July 2024

Computer Vision News

& Medical Imaging News

The Magazine of the Algorithm Community

BEST OF CVPR

Generative Image Dynamics

Full review of Best Paper Award and many more - Read about brilliant work!

A publication by RSIP Vision

CVPR Best Paper!

Generative Image Dynamics

Zhengqi Li is a research scientist at Google DeepMind, working on computer vision, computer graphics, and AI.

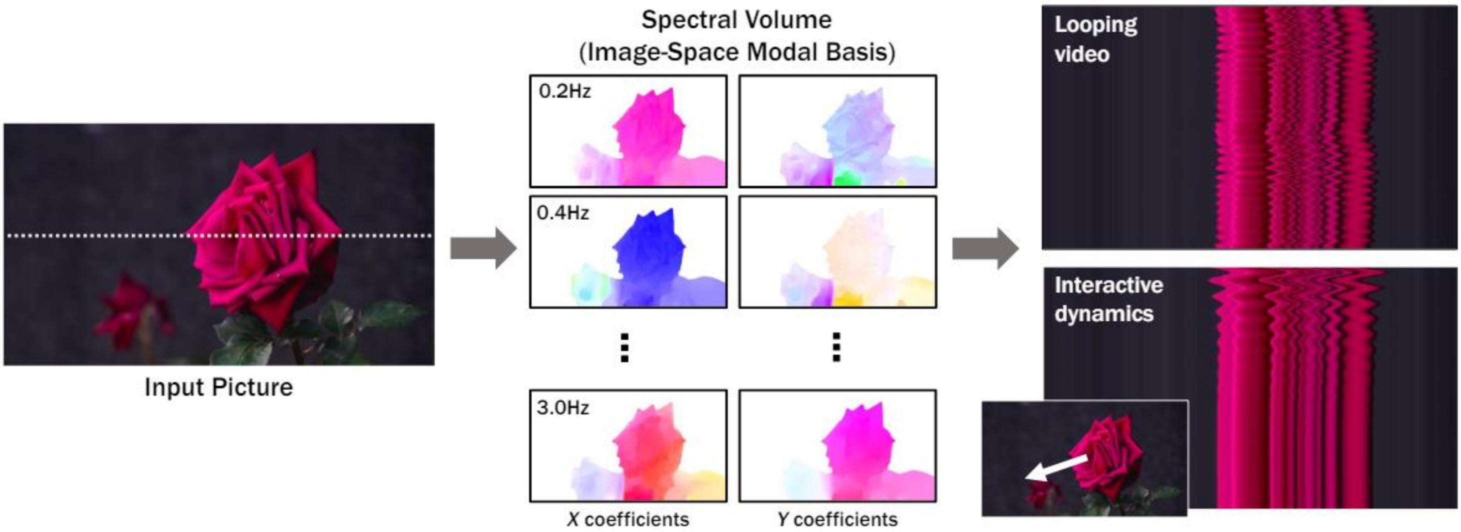
His paper on Generative Image Dynamics has not only been selected as a highlight paper this year but is also in the running for a best paper award.

He is here to tell us more about it before his oral and poster presentations.

NOTE: this article was written before the announcement of the award winners. Which explains why it keeps mentioning a candidate and not a winning paper. Once again, we placed our bets on the right horse! Congratulations to Zhengqi and team for the brilliant win! And to the other winning paper too!

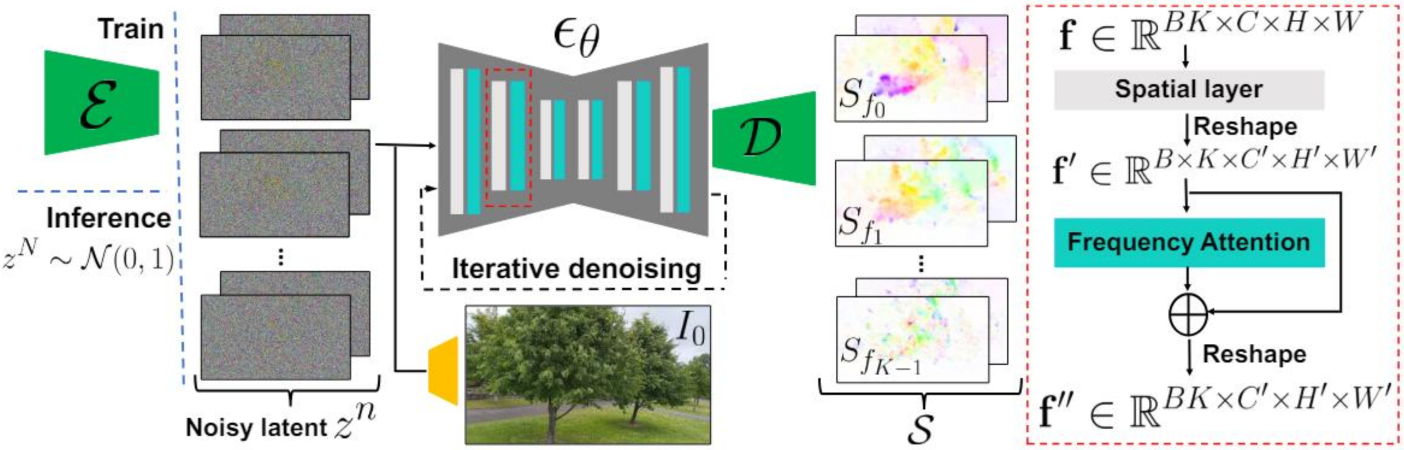
Imagine looking at a picture of a beautiful rose and visualizing how it sways in the wind or responds to your touch. This innovative work aims to do just that by **automatically animating single images without user annotations**. It proposes to solve the problem by modeling what it calls image-space motion priors to generate a video in a highly efficient and consistent manner. *"By using these representations, we're able to simulate the dynamics of the underlying thing, like flowers, trees, clothing, or candles moving in the wind,"* Zhengqi tells us. *"Then, we can do real-time interactive simulation. You can use your mouse to drag the flower, and it will respond automatically based on the physics of our world."*

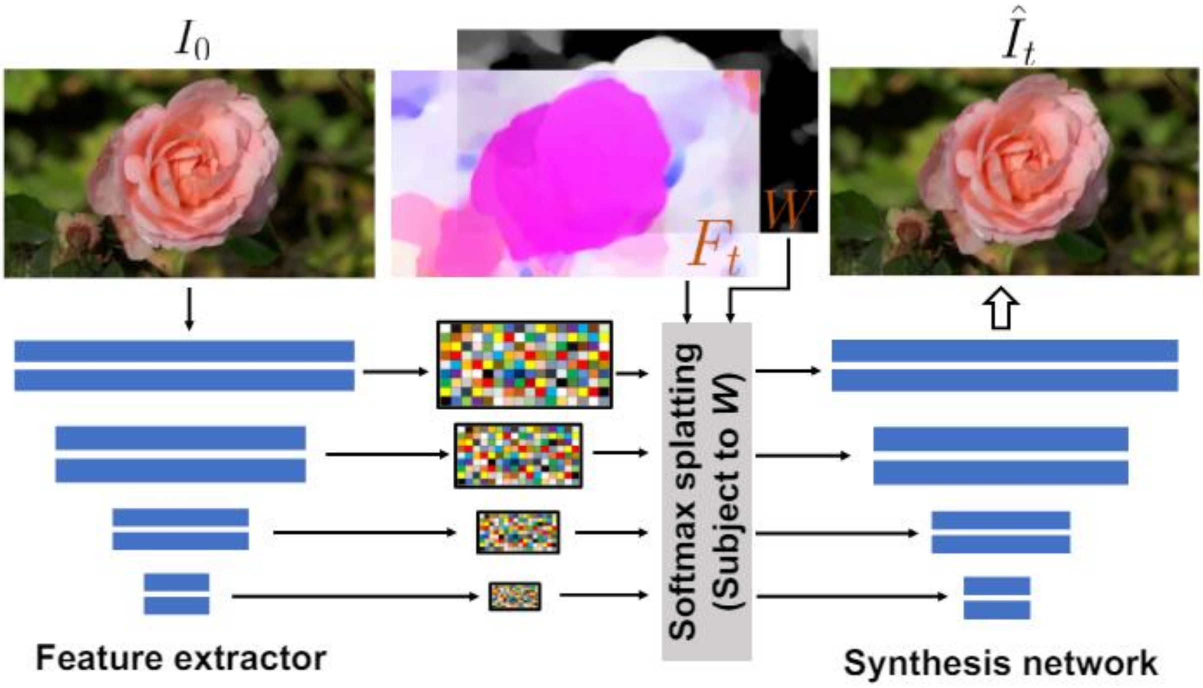
The applications of this technology are already promising. Currently, it can model small motions, similar to a technique called **cinemagraph**, where the background is typically static, but the object is moving. A potential application for this would



be **dynamic backgrounds for virtual meetings**, providing a more engaging and visually appealing alternative to static or blurred backgrounds but without excessive motion that could be distracting. *"Moving to model larger motion, like human motion or cats and dogs running away, is an interesting future research direction,"* Zhengqi points out. *"We're working on that to see if we can use a better and more flexible motion representation to model those generic motions to get better video generation or simulation results."*

Most current and prior mainstream approaches in video modeling involve using a deep neural network or diffusion model to directly predict large volumes of pixels representing video frames, which is computationally intensive and expensive. In contrast, **this work predicts underlying motion, which lies on a lower-dimensional manifold, and uses a small number of bases to represent a very long motion trajectory**. *"You can use a very small number of coefficients to represent very long videos,"* Zhengqi explains. *"This allows us to use this*


Train
$\mathcal{E}$
Inference
$z^N \sim \mathcal{N}(0,1)$
Noisy latent z^n
ϵ_θ
Iterative denoising
I_0
$\mathcal{D}$
S_{f_0}
S_{f_1}
$S_{f_{K-1}}$
$\mathcal{S}$
$\mathbf{f} \in \mathbb{R}^{BK \times C \times H \times W}$
Spatial layer
Reshape
$\mathbf{f}' \in \mathbb{R}^{B \times K \times C' \times H' \times W'}$
Frequency Attention
Reshape
$\mathbf{f}'' \in \mathbb{R}^{BK \times C' \times H' \times W'}$



representation to produce a more consistent result more efficiently. I think that's the main difference compared with other video generation methods you might see."

The novelty of this approach has not gone unnoticed, with the work being picked as a top-rated paper at this year's CVPR, given a coveted oral presentation slot, and recognized as one of only 24 papers in line for a best paper award. If we were placing bets on the winners, this work, with its stellar team of authors, would be our hot tip. What does Zhengqi believe are the magic ingredients that have afforded it such honors? *"There are a few thousand papers on video generation dynamics, and they all have similar ideas,"* he responds. *"They predict the raw pixel, and* ***we're going in a completely different direction predicting the underlying motion****. That's something the research community appreciates because it's unique. I guess they believe this might be an interesting future research direction for people to explore because, for generative AI, people are more focused on how you can scale those big models trained on 10 billion data while* ***we're trying to use a different representation that we can train more efficiently to get even better results****. That's a completely different angle, and the award community might like those very different, unique, special angles."*

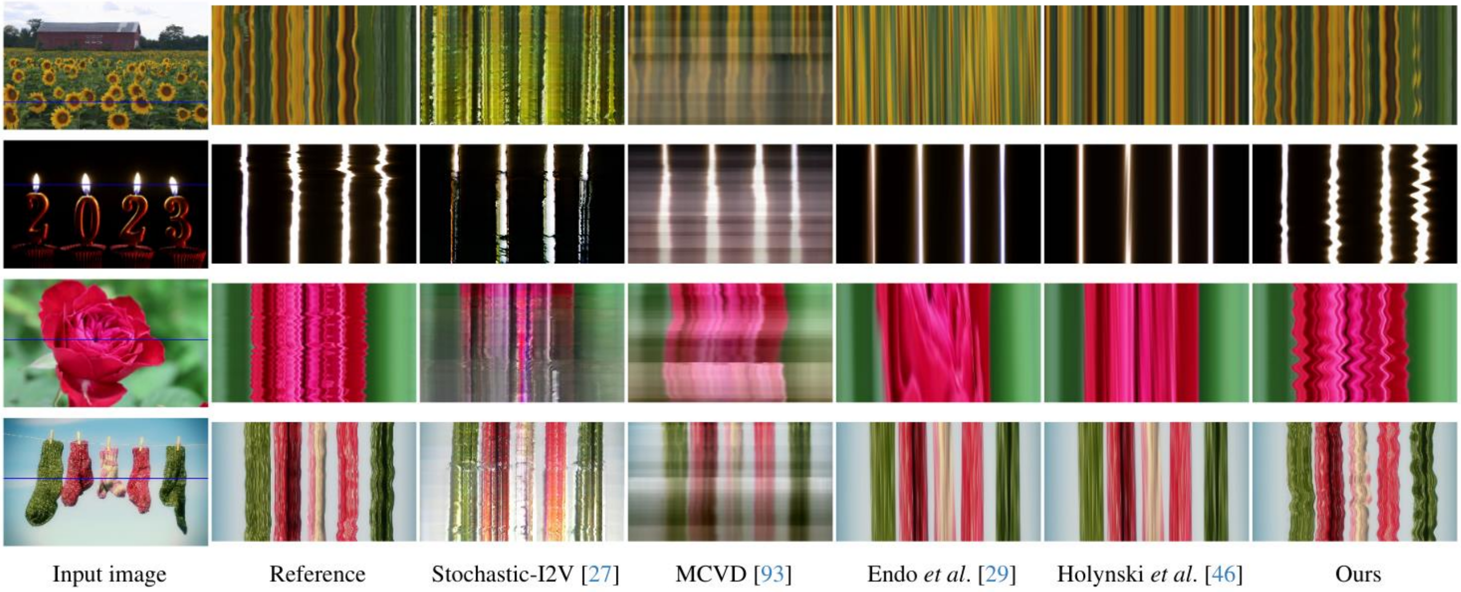
"If you don't have data, you can't train your model to get good results!"

