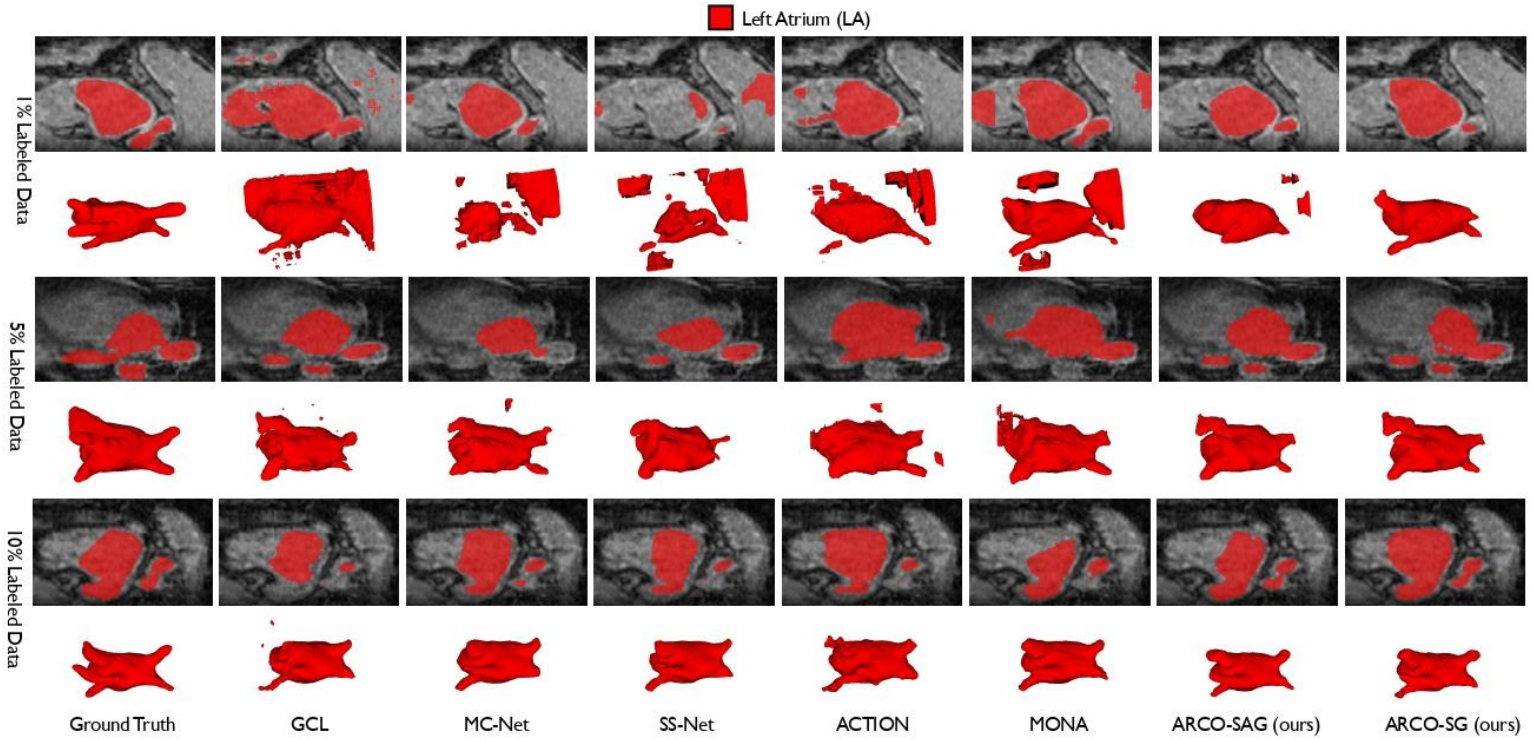
Chenyu's new sampling method aims to mitigate these issues and improve the reliability of the model. He reports positive outcomes, emphasizing two crucial medical image analysis and machine learning properties: segmentation robustness and label efficiency.

"*We evaluated our method from these two perspectives,*" he explains. "*For segmentation robustness, we improved our method compared to the state of the art by 5-10%, and 30% compared to the baseline. Another thing we achieved was an extremely limited label setting.*

We only require 1% of labels in the medical data to achieve results comparable to the supervised method. These are very good improvements, especially for medical image analysis."