However, the road to this point was not without its challenges. Collecting sufficient data to train the model was a significant hurdle the team had to overcome. They searched the Internet and internal Google video resources and even captured their own footage to gather the necessary data, taking a camera and tripod to different parks to capture thousands of videos. “*The hardest part was we spent a lot of time working on it, but that’s the key ingredient that made our method work,*” Zhengqi recalls. “*If you don’t have data, you can’t train your model to get good results.*”

While other works use optical flow to predict the motion of each pixel, **this work trains a latent diffusion model, which learns to iteratively denoise features starting from Gaussian noise to predict motion maps rather than traditional RGB images**. Motion maps are more like coefficients of motion. The model uses this to render the video from the input picture, which is very different from other works that directly predict the video frame from the images or text. “*That’s something quite interesting,*” Zhengqi notes. “*We’re working from more of a vision than a machine learning perspective. I think that’s why people like it in computer vision communities.*”

Outside of writing award-candidate papers, Zhengqi’s work at Google mainly focuses on research but has some practical applications, including assisting product teams with video processing. He also advises several PhD student interns. “*We work together on interesting research projects to achieve very good outcomes,*” he reveals. “***That’s our daily goal as research scientists at Google DeepMind!***”

CVPR Best Paper!



Objects as Volumes: A Stochastic Geometry View of Opaque Solids

Bailey Miller is a PhD student at Carnegie Mellon.

His novel paper, which breaks new ground in volume rendering, has been selected from thousands of accepted papers as a conference highlight and as a candidate for a coveted Best Paper Award.

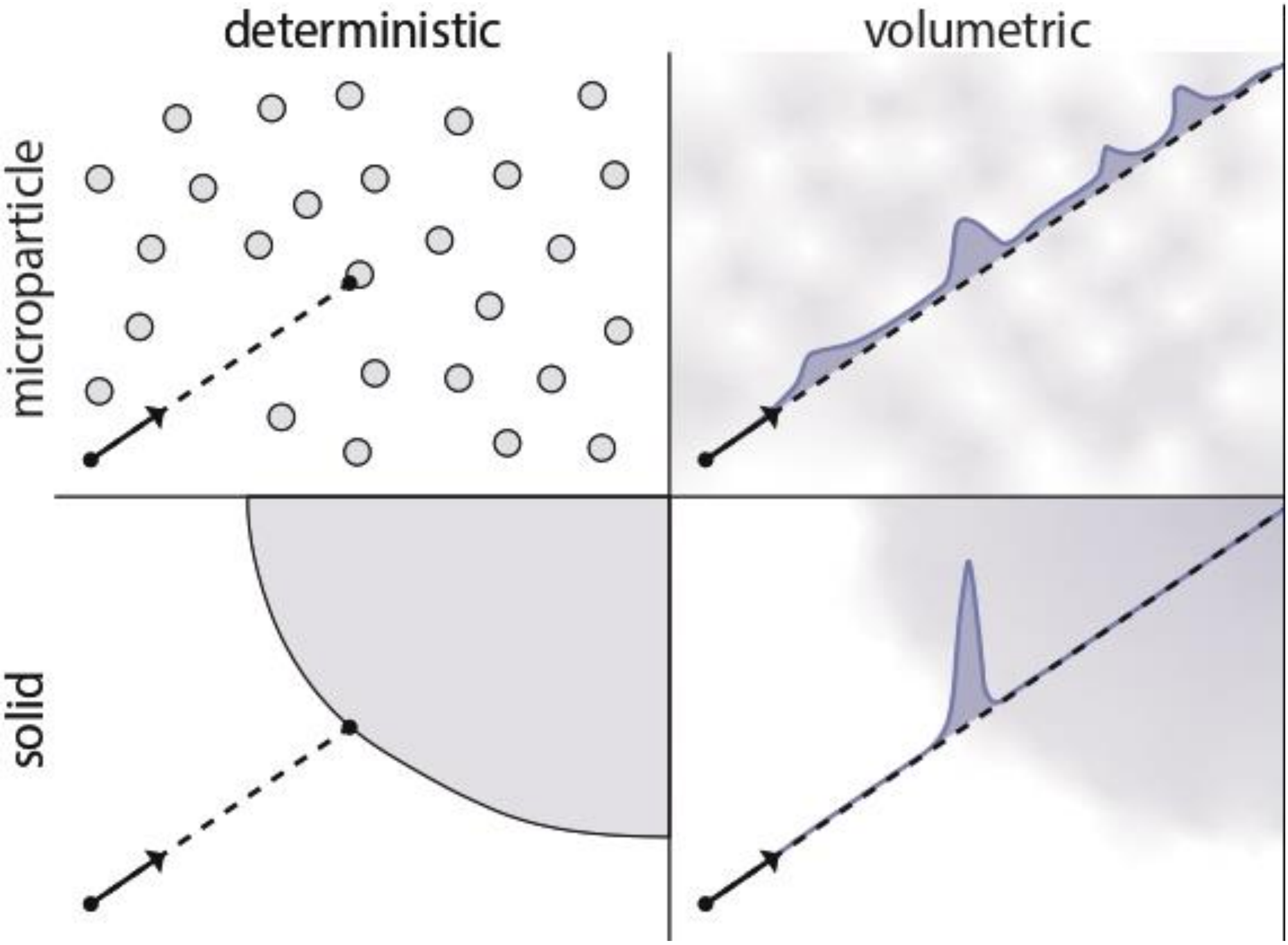
Bailey spoke to us before his oral presentation at CVPR 2024.

For several decades, **volume rendering techniques** have been a popular class of methods for simulating light transport in translucent media such as clouds, smoke, and tissue, with various applications in graphics and physics. In the past five years, there has been a shift towards using these methods to model more familiar, everyday objects such as solid, opaque items.

"Our paper is about figuring out why these volume rendering methods, originally developed for clouds, can work on things like a Lego truck," Bailey begins. *"**We've developed a stochastic geometric theory that explains the connection between these two different models.**"*

The initial challenge Bailey faced was understanding the foundational principles of volume rendering, which have been obscured or had a black box put around them by the numerous successful yet complex methods developed in recent years. *"Revisiting its roots, you see that in classic volume rendering, scenes are modeled as a collection of microparticles,"* he reveals. *"Once we could understand it in this very principled manner, we could start to develop **ideas and approaches for considering volume rendering on stochastic opaque solid objects.**"*

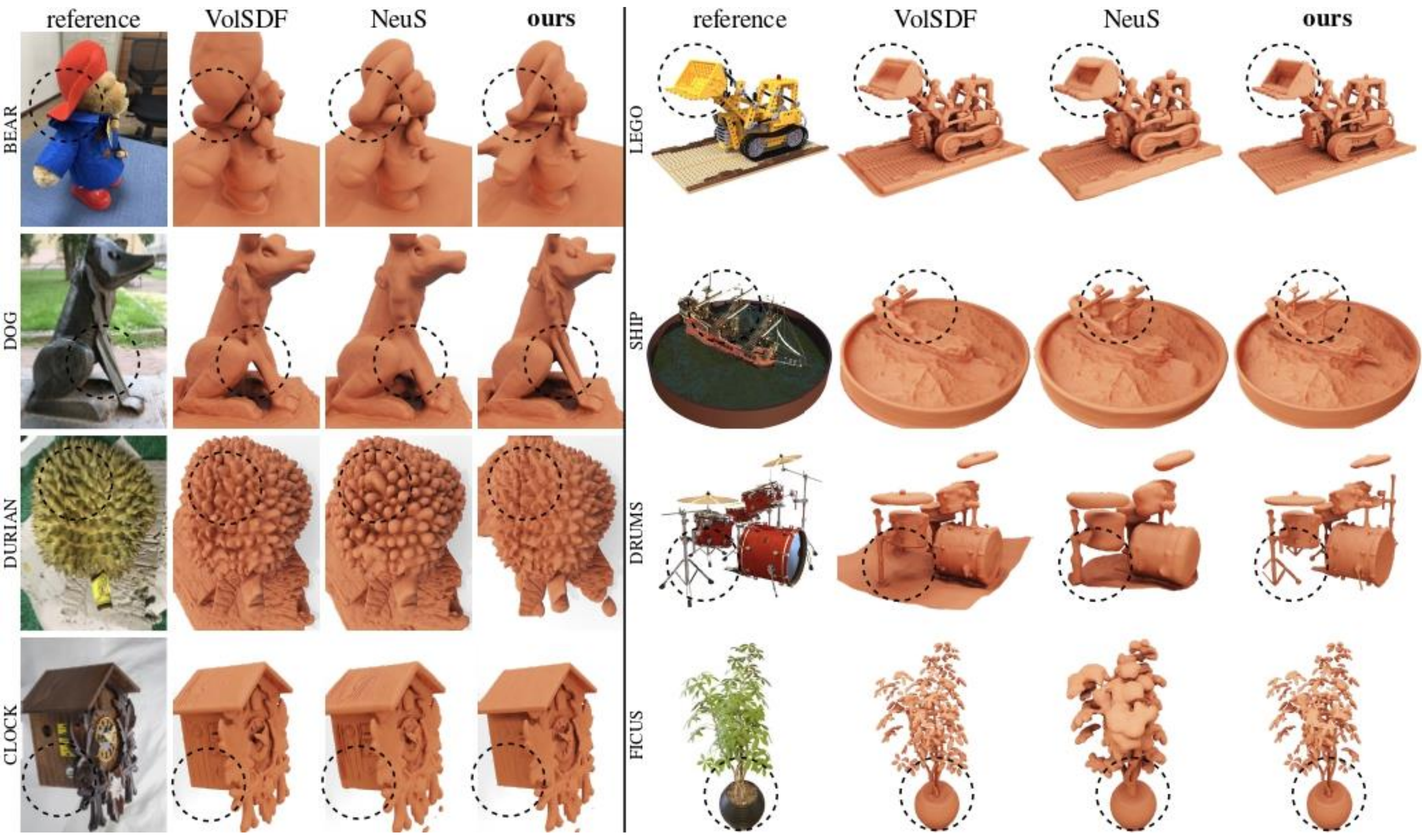
Did he solve the problem? *"Part of it,"* he tells us. *"I think we've opened some new doors. We show how you can develop these rendering algorithms for a very*



limited set of new stochastic geometry. There's a lot of work to be done in extending these methods to even more extensive types of geometry and scenes."

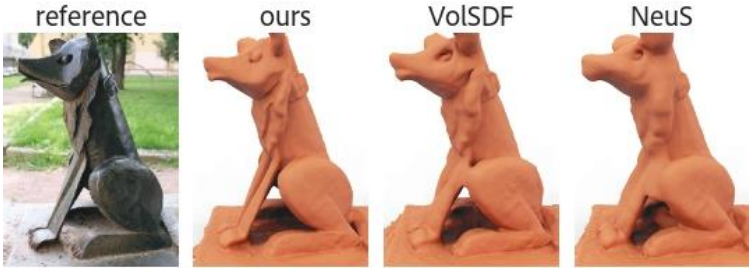
One of the primary applications of this work is in **surface reconstruction**. Essentially, this involves taking a collection of images and trying to understand the geometry of the world that gave rise to them. The connection becomes clearer through light transport. *"Light bounces around the world, and depending on how it interacts with the geometry, it gives different images,"* Bailey explains. *"By introducing a new way of modeling the geometry in scenes and considering how light interacts with that geometry, we can improve the surface reconstruction algorithms that have been using volume rendering over the past several years* ***to get all sorts of performance improvements****."*

The implications of this research extend to various fields, particularly settings such as **robotics** or **autonomous driving**, where there is a benefit in having a notion of uncertainty in solid objects and the



trustworthiness of algorithmic results. **In these scenarios, agents might leverage images or video to build a probabilistic model of the surrounding world, which can significantly enhance reliability, safety, and efficiency**.

Being chosen as an **award candidate for the first time** is a high enough achievement, but it is even more special for Bailey, as **this is his first CVPR**. He puts the positive reception down to how timely the work is given recent advancements in surface reconstruction and novel view synthesis algorithms based on volume rendering, which have worked well but for unclear reasons. “*By providing this perspective and saying, ‘Here’s an explanation for why all these volume rendering methods work so well,’ we can understand why we can model the world in this particular way and why we’ve had success with it,*” he points out. “*We can understand the current state of the art but also develop a perspective on these methods that allows us to continue improving. I love thinking about how we model the world and how the underlying assumptions we make in our models impact the algorithms and methods we ultimately develop.*”

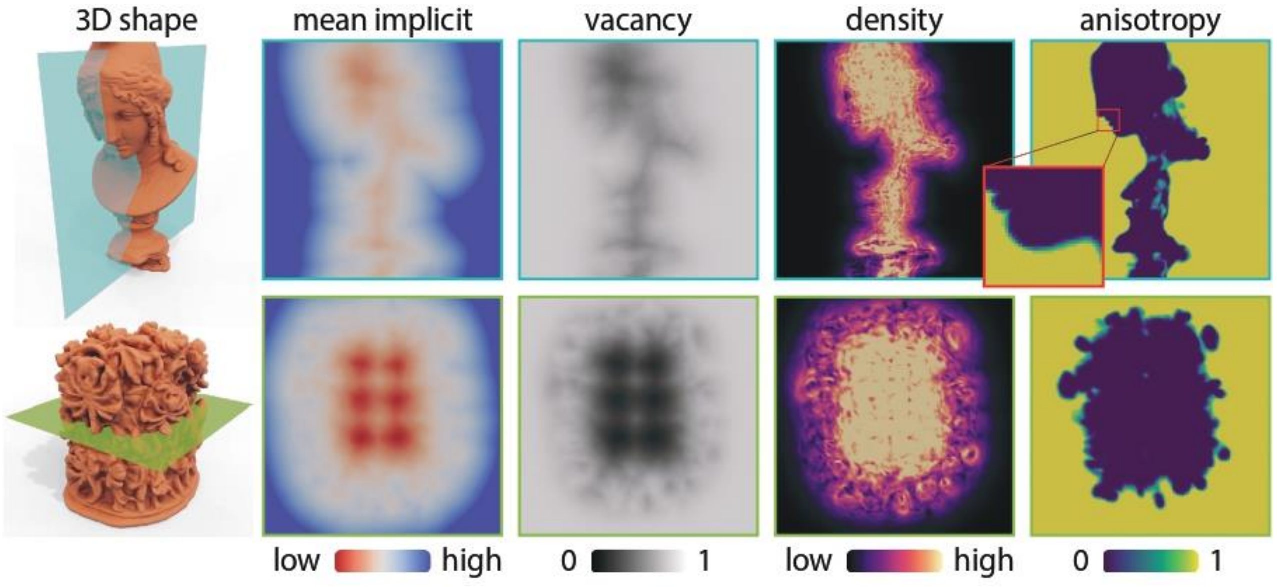


Reflecting on the growing emphasis on the role of **explainability in research**, Bailey says it is a trade-off: *"You need to push forward and figure out what works in practice, and then you need to step back and ask, 'Why do these methods work so well?' It's a constant process moving back and forth between the two."*

In addition to his work on stochastic geometry, Bailey is exploring **Monte Carlo PDE solving**, which involves adapting Monte Carlo methods that work really well for light transport and simulating light to other types of physics, such as **heat transfer, acoustics, and wave equations**. *"I don't think this has been as present in the vision community yet, but it's been starting to gain some attention in graphics,"* he tells us. *"I think, eventually, these algorithms will be of interest in the computer vision community because we're seeing the development of all sorts of new imaging modalities or renewed interest in modalities like* ***thermal imaging****. Good ways to simulate those should help vision researchers and practitioners* ***develop algorithms that work with physics beyond just light****."*

Looking ahead, **Bailey is excited about the potential for developing this work further, including extending the stochastic geometric approach to a broader range of stochastic models and probabilistic assumptions about the world or scenes**. Also, the core idea of stochastic geometry has applications beyond light transport algorithms, which opens a range of possibilities for future research.

Could we be sensing the first hints of next year's award paper? *"I'd love that, but I'm happy with the one this year for now!"* he laughs. *"I feel very fortunate to have had our paper selected. I hope everyone who reads it enjoys it and takes something away from this stochastic geometry perspective."*



Paula Ramos (on my left) and Raymond Lo (on my right) are AI Evangelists at Intel.

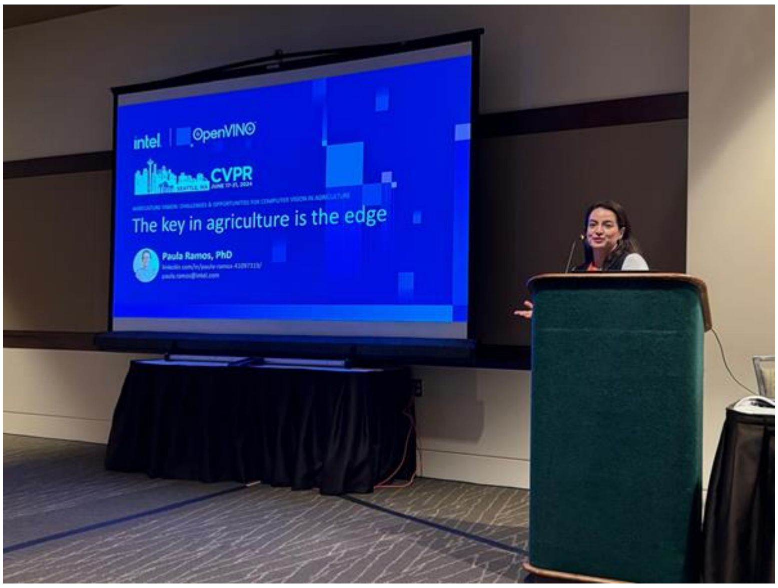
We catch up with them at CVPR 2024 to learn more about Intel's innovative AI-powered solutions, their authoritative engagement at the conference, and the community's enthusiastic response.

Raymond and Paula's CVPR journey began three years ago with a vision to bridge the gap between academia and industry regarding AI. *"It was a dream that Paula and I had,"* Raymond, the global lead of the **Intel AI Evangelist team**, recalls. *"When I was a researcher and had an idea, people always said, '10 years later, it will come to the market.' It was discouraging to me. I was like, no, I want this to be in the market tonight!"*

That vision has materialized with Intel making significant strides in AI, notably **through OpenVINO, its open-source toolkit for AI inference**, allowing developers to optimize and deploy models efficiently across various platforms. These days, AI innovations can progress from research into users' hands within a year. *"The new AI trends are moving super fast,"* Paula tells us. *"If we have a new model, we want to deploy it at the edge, in the cloud, or* *on client devices. We're trying to bridge the gap between those new AI trends and how developers can deploy on their own laptops so they don't need to create or use new infrastructure."*

Intel Labs researchers presented 24 papers at CVPR this year, including six in the main conference, and co-organized **three workshops**. Intel had a booth and a tutorial. AI Research Engineer Samet Akcay was a keynote speaker at the **Anomaly Detection workshop**, while Paula was a keynote speaker at the **Agriculture-Vision workshop**. Intel also sponsored the **AI Summit**, a networking meetup, and the **Visual Anomaly and Novelty Detection 2024 Challenge**.

"When we started at CVPR, the demos were rough because there were only a few of us trying to make it work," Raymond remembers. *"A success story I can tell is that the community came together with hundreds of people working on the same thing. Hundreds. That's the difference we see.* ***People appreciate that we come back and are better at repeating the effort!***"



This community spirit is reflected at the booth and the annual party Intel hosts at CVPR, which has proved to be a popular way to connect people. Creating a fun and engaging environment ensures that people remember the company from a human perspective, not just as a technology provider. *"My perspective is that* ***I'm here to show you the best of everything we have****, but we can't forget that we're walking in the footsteps of all these giants here at CVPR, all these talents,"* Raymond recognizes. *"The whole purpose of the party is to put the right people together at the right time. Of course, it has to be fun!"*

"OpenVINO Notebooks repository has literally hundreds of tutorials!"

He points out that Intel's commitment to the community goes beyond superficial engagement. It brings hardware and software engineers to CVPR for hands-on support and expertise. *"We care by showing up,"* he attests. *"In our booth, we have the people that do the actual development. Two engineers with PhDs in Optimization will show you how we optimize the hardware.* ***It's not just telling you to read the documentation; the engineers are here to talk to you if you want extra help. We're here for you!****"*

Raymond says they regularly add people to LinkedIn, give them demos, and this year, even **gave away Discrete Arc GPUs, AI PC DEV KITs, and AI PCs featuring the Intel Core Ultra, which combines NPU, CPU, and GPU capabilities**. *"It's a very powerful machine in the sense that, 10 years ago, if I saw it, I'd be like, 'Oh my God!'"* he exclaims. *"For students, this is actually their laptop in the coming years. We're shipping hundreds of millions of these."* He stresses to students the importance of making their innovative solutions accessible and repeatable. *"Great, you have your solution, but have you thought about giving it to your friends and your friends' friends?"* he asks. *"Think about the layer of people that can benefit from your work. That's why I say these kinds of laptops are not just laptops; they're tools for expanding your ideas."*