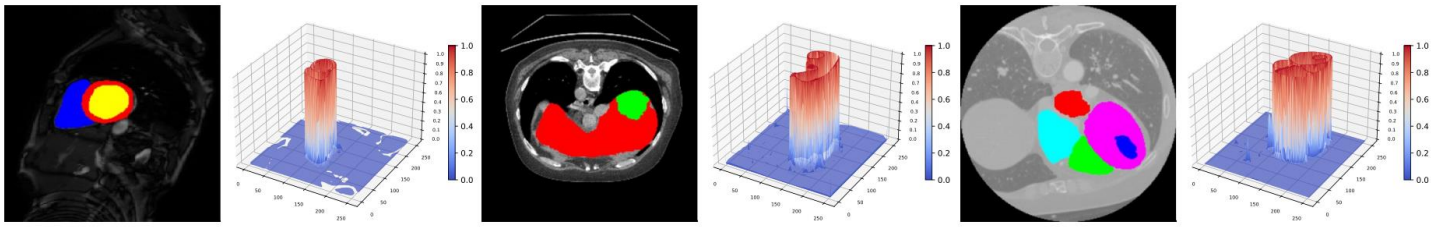
The impact of this work is not confined to medical image segmentation alone; Chenyu reveals that the method is extendable to the broader computer vision community. When tested on three semantic benchmarks, the algorithm consistently outperformed existing methods, achieving state-of-the-art results across the board.

The key to these advancements lies in the **MONA framework**, a contrastive learning framework proposed by Chenyu last year and built on the principles of tailness, diversity, and equivalence.

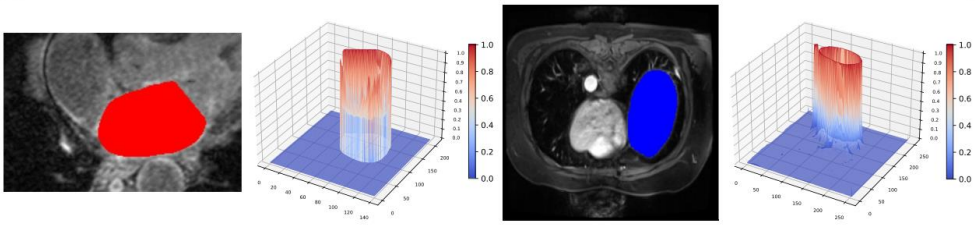


"*We tried to address medical segmentation based on these three principles but found that if we randomly or naively sampled the pixels, it resulted in model collapse issues,*" he reveals. "*We introduced two new sampling strategies to address this. Also, we provided a theoretical guarantee to demonstrate that these sampling strategies can effectively reduce the variance issues and mitigate model collapse.*"

MONA has already been extended to other computer vision domains, showcasing excellent performance. It initially improved the label setting from 5% to 1% and demonstrated a remarkable 10% performance increase across four datasets in both



(1) ACDC (2) LiTS (3) MMWHS



(4) LA (5) MP-MRI

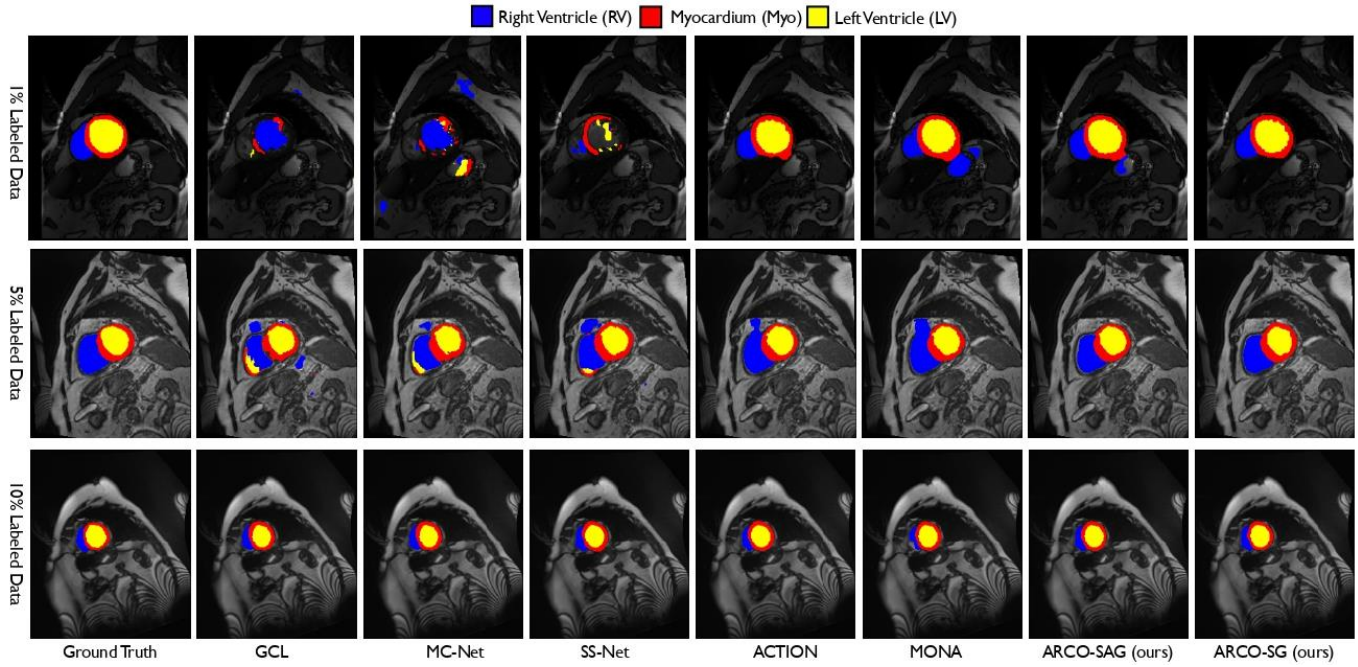
2D and 3D domains.