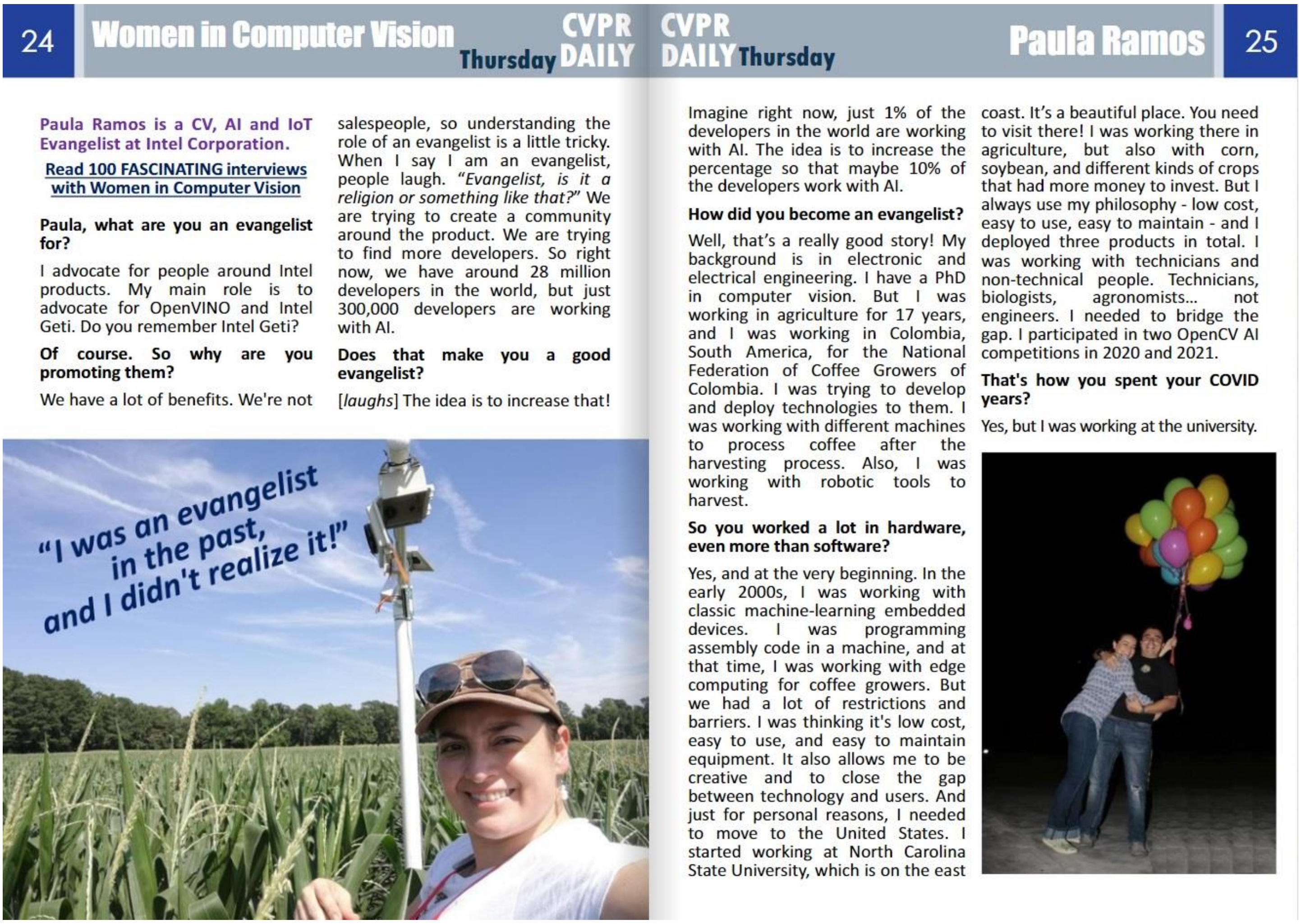
With tools like OpenVINO, people can run high-performance AI models on modest hardware, making advanced capabilities available without significant financial investment. “*Today, people came to my booth to say, ‘Hey, Ray, what are you running?’*” he tells us. “*I’m running a large language model, Llama 3, eight billion parameters, that never runs properly on their machines. I use OpenVINO to compress it, and now it fits into the RAM, and I can fit it into the GPU of this little machine.*” That is how the 10-year route-to-market cycle that Raymond talked about earlier gets shorter. “*Your everyday machine becomes your AI engine,*” he confirms.

> ***“It really warms my heart when people come back to tell me I tried it, I used it, I made an impact...”***

Intel is developing user-friendly APIs and extensive tutorials to help computer vision professionals benefit from these advancements. **Its OpenVINO Notebooks repository has literally hundreds of tutorials!** It has Edge AI Reference Kits designed for specific use cases across industries like manufacturing and retail. It creates content users can follow step by step to optimize, quantize, and apply these new AI trends in their infrastructure. “*We don’t just give you cookies,*” Raymond adds. “*You get the recipes for making your own cookies!*”

Feedback from the field has been overwhelmingly positive. “***It really warms my heart when people come back to tell me I tried it, I used it, I made an impact,***” he recounts. “*One person told me, ‘I can make a business from it. I can make an impact on millions of people using this product.’ He became a top scientist from Stanford and told me, ‘I could accelerate some of this model as well.’ That’s heartwarming because we don’t just show people something, and then they forget, and it never gets used anywhere else; they take it to the next level!*”

Intel’s commitment to the community extends beyond CVPR, with regular virtual events called



Paula Ramos is a CV, AI and IoT Evangelist at Intel Corporation.

Read 100 FASCINATING interviews with Women in Computer Vision

Paula, what are you an evangelist for?

I advocate for people around Intel products. My main role is to advocate for OpenVINO and Intel Geti. Do you remember Intel Geti?

Of course. So why are you promoting them?

We have a lot of benefits. We're not salespeople, so understanding the role of an evangelist is a little tricky. When I say I am an evangelist, people laugh. *"Evangelist, is it a religion or something like that?"* We are trying to create a community around the product. We are trying to find more developers. So right now, we have around 28 million developers in the world, but just 300,000 developers are working with AI.

Does that make you a good evangelist?

[*laughs*] The idea is to increase that!

"I was an evangelist in the past, and I didn't realize it!"

Imagine right now, just 1% of the developers in the world are working with AI. The idea is to increase the percentage so that maybe 10% of the developers work with AI.

How did you become an evangelist?

Well, that's a really good story! My background is in electronic and electrical engineering. I have a PhD in computer vision. But I was working in agriculture for 17 years, and I was working in Colombia, South America, for the National Federation of Coffee Growers of Colombia. I was trying to develop and deploy technologies to them. I was working with different machines to process coffee after the harvesting process. Also, I was working with robotic tools to harvest.

So you worked a lot in hardware, even more than software?

Yes, and at the very beginning. In the early 2000s, I was working with classic machine-learning embedded devices. I was programming assembly code in a machine, and at that time, I was working with edge computing for coffee growers. But we had a lot of restrictions and barriers. I was thinking it's low cost, easy to use, and easy to maintain equipment. It also allows me to be creative and to close the gap between technology and users. And just for personal reasons, I needed to move to the United States. I started working at North Carolina State University, which is on the east coast. It's a beautiful place. You need to visit there! I was working there in agriculture, but also with corn, soybean, and different kinds of crops that had more money to invest. But I always use my philosophy - low cost, easy to use, easy to maintain - and I deployed three products in total. I was working with technicians and non-technical people. Technicians, biologists, agronomists… not engineers. I needed to bridge the gap. I participated in two OpenCV AI competitions in 2020 and 2021.

That's how you spent your COVID years?

Yes, but I was working at the university.

Paula was featured last year as Woman in Computer Vision.

DevCon webinars and its big conference, **Intel Innovation**, which takes place September 24-25 in San Jose, California. Despite such broad engagement, there is still a common misconception that Intel is solely a hardware company. *"People ask: 'You're a hardware company; why are you doing AI?'"* Paula reveals. *"**We're doing AI because we want to enable AI in the hardware, but we're also working beyond that!**"*

From the viewpoint of their booth, where the team has been busy teaching people how to use AI with Intel, the team has noticed that the awareness of its AI solutions has grown significantly this year. Many more people are familiar with OpenVINO, and generative AI is an exciting new trend. *"There's an awareness that all our solutions are better right now,"* Paula notes. *"It will keep getting better year by year."*

"… the opportunity is today! Grab it and stay strong with the CVPR community! "

Intel has a unique capability to create **a comprehensive environment for AI development**. With millions of devices everywhere, it can write software that millions of people can start benefiting from overnight. Its end-to-end capabilities in designing, manufacturing, shipping, and inferencing AI solutions place it in a position of **significant responsibility**, where it must ensure ethical AI practices and develop explainable AI solutions. **Work like Ilke Demir's on deepfake detection is important, as are its trust and security services**.

As our interview concludes, Raymond offers a powerful message to the community, reflecting on the rapid advancements in the field. "*Opportunities come every year,*" he says. "*When I was a researcher, everyone said 10 years. With this new generation, my advice is:* ***the opportunity is today! Grab it and stay strong with the CVPR community!*** *The community defines this field, and the CVPR community is the leader.* ***We should lead and then make your dreams come true!***"

"... No, I want this to be in the market tonight!"

SpiderMatch: 3D Shape Matching with Global Optimality and Geometric Consistency

Paul Roetzer (left) is a PhD student under the supervision of Florian Bernard (right), an Associate Professor at the University of Bonn and the Head of the Learning and Optimisation for Visual Computing Group.

Before their oral presentation this afternoon, they speak to us about their highlight paper on 3D shape matching, which has also been chosen as a Best Paper Award candidate.

The problem of **3D shape matching** involves identifying correspondences between surfaces of 3D objects, a task with applications in medical imaging, graphics, and computer vision. This work's main novelty is that it accounts for **geometric consistency, a property often neglected in previous 3D shape matching methods due to its complexity**.

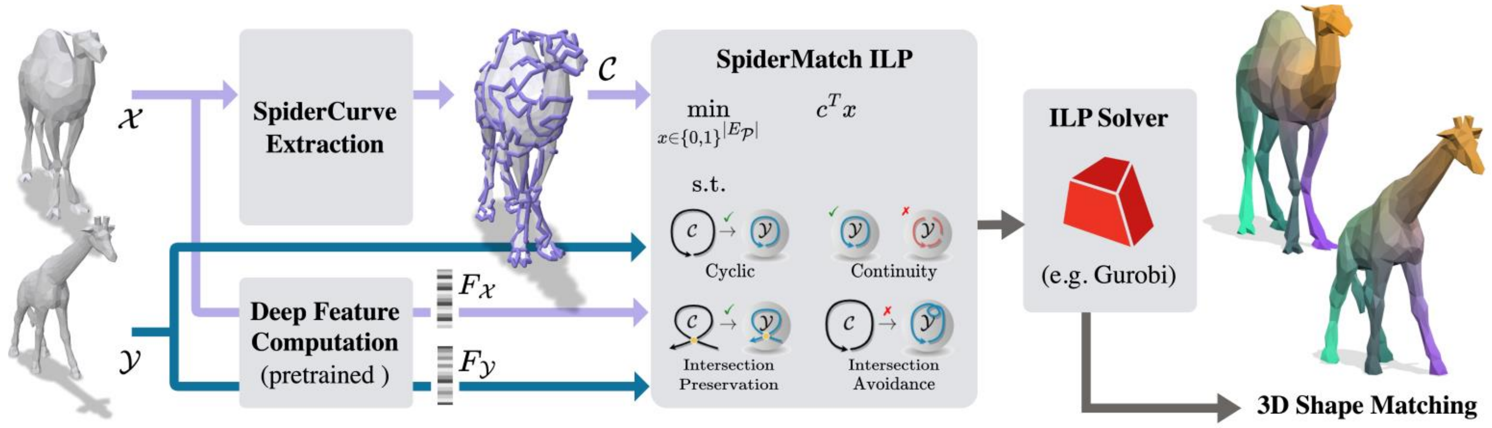
Geometric consistency ensures that when matching the surface of one shape to another, **the neighboring elements are matched consistently,**

Computer Vision News
Publisher: **RSIP Vision**
Copyright: **RSIP Vision**

Editor: **Ralph Anzarouth**

All rights reserved
Unauthorized reproduction is strictly forbidden.

Our editorial choices are fully independent from IEEE, CVPR and all conference organizers.



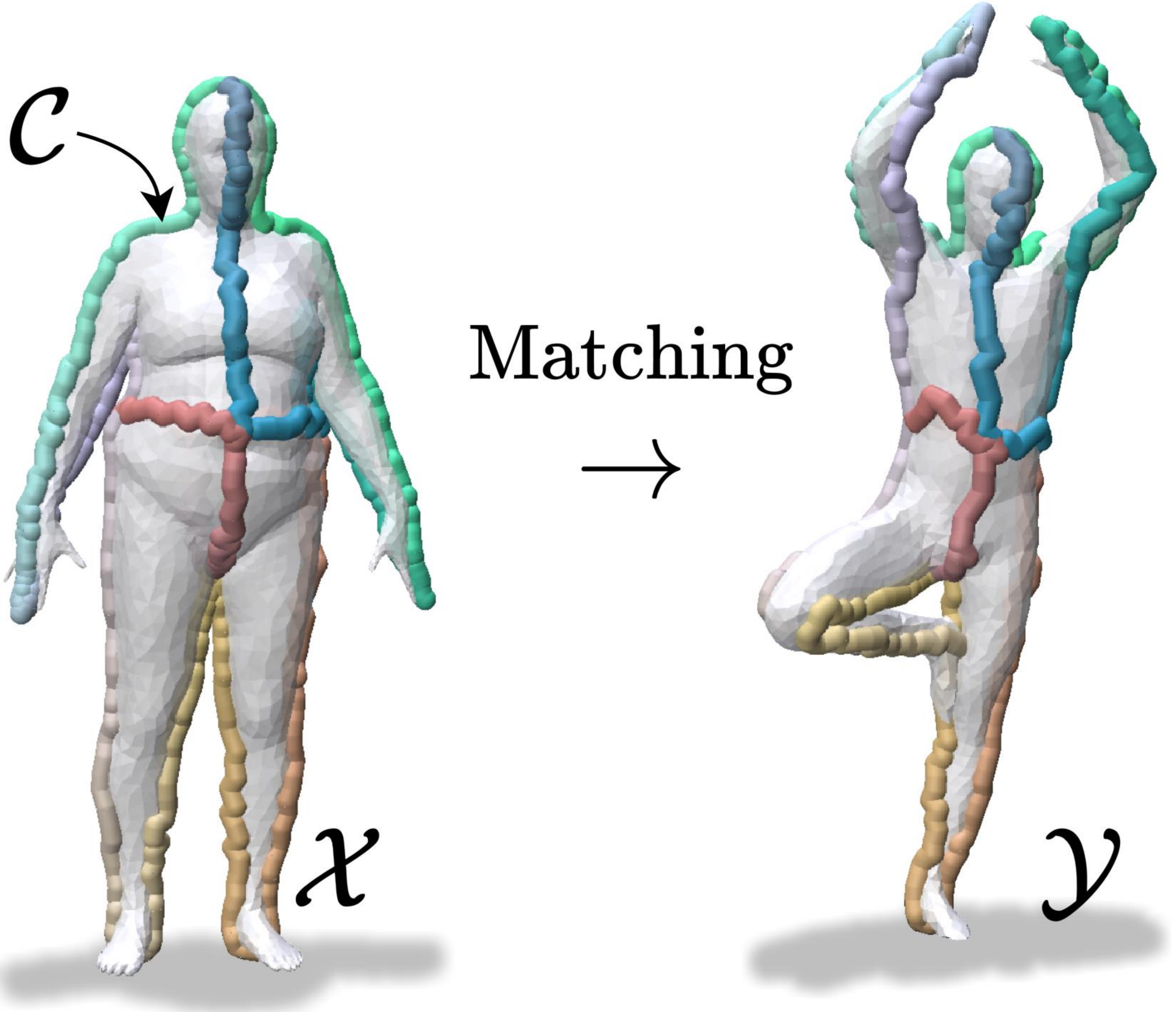
preserving neighborhood relations. “*Imagine two organs, like the liver, heart, or lungs, and you match them from different people,*” Florian explains. “You *take the shapes and want to train a statistical shape model. If you didn’t have this geometric consistency, the deformation from one to the other would lead to self-intersecting sections which aren’t anatomically plausible.*”

Many existing approaches to 3D shape matching do not enforce geometric consistency as a hard constraint but optimize it as a soft objective, often framed as graph matching or quadratic assignment problems. “*This is a problem class well known to be* ***NP-hard****, making it extremely challenging to solve for large instances in practice,*” Florian tells us. “We *find a different representation that makes the problem easier to solve.*”

Florian and Paul propose **a novel path-based formalism**, representing one of the 3D shapes (the source shape) as a long, self-intersecting curve (‘**SpiderCurve**’) that traces the 3D shape surface. This alternative discretization simplifies the 3D shape matching problem to find the shortest path in the product graph of the SpiderCurve and the target 3D shape. “*This switch of the discretization is what makes our paper novel,*” Paul points out. “*We think differently about a problem, making a very complicated task to solve a simpler one.*”

This formalism leads to **an integer linear programming problem**, which the team demonstrates can be efficiently solved to global optimality. The result is competitive with recent state-of-the-art shape matching methods and guarantees geometric consistency. “*For the first time, we can find geometrically consistent shape matchings while also finding global optima in practice,*” Florian reveals. “*Within the framework of our optimization formulation, in all the instances that we’ve evaluated, we know that* ***we have the best possible solution among all potential solutions!***”

3D shape matching is just one of a class of matching problems that are fundamental to computer vision. Could devising an innovative new approach to solving such a fundamental problem be part of the reason the paper has been chosen as *a* **candidate for a Best Paper Award**? *"We have conceptually a pretty simple idea,"* Florian responds. "*Instead of representing a 3D shape using triangles as discretization, we simply discretize the 3D shape using a one-dimensional curve that traces the surface while visiting all the vertices.*



Idea: Match Cyclic Path $\mathcal{C}$ on $\mathcal{X}$ to $\mathcal{Y}$

*By looking at a different representation of the 3D surface, we can build on well-established frameworks for global optimal matching problems that lead to geometric consistency. I think **the secret is the simplicity** and the fact that it's very fast in practice."*

Away from writing top-rated papers, Florian works at the intersection of machine learning and mathematical optimization in visual computing. Meanwhile, Paul explores solutions to 3D shape matching problems with optimization methods.

Looking ahead, Florian acknowledges an unresolved challenge: *"The most critical open problem is **whether an algorithm exists to solve this problem in polynomial time,**"* he ponders. *"What we have is fast in practice, but the worst-case time is still exponential. The next step would be to investigate if it's possible to come up with a similar formalism that could lead to a polynomial time algorithm that is provably fast."*

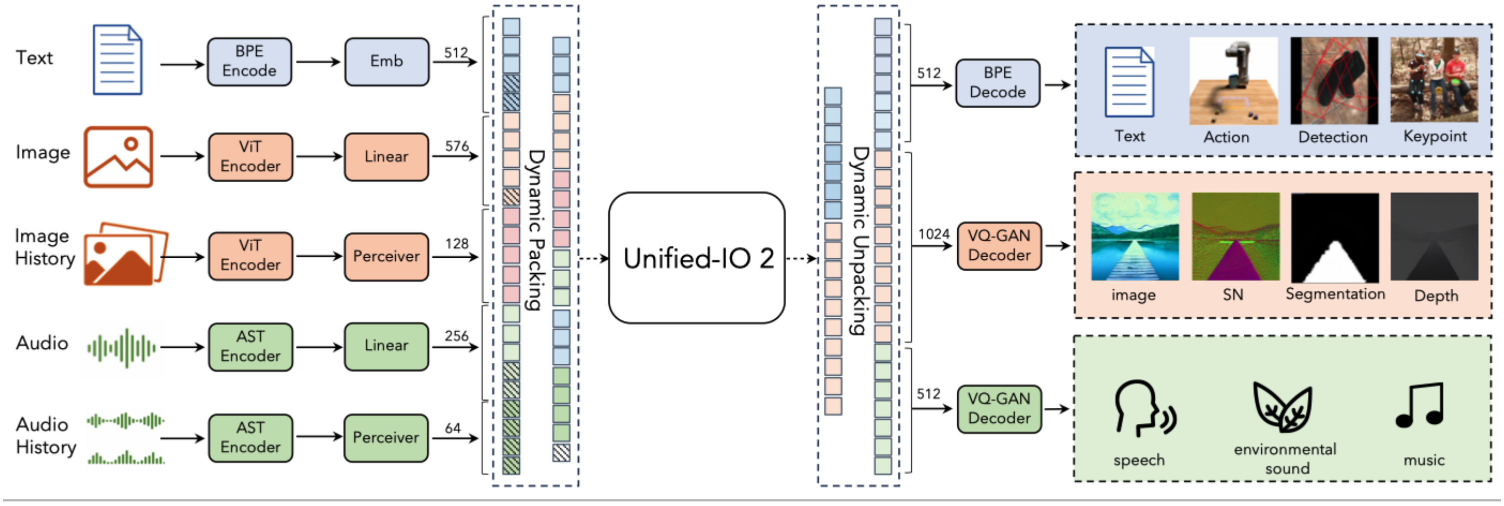
The **BEST OF CVPR 2024** continues on the next page with another exceptional Oral Paper!

Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision Language Audio and Action

Aniruddha Kembhavi (top left) is a Senior Director at the Allen Institute for AI (AI2), leading the Perceptual Reasoning and Interaction Research (PRIOR) team, where Christopher Clark (center) and Jiasen Lu (top right) are Research Scientists, Sangho Lee (bottom left) is a Postdoctoral Researcher, and Zichen "Charles" Zhang (bottom right) is a Predoctoral Young Investigator.

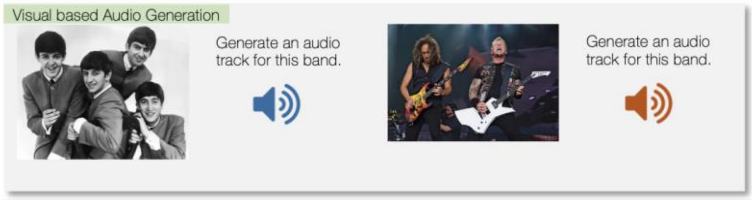
They spoke to us about their highlight paper proposing Unified-IO 2, a versatile autoregressive multimodal model.

Unified-IO 2 is the first autoregressive multimodal model capable of understanding and generating images, text, audio, and action. It can handle multiple input and output modalities and incorporates a wide range of tasks from vision research. Unlike traditional models with specialized components for different tasks, it uses a single encoder-decoder transformer model to handle all tasks, with a unified loss function and pretraining objective.



"It's a super broad model," Christopher tells us. *"It can take many different modalities as input and output. It can do image, text, audio, and video as input and can generate text, image, and audio output. Within those modalities, we basically threw in every task we could think of that vision researchers have been interested in. It's a super, super broad model. I think it's one of the most broadly capable models that exists today."*

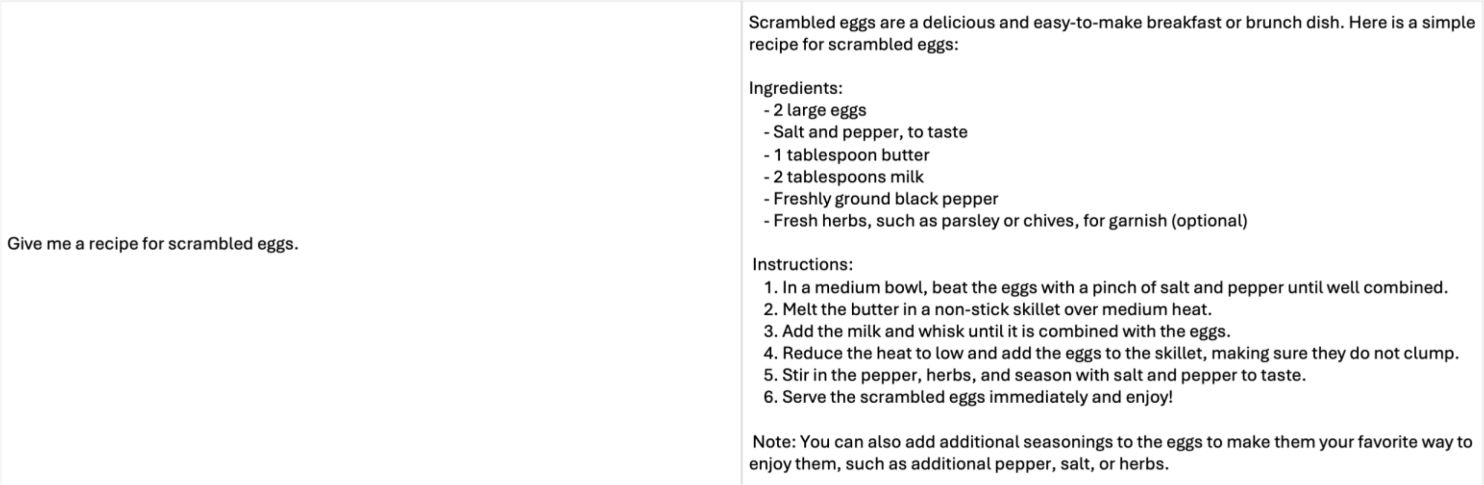
While language models can perform many tasks and input and output all kinds of structured language, handling diverse inputs and outputs in computer vision is more challenging. *"When it comes to computer vision, it's a mess,"* Aniruddha says bluntly. *"Sometimes, you have to input an image. Sometimes, you have to output a bounding box. Sometimes, you have to output a continuous vector like a depth map. Inputs and outputs in computer vision are very heterogeneous. That's why, for the last 10 years, people have been building models that can do one or two things."*


Visual based Audio Generation
Generate an audio track for this band.
Generate an audio track for this band.

Visual Audio Localization
(drum sounds)
Identify the locations of the instruments producing the given sound.

Unified-IO 2 builds on the foundations laid by its predecessor, Unified-IO, aiming to create a model that can truly input and output anything. Training such a comprehensive model, especially with limited resources, has been incredibly tough. The team's first major challenge was collecting the pretraining and instruction tuning data. The second was training a multimodal model from scratch rather than adapting existing unimodal models.

"We tried a few months of tricks to stabilize the model and make it train better," Jiasen recalls. *"We figured out a few key recipes that were used by later papers and shown to be very effective, even in other things like image generation.* ***We're training on a relatively large scale with 7B models and over 1 trillion data. More than 230 tasks were involved in training these giant models.****"*



Give me a recipe for scrambled eggs.