"These three principles will guide future community researchers to address their medical analysis problems," he adds. *"My colleagues doing registration incorporate this framework into their registration tasks, and those doing reconstruction incorporate it into their reconstruction tasks."*

As we conclude our interview, Chenyu tells us how much he has enjoyed talking with us and thanks us for the invite. Finally, he has some words of gratitude for his colleagues.

"I'd like to thank my advisor, Professor James Duncan, for his guidance, and my collaborator Lawrence, who is also in our Image Processing and Analysis Group at Yale University," he remarks. *"Thank you, guys, for all the help!"*

Do you want to learn more about Chenyu's work? Visit him during the NeurIPS poster session 5 on Thu 14 Dec starting at 17:45 CST.



Jianan Chen has recently completed his PhD at the University of Toronto, under the supervision of Anne Martel.

His research focused on improving outcome prediction for cancer patients using machine learning and liver MRI.

Jianan has recently accepted a postdoc position at University College London, where he will continue to study cancer, this time using the integration of genomics and digital pathology. Congrats, Doctor Jianan!

Medical image-based outcome prediction can provide information for the risk stratification and precision medicine of cancer patients. Machine learning plays an important role in the analysis of medical images and have demonstrated success in many outcome prediction tasks. However, machine learning-based outcome prediction still faces many challenges, limiting its application in the clinic.

In his PhD thesis, Jianan discussed several major challenges in the clinical translation of outcome prediction algorithms and developed machine learning techniques to address them.

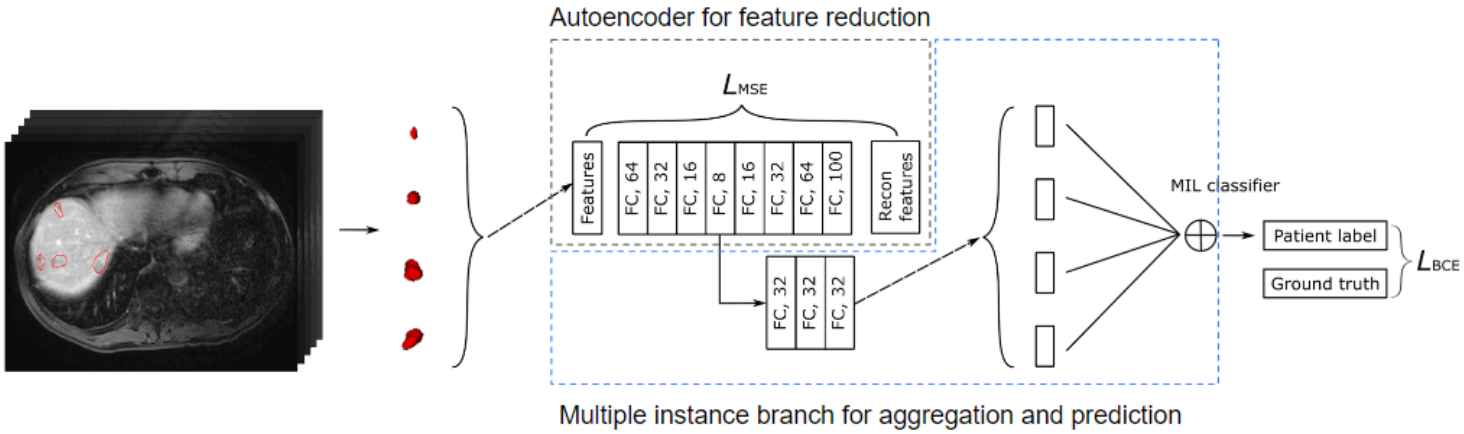
Jianan first tackled the problem of low generalizability of radiomics models caused by inter-institution variabilities. The differences in study population, imaging equipment and imaging protocols can lead to large variations in the enhancement patterns of contrast-enhanced MRIs. Jianan proposed a feature selection strategy informed by the mechanism of contrast agents to select contrast-agnostic features. The model trained with the selected features were successfully validated in multiple cohorts from multiple hospitals. This study showed the possibility of training a widely generalizable radiomics biomarker by keeping in mind the issue of overfitting and

utilizing prior knowledge.

Next, to further reduce the possibility of overfitting, Jianan presented an unsupervised clustering algorithm for identifying imaging phenotypes of colorectal cancer lesions. Inspired by the observation that imaging phenotypes do not necessarily correlate with molecular subtypes, he hypothesized that different subtypes of tumors may have similar appearances. With an autoencoder-based Gaussian mixture model, it became possible to address the overlapping appearance problem while performing unsupervised clustering.

Lastly, Jianan proposed an outcome prediction algorithm specifically designed for multifocal and metastatic cancer. It remains unclear how multiple tumors contribute to patient outcome, and most medical imaging-based prognostic models focus solely on the primary lesion. Jianan proposed a multiple-instance neural network for integrating information from all tumor lesions. His hypothesis that using all tumor lesions for outcome prediction improves prediction accuracy was empirically validated. This approach also enabled the analysis of individual lesion aggressiveness without labels, with implications for understanding disease mechanism and clinical decision making.

Predictions based on the approaches proposed in Jianan's thesis achieved better prognostic value when compared to existing clinical and imaging biomarkers. In the future, Jianan plans to validate his findings in larger datasets and different types of cancers. He will also explore the integration of other data modalities to develop biologically-relevant imaging biomarkers.



Breamy: An augmented reality mHealth prototype for surgical decision-making in breast cancer

Niki Najafi is a graduate student in Computer Science at Concordia University supervised by Marta Kersten-Oertel.

She recently presented an augmented reality mobile app, Breamy, for more informed surgical decision-making in breast cancer at the AE-CAI workshop (MICCAI 2023).

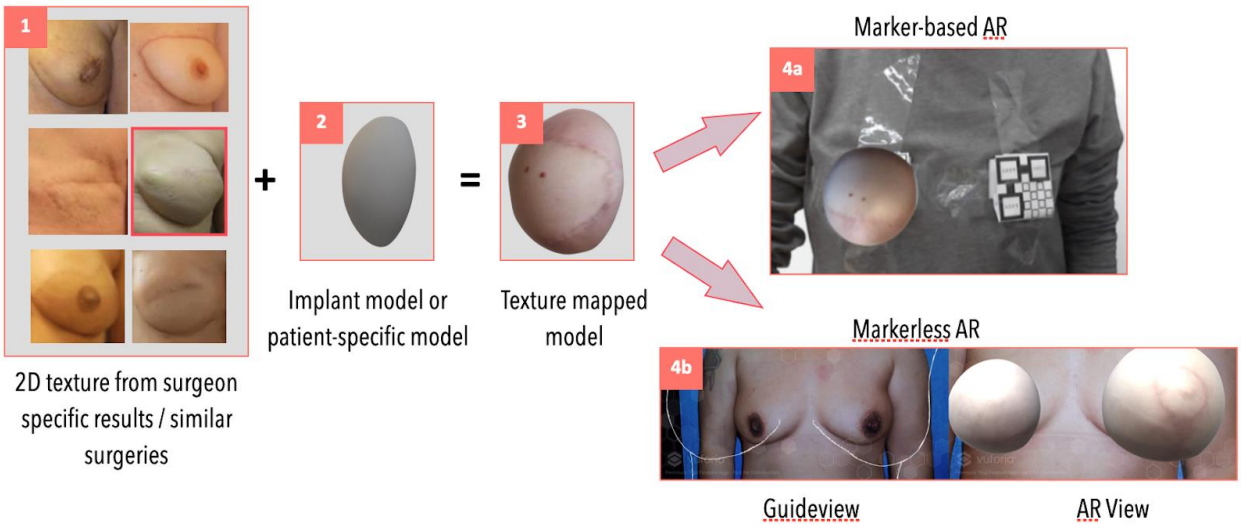
The idea of Breamy emerged from a team of graduate students (Miranda Addie, Chelsea De Bellis, Ujjwala Naithani, Niki Najafi, and Tanushree Paul) taking part in the NSERC CREATE Surgical Innovation program (delivered jointly by McGill, Concordia and École de technologie supérieure) with clinical guidance from Sarkis Meterissian, MD, Director of the Breast Center at McGill University Health Centre (MUHC).