Scrambled eggs are a delicious and easy-to-make breakfast or brunch dish. Here is a simple recipe for scrambled eggs:

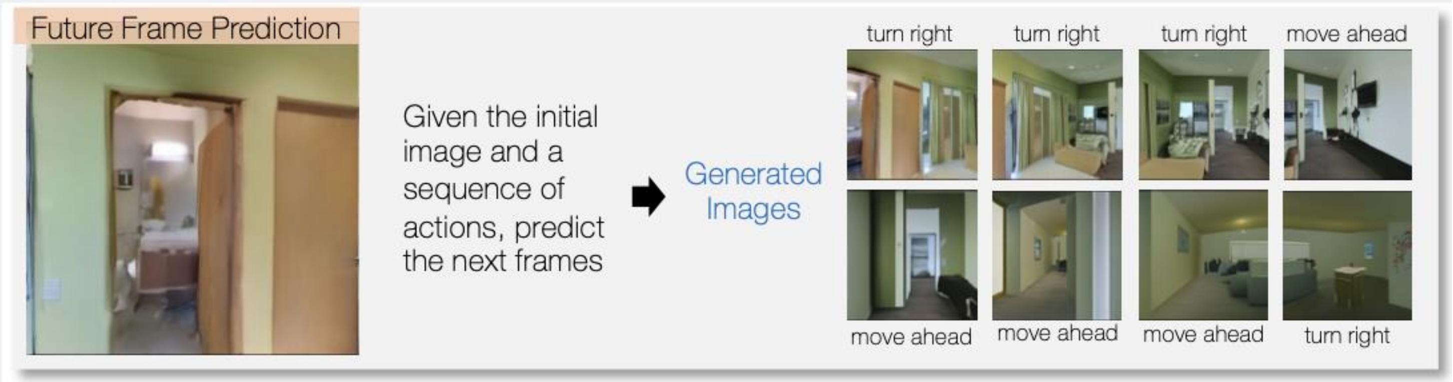
Ingredients:
- 2 large eggs
- Salt and pepper, to taste
- 1 tablespoon butter
- 2 tablespoons milk
- Freshly ground black pepper
- Fresh herbs, such as parsley or chives, for garnish (optional)

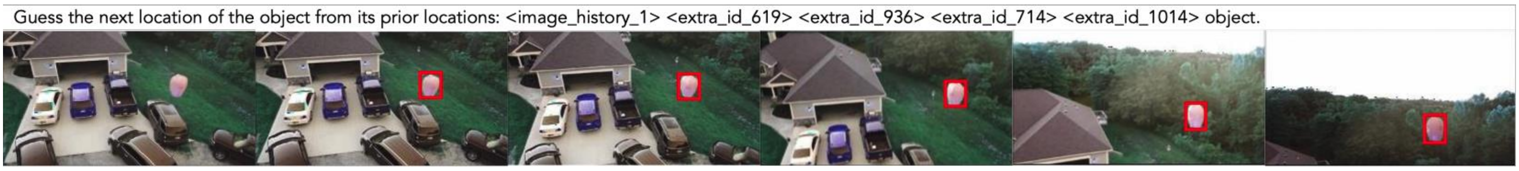
Instructions:
1. In a medium bowl, beat the eggs with a pinch of salt and pepper until well combined.
2. Melt the butter in a non-stick skillet over medium heat.
3. Add the milk and whisk until it is combined with the eggs.
4. Reduce the heat to low and add the eggs to the skillet, making sure they do not clump.
5. Stir in the pepper, herbs, and season with salt and pepper to taste.
6. Serve the scrambled eggs immediately and enjoy!

Note: You can also add additional seasonings to the eggs to make them your favorite way to enjoy them, such as additional pepper, salt, or herbs.

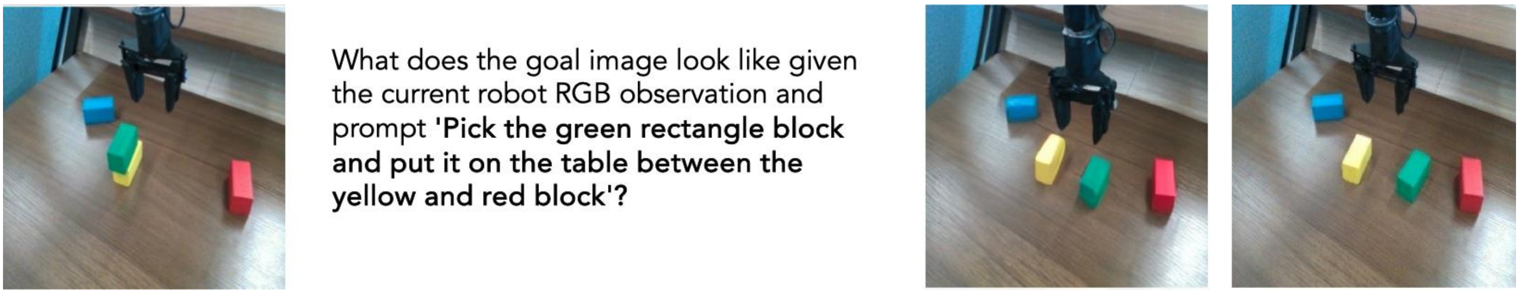
The development of Unified-IO 2 has been a collaborative effort involving the four first authors: Jiasen, Christopher, Sangho, and Charles. Aniruddha is keen to ensure they get the recognition they deserve for the feat they have pulled off. *"This project is a Herculean effort by these four people,"* he points out. *"Usually, people will take a large language model, then put a vision backbone, and then finetune that on some computer vision tasks. In this model, the language model is also trained from scratch. Think of large companies with hundreds of researchers trying to train a language model. Contrast that with this paper, which has four first authors trying to train a model that does everything. These four gentlemen have toiled night and day for many, many months. I can testify to that."*

Everything about Unified-IO 2 is open source. If you visit the team's poster today, you can feel safe knowing they are willing to share every aspect of the project. *"We've released all the data, the training recipes, the challenges, especially in stabilizing the model training, and all the evaluation pipelines,"* Sangho confirms. *"If you come to our poster booth, we'll be very happy to* ***share all the recipes and know-how for training this special kind of multimodal foundation model****."*





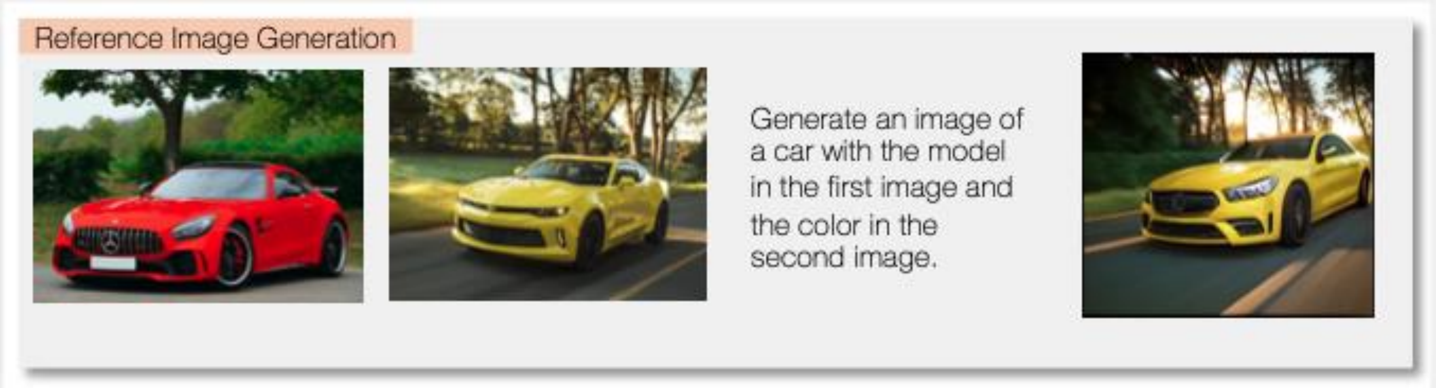
Guess the next location of the object from its prior locations: <image_history_1> <extra_id_619> <extra_id_936> <extra_id_714> <extra_id_1014> object.



What does the goal image look like given the current robot RGB observation and prompt **'Pick the green rectangle block and put it on the table between the yellow and red block'**?

During the evaluation stage, the team discovered that Unified-IO 2 could perform well in tasks they had not initially targeted, such as video tracking and some embodied tasks. They will showcase these surprising results with iPad demonstrations at their poster session. *"We've tested the model multiple times, but maybe only with a few modalities and target tasks,"* Charles reveals. *"It's a surprise that the model is so good at other tasks we've not focused on before. There are lots of interesting behaviors of the models and some very cool visualizations that the model can follow some novel instructions."*

The paradigm behind Unified-IO 2, where all modalities are integrated into a single transformer without relying on external unimodal models, is a promising direction for future AI research. *"It's in contention with other ways of training generalist models, and people are still exploring and building on that,"* Christopher adds. *"I think Unified-IO 2, in particular, has a lot of modalities and tasks and really* ***pushes that way of building models to an extreme****."*



Generate an image of a car with the model in the first image and the color in the second image.

Russian Invasion of Ukraine

CVPR condemns in the strongest possible terms the actions of the Russian Federation government in invading the sovereign state of Ukraine and engaging in war against the Ukrainian people. We express our solidarity and support for the people of Ukraine and for all those who have been adversely affected by this war.

Denys Rozumnyi (top left) has won the Structured Semantic 3D Reconstruction Challenge, which was held as part of the Urban Scene Modeling workshop here at CVPR 2024. Denys is open to work offers. He's a great catch! Grab him before it's too late.

Yaroslava Lochman (top right) is a PhD student at Chalmers University of Technology in Gothenburg, Sweden. She presented her paper, which explores the problem of motion segmentation, during her poster session at CVPR 2024.

Sophia Sirko-Galouchenko (here on the left) is a first-year PhD student at Sorbonne University in Paris and Valeo.ai. Her paper on bird's-eye-view perception in autonomous driving was presented by colleagues during a poster session at the CVPR Workshop on Autonomous Driving.

Read 150 FASCINATING interviews with Women in Science

"Now suddenly, everything is converging into one big model, which can do these various things, which is exciting but also maybe a bit scary!"

Ivana Balažević is a Research Scientist at Google DeepMind.

She spoke to us at CVPR, right after her talk and panel at the Prompting in Vision workshop.

Where does Balažević come from?

Croatia.

I am not very far away. I am Italian.

Ah, okay, so we're neighbors! [*she laughs*]

What is your work about?

Well, various different things. I have mainly, in the past couple of years, worked on multimodal image and video understanding. As of recently, I moved into Gemini, working more on language. But, yeah, Gemini, super secret, can't talk about it – you know how it is!

Is convergence of multimodalities really happening? Text with vision, video, audio, all these things together?

I think it is, especially in the past couple of years. I finished my PhD in 2021, and there were just these small models doing various different tasks. Everyone was working on their little model in their PhD or in whichever company, and now suddenly, everything is converging into one big model, which can do these various things, which is exciting but also maybe a bit scary. I don't know. Mainly exciting, I would say.

Why exciting, and why scary?

That's a very good question! In my mind, exciting because it unlocks a whole world of possibilities for what we can possibly do with these models in some possibly distant future. I don't know because I didn't think we'd be where we are now, but we would maybe be able to learn from these models or learn something new that we don't know. These models might be able to make some sort of inferences, like combining various modalities to teach us things. How amazing would it be if we had a model that would be able to read scientific papers and come up with a new paper that is actually correct and that teaches us something new? Or a model that takes all our knowledge about medicine and biology and chemistry and finds a cure for cancer or something like that? Some of these things would be really, really amazing. Scary because, well, as with anything, people can abuse these sorts of models.

And you cannot control the person who wants to abuse.

Exactly. Any tool probably in human history – not any, but a lot of them can be used for good and for bad things. You have a kitchen knife that you can cut vegetables with or something…

Or cut the neighbor!

Yeah, exactly! [*both laugh*] It's the same with the technology nowadays.

Do you know what Nobel invented? Dynamite!

Yeah, Nobel, exactly. There we go! This is a very good example.

Would you agree with Yann LeCun, whom I have interviewed twice? He says that today's AI is no smarter than a home cat.

Than a cat? Well, probably not at this point, but we will see how much time will pass until they are smarter than a cat, so we'll see! [*she laughs*]

Is it exactly what you wanted to do to be at the intersection of all these nice things at the moment they are going to intercept?