Did you know that **1 in 8 women** will be diagnosed with breast cancer? Most diagnosed women will require surgery as part of their treatment. If women, with their surgeons, decide to proceed with a mastectomy, i.e., surgery to remove the breast, they must decide whether they desire breast reconstruction. However, the time from diagnosis to surgery is short – between 4-8 weeks. Coupled with the stress of a cancer diagnosis, these patients often feel overwhelmed by the life-altering decisions they have to quickly make. This type of time-sensitive decision-making process can lead to anxiety, decisional regret, and revision surgeries.

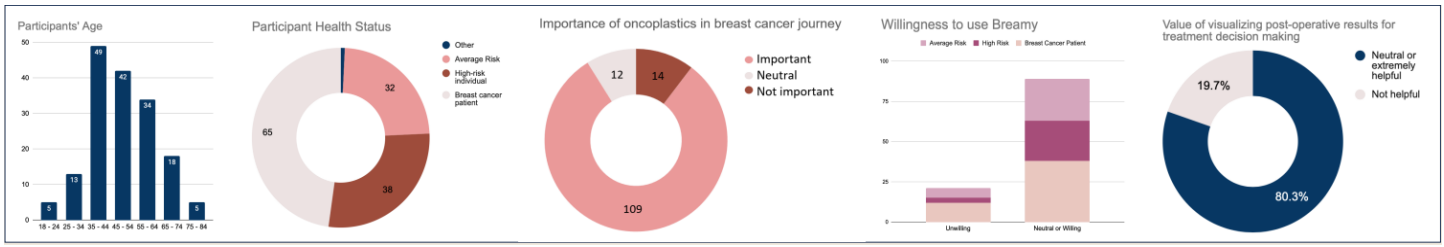
Decision aids (DAs) help patients by providing detailed medical information, clarifying values, and engaging them in active decision-making. The use of DAs has been shown to aid collaborative and informed decision-making. Our goal is to develop smartphone technology to empower women to make more informed and personalized decisions. Our prototype, Breamy, uses AR to display various reconstructive options on a patient's body for more tailored decisions.

In Breamy's AR feature, the surgeon selects photographs that are most likely to represent the patient's surgical outcome. These are mapped onto a 3D model, allowing patients and surgeons to see treatment options in real-time. We developed 2 ways of using AR:

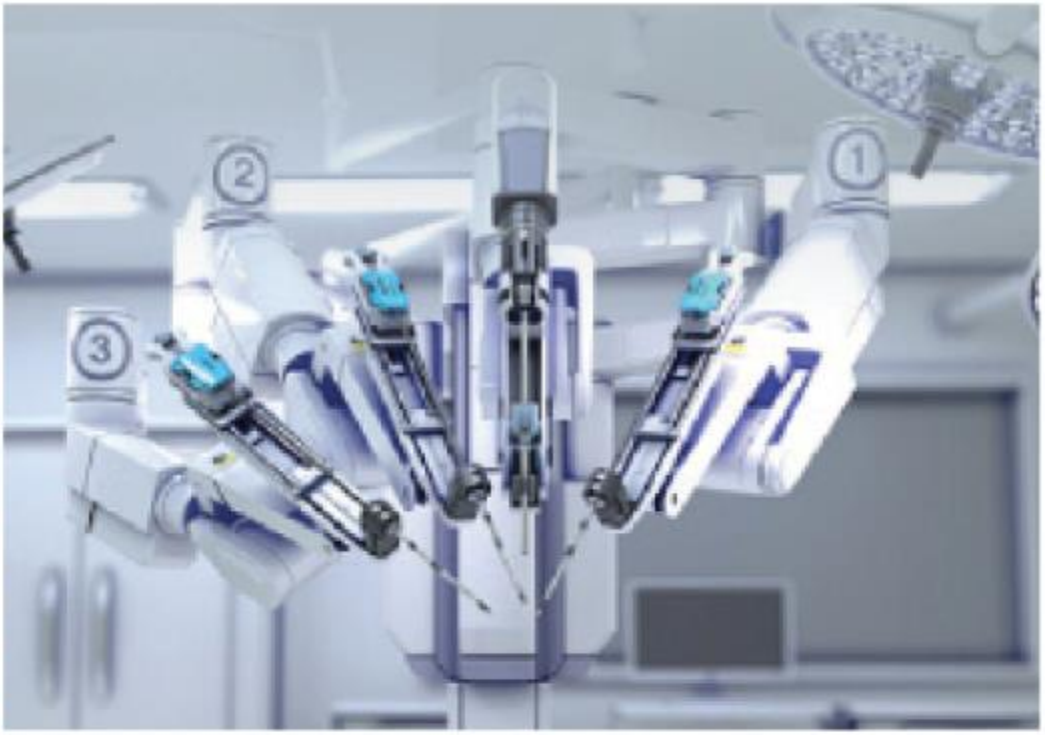
Marker-based method where the 3D breast model is projected upon detection of an AR marker (a QR code). We envision this could be used when the patient is wearing clothing for example in the surgeon's office. Second, we have a **markerless method** where we detect the patient's breasts directly by using a "guideview" (an outline of the patient's breast 3D model). The patient uses the camera of their phone and adjusts their device until the guideline fits their breasts. This initiates body tracking and enables visualization of different projected surgical models on them. Watch here.



To assess the perceived usefulness of Breamy, we surveyed 166 people including breast cancer patients, high risk individuals, and average risk women. Around 80% of the respondents expressed their willingness to use Breamy to view different surgical treatment options projected onto their bodies if they had to make a decision about their own breast surgery treatment.



These initial findings show that using AR for breast cancer patients can improve understanding and assist in making better decisions. This may result in fewer revision surgeries, less regret, and improved overall patient well-being. Breamy aims to enhance patient education and help survivors feel confident about their bodies.



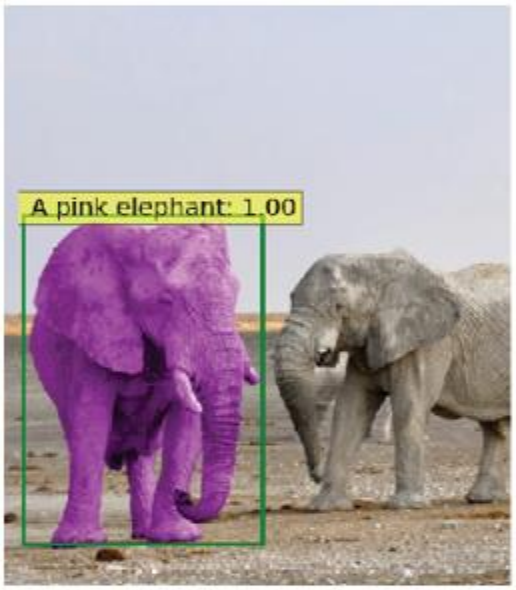
IMPROVE YOUR VISION WITH

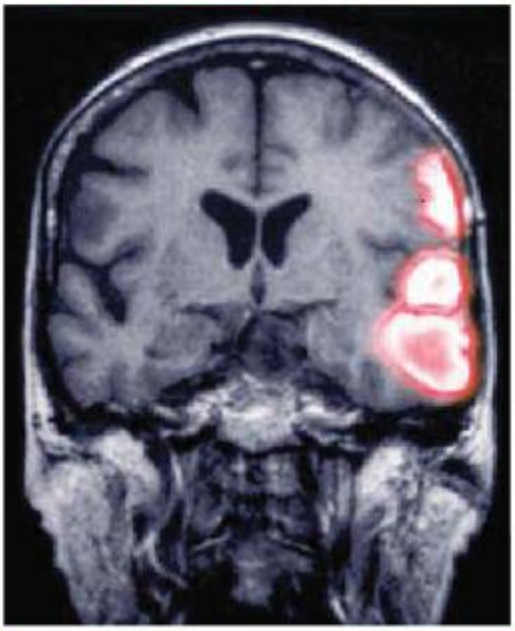
Computer Vision News

SUBSCRIBE

to the magazine of the algorithm community

and get also the new supplement Medical Imaging News!





RSIP VISION