Yeah, I think it's a nice place to be. It's sometimes a bit uncomfortable because things are moving really, really fast. As I said, during my PhD time, everything felt more chill, whereas nowadays, there's a lot going on, but it's also very fun.

Is this why you do research to have fun in something innovative?

I mean, partially. There are various different reasons why I do research. I think I'm bored quite easily, so I like to do things that are constantly new, so that your day to day isn't just repeating the same things over and over again. Also, because of what I just mentioned, all these amazing

things that these tools that we're developing can potentially help us unlock. But also to have fun.

I am sure you are not afraid you are going to be bored by AI in the coming years.

No, I don't think so. The opposite, actually.

In your short career until now, what in AI particularly made you say, 'wow'?

That's a very good question. I remember we had this Flamingo model in DeepMind. Before it came out, we could play around with it internally a bit, and I remember uploading a picture of my dad's cat and asking it various questions about the picture, and it could answer everything. I was like, how is this possible? That was my first moment like, wow, these things actually work! Because I was pretty much a sceptic before that about this kind of increase the model size, increase the data size. I wasn't thinking this is actually going to work. I mean, there's probably a limit. I still think there's a limit, but we're still pushing this frontier it would seem.

What is the next 'wow' you are going to say?

If we have actual embodied agents. If we can integrate all of these current assistants we have and put it into a robot This is the point where I would actually find it scary as some sort of actual sci-fi moment, but it would also be like, wow.

What would be your dream for the next 10 years?

Personally, I would like to move into the more application side of things, where you can do something good with these models. The model development is very, very interesting as well, but this has kind of been a change in the past couple of years for me, where we've moved on from models that don't really work to models that do work, and now it's unlocked this whole world of possibilities where you can actually use it to solve real-world problems.

The spectrum of things that you can do is infinite. If I only think about climate change, the number of issues that AI can help on is infinite.

Exactly.

Will you pick one of those?

Good question. I think we have so many open problems in our world, from various political problems to climate change. I think climate change is a good one, maybe something medical. I find some of the companies that do research into longevity, not necessarily extending the length of human life but improving the quality of life at a later stage where I think there's a lot of ML used there, I think these are the things that I find interesting.

You are preparing for retirement.

Yeah, exactly. I want to retire in a nice way! [*both laugh*] Still be able to go on hikes or something.

Scuba diving?

Yeah, perfect. Scuba dive at 90 or something like that! [*Ivana laughs*]

We have spoken about the future; we did not speak much about the past. Why did you leave Croatia?

Well, that was a very long time ago now. It was 12 years ago. Yeah, that's when I left Croatia. I wanted to see what there is outside Croatia. I wanted to experience life outside Croatia, but I didn't think I was going to stay. I initially thought it was going to be a couple of years, but then, here we are.

What did you do for this couple of years?

I did a Master's in Berlin at first. Then, I was in California for a year doing an internship at a startup, and then I started doing my PhD in Edinburgh, Scotland.

So, it was not the plan?

No, not initially. It was kind of let's go and see sort of thing.

Is this the way you decide things?

When I was young, for sure. Now, sometimes!

Are you going to use the same criteria for things happening in the future?

I mean, not so much. Now, I tend to overthink things a bit more than when I was 21.

What is the thing that you did until now that you are the most satisfied with?

I'm going to say something that's unrelated to my career. I ran a marathon a month and a half ago! Yeah, it was hard, but also, it made me feel like, oh, I can do this.

> ***"I think it's this thing about not giving up. Knowing that you can do more than you think you can!"***

Wow, I'm so jealous. I registered for three marathons. I was never able to start one. I always got injured.

No way. It's very annoying. I'm still a bit injured now. My feet hurt, but other than that, I enjoyed it a lot. It was one of those things because I often doubt myself, and I'm like, this is too hard, I cannot do this, I cannot do that, and it's one of those things where you're like, I can actually do this.

Four hours?

3:57.

Tell me one thing that our readers can learn from your marathon.

I think it's this thing about not giving up. Knowing that you can do more than you think you can because I feel like it's one of those things where you have to believe that you can do it. But yeah, as I said, I'm not that kind of person.

Some people will say that life is already competitive enough.

True, but I think it makes you more robust to everyday life challenges because it teaches you, yeah, I can do this. I can overcome this thing that I wasn't able to do.

I am sure that finishing a marathon makes you very, very high.

Yeah, it does. I was laughing and crying at the same time. I was in this weird state. [*Ivana laughs*]

How many CVPRs have you attended?

This is my first one. I have never been to CVPR before. I normally go to ML conferences.

How are you finding it?

Yeah, I think it's nice. It's great. I don't know how big it is, but from what I've heard, it's bigger than NeurIPS.

11,500 this year, including online.

Yeah, I'm excited for it. Let's see.

What did you learn from the workshop?

That there are a lot of different opinions and ways people do prompting. I don't know if that's a good thing or a bad thing. In language, prompting is a well-defined thing, whereas here, it's very much an open research problem open for discussion. It's possibly a good thing. There might be good things coming out of it.

Is there a lifelong dream or goal you hope to achieve before you retire?

I want to work on a particular problem that matters right now. I'm now doing research for the sake of research because it’s interesting. I haven’t decided which of these important problems we talked about yet is that one problem, but I want to do something like that and focus and try to help achieve that.

Tell us one thing about you that we do not know.

Again, that’s a very difficult question.

I do not know much!

I don’t think I like to talk about myself that much! [*she laughs*]

Oh, so you had a great idea of accepting my offer for an interview.

Yeah, to be honest, I was not sure what I was going to say!

Now, it’s too late!

I know. People have convinced me otherwise.

Who convinced you?

Well, my manager and my boyfriend. They were both like, *“No, you should do it. It’s a good thing. It inspires people!”* I’m like: *“I don’t know if I like doing these self-promotion things.”*

I often discover the most surprising things about people I do not know.

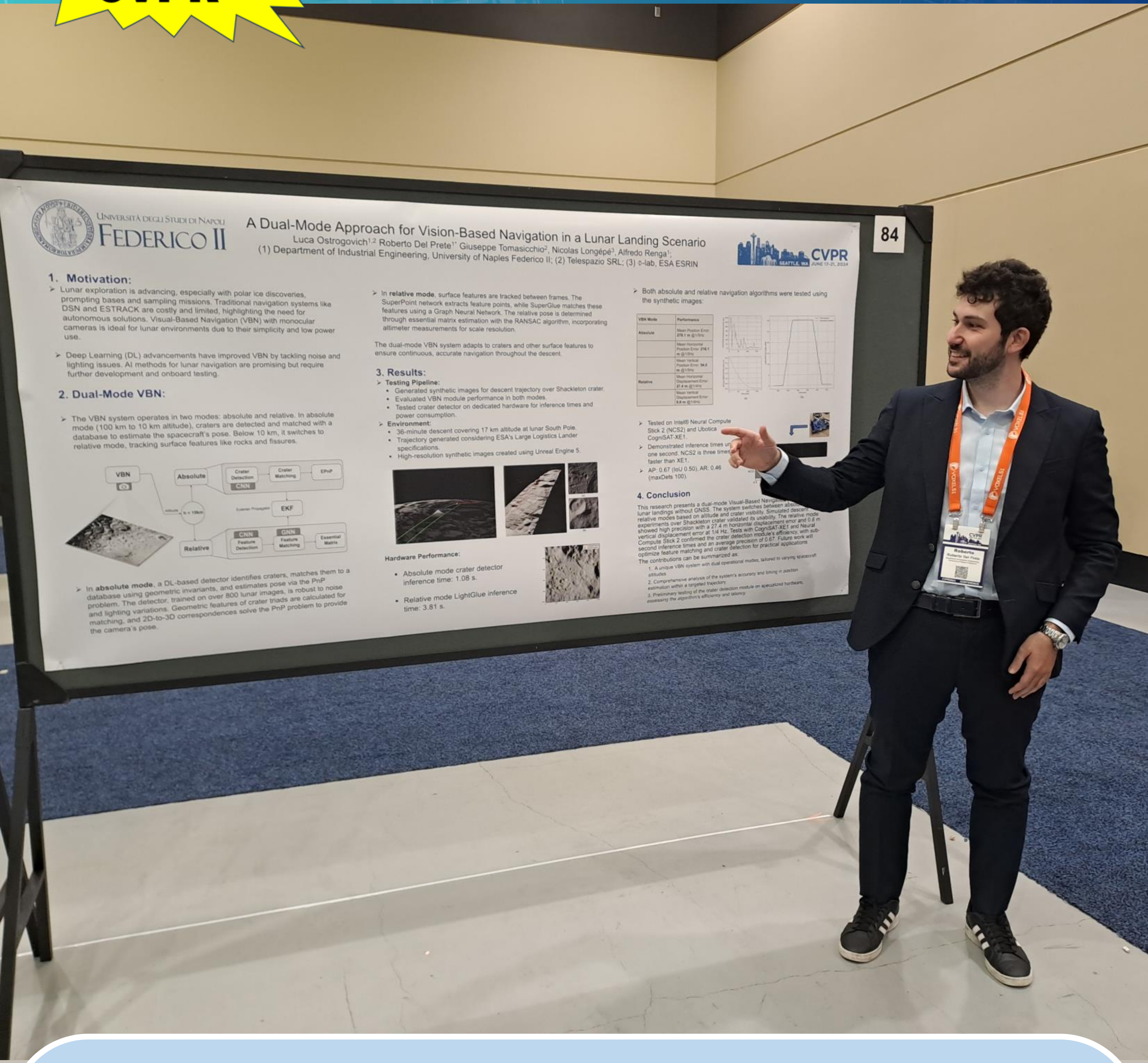
Yeah, well, there we go. That’s one thing you didn’t know about me! [*both laugh*]

Do you have a final message for the community?

Do big things that matter because I think you can achieve it. There are so many smart people in this field and there are many, many interesting problems to be solved. I think if we put our minds together, we can improve this world that we live in.

Read 150 FASCINATING interviews with Women in Computer Vision!

AI4Space Workshop Poster


A Dual-Mode Approach for Vision-Based Navigation in a Lunar Landing Scenario
Luca Ostrogovich[1,2] Roberto Del Prete[1*] Giuseppe Tomasicchio[2], Nicolas Longépé[3], Alfredo Renga[1];
(1) Department of Industrial Engineering, University of Naples Federico II; (2) Telespazio SRL; (3) ɸ-lab, ESA ESRIN
Università degli Studi di Napoli Federico II
CVPR
84

Roberto Del Prete, a PhD student from the University of Napoli, who is set to complete his PhD in November, is proud to present at his first CVPR conference.

His poster, presented at the AI4Space workshop, discusses a novel approach for autonomous lunar landing, leveraging visual information.

The workshop highlighted the space capabilities that draw from and/or overlap significantly with vision and learning research, outline the unique difficulties presented by space applications to vision and learning, and discuss recent advances towards overcoming those obstacles.

Test-Time Adaptation with SaLIP: A Cascade of SAM and CLIP for Zero-shot Medical Image Segmentation

Sidra Aleem is a final-year PhD researcher at ML-Labs, Dublin City University, focusing on domain adaptation for biomedical imaging using foundation models.

Following her oral presentation on Monday at the CVPR 2024 Workshop on Domain adaptation, Explainability, and Fairness in AI for Medical Image Analysis (DEF-AI-MIA), she speaks to us about her paper on test-time adaptation with foundation models.

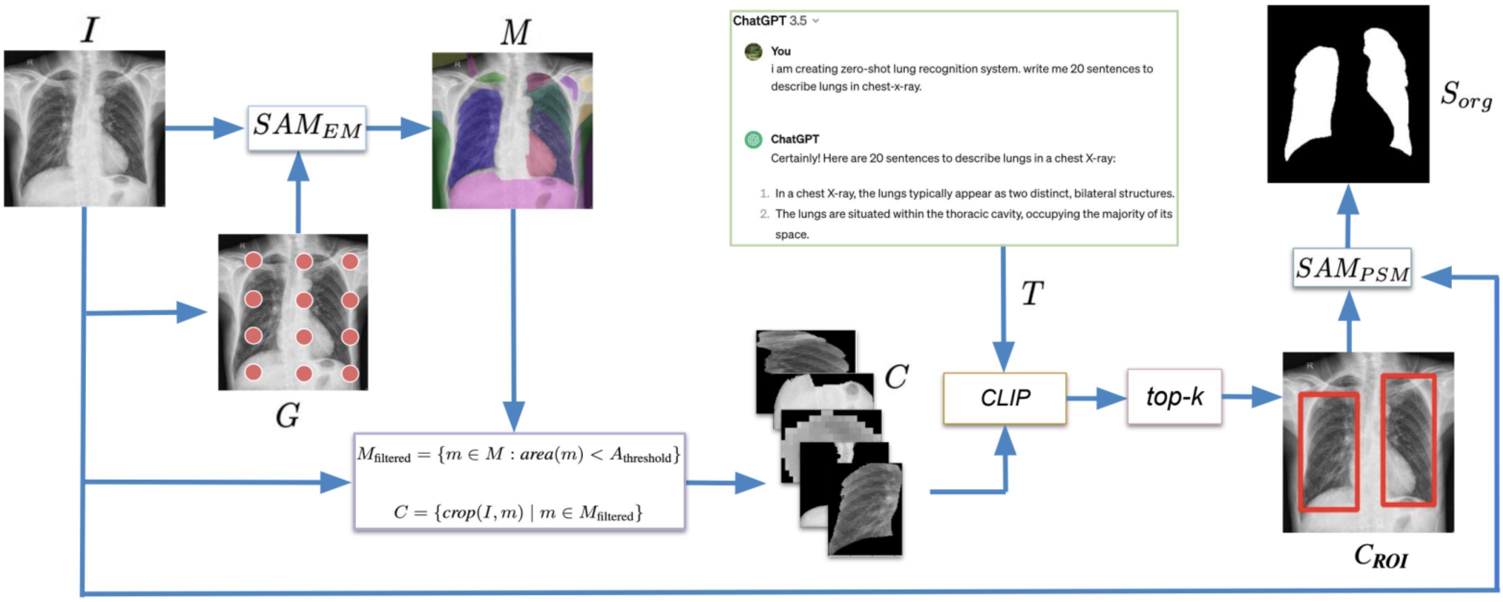
Medical image segmentation, critical for clinicians in diagnosis and prognosis, is the focus of Sidra's innovative work. She proposes a novel cascade of two foundation models, Meta's **Segment Anything Model (SAM)** and OpenAI's **CLIP**, leveraging their unique capabilities to enhance zero-shot organ segmentation accuracy in medical imaging. *"These foundation models have completely revolutionized the world around us,"* she tells us. *"While they've been predominant in natural imaging, their effective application has yet to be explored in medical image segmentation."*

Her approach involves using SAM to generate all the different region proposals from medical images. She then employs CLIP, a multimodal model designed to process text and images, to identify the specific organ for segmentation. While CLIP has already been extensively tested on natural images, Sidra has successfully adapted it to the unique challenges of medical imaging. *"As we know, one of the widely used applications of CLIP is image retrieval,"* she explains. *"My objective was to utilize CLIP to get the region of interest from all these region proposals. For the text part, in medical imaging, we need domain knowledge. **To mitigate that issue, I generated text prompts using ChatGPT**."*

Regarding lung segmentation, Sidra used ChatGPT to generate 20 attributes to describe the lungs in a chest X-ray. These prompts were then fed into CLIP’s text encoder, which calculated the similarity of all the SAM-generated region proposals and the text prompts and retrieved the relevant mask from the pool of SAM-generated masks. Finally, SAM was prompted by the retrieved region of interest to segment the lung.

The **SaLIP** model has been tested on medical imaging datasets encompassing MRI scans, ultrasound, and X-ray images and diverse segmentation tasks, including brain, lung, and fetal head. *“The lung segmentation was more challenging, as there were two regions of interest – the left and right lung,”* she recalls. *“As reported in the paper, the performance was really good!”*

A significant contribution of this work is that **it employs both the ‘segment everything’ and ‘promptable’ modes of SAM connected through CLIP**. This dual-mode approach allows for more precise and varied applications in clinical settings. Previous work adapting SAM in medical imaging has used it to segment everything in the image. However, clinicians want to focus on specific regions of interest, which vary depending on the clinical need. *“First, I utilized the segment everything mode of SAM to create region proposals for every region in the image,”* Sidra explains. *“Then, I used the promptable mode, where we use specific prompts to segment a specific region. To connect both these modes, I used CLIP as a bridge between them. To the best of my knowledge, none of the work in medical imaging has utilized both modes of SAM. The literature mainly focuses on finetuning either the segment everything or promptable mode.”*


I
SAM_{EM}
M
G
$M_{\text{filtered}} = \{m \in M : area(m) < A_{\text{threshold}}\}$
$C = \{crop(I, m) \mid m \in M_{\text{filtered}}\}$
C
ChatGPT 3.5
T
CLIP
top-k
C_{ROI}
SAM_{PSM}
S_{org}

You
i am creating zero-shot lung recognition system. write me 20 sentences to describe lungs in chest-x-ray.

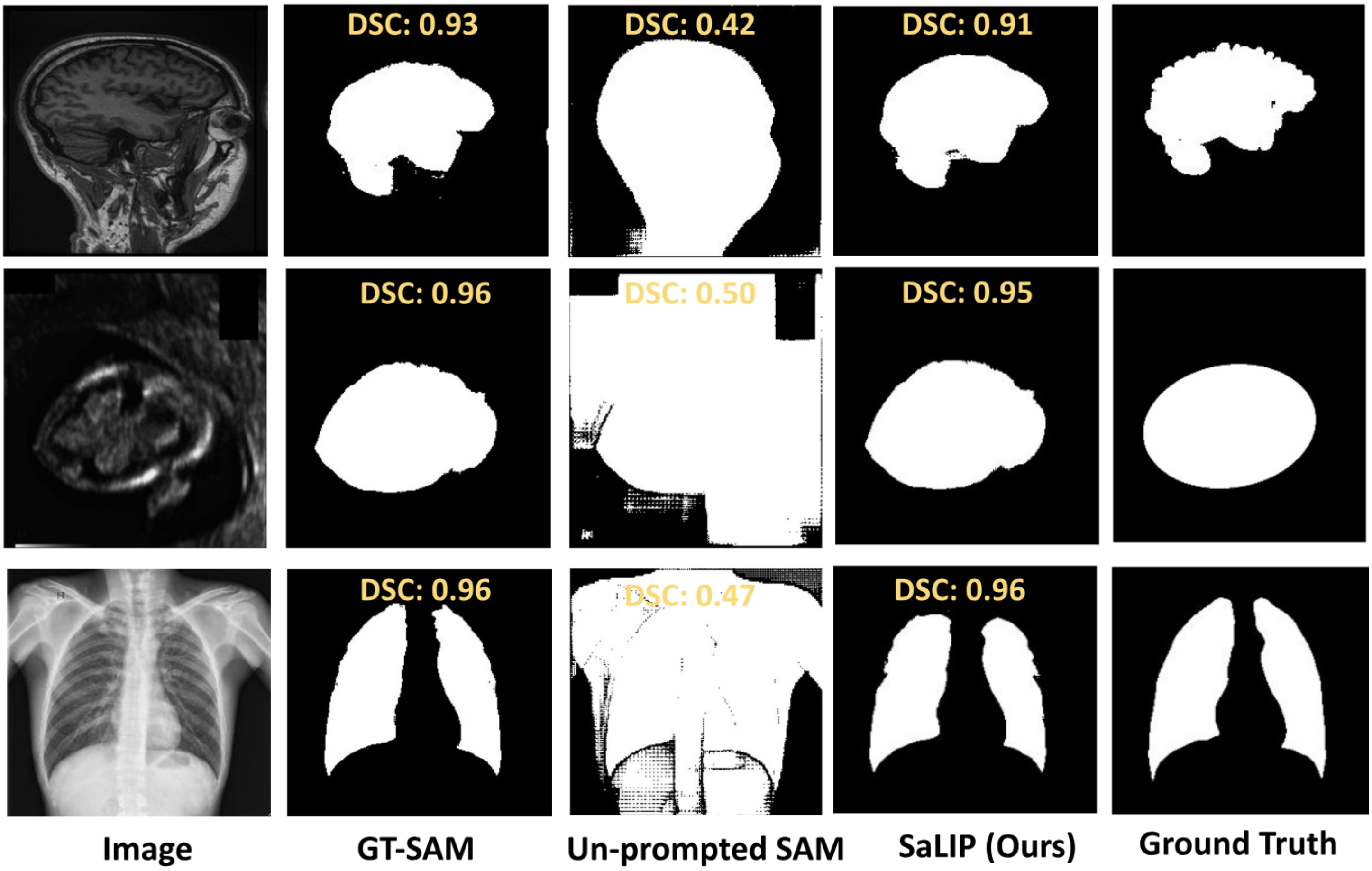
ChatGPT
Certainly! Here are 20 sentences to describe lungs in a chest X-ray:

1. In a chest X-ray, the lungs typically appear as two distinct, bilateral structures.
2. The lungs are situated within the thoracic cavity, occupying the majority of its space.

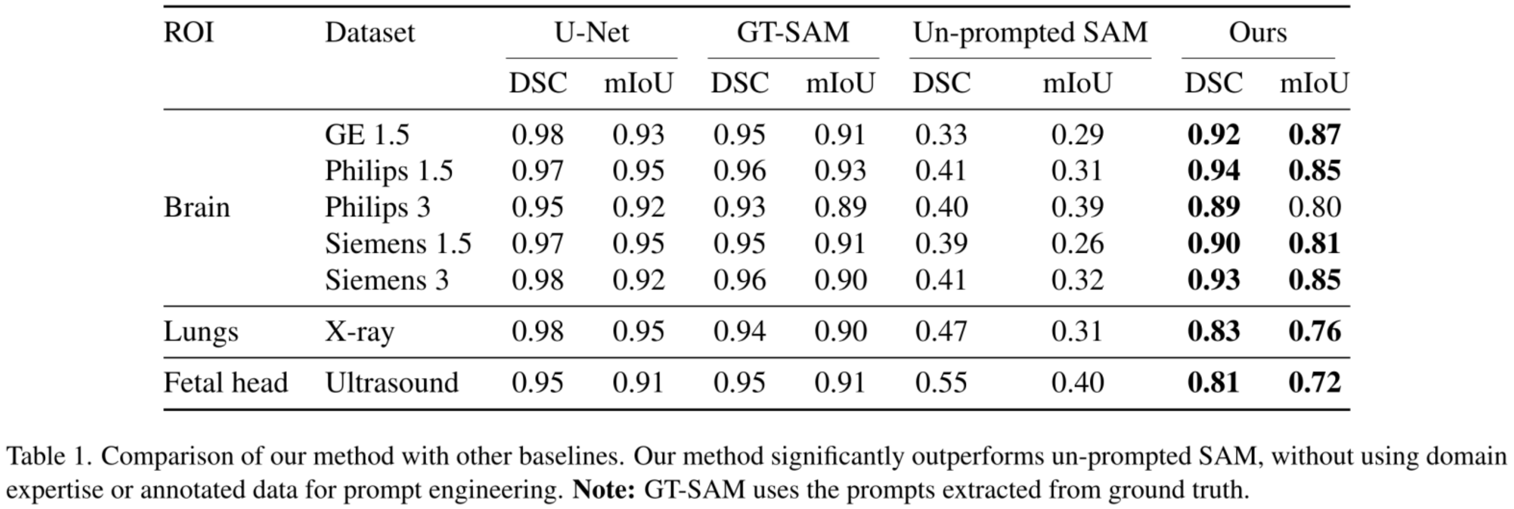
Overview of SaLIP

Another major contribution is that **this unified framework is fully adapted at test time for zero-shot organ segmentation, meaning no training is involved**. Traditional methods of adapting SAM to medical imaging involve finetuning or transfer learning, which require substantial computational resources and large datasets. A common challenge in medical imaging is the scarcity of available data. A training-free approach bypasses the need for annotated data or human experts and resolves privacy concerns and resource limitations in medical imaging. Harnessing the capabilities of LLMs also eliminates the need for engineering.

As our interview draws to a close, Sidra shares some highlights from her recent experience at the **International Symposium on Biomedical Imaging (ISBI) in Athens**. *"I've been to a lot of conferences, but there were a few things I experienced for the first time at ISBI,"* she says. *"The highlight for me was the lunch with leaders, where you sit at a table with a specific leader and speak one-on-one about your career. As I'm nearing the completion of my PhD, I need professional guidance from someone in this field."*



Qualititative Results. DSC refers to Disce Score Coefficient



ROI	Dataset	U-Net		GT-SAM		Un-prompted SAM		Ours	
		DSC	mIoU	DSC	mIoU	DSC	mIoU	DSC	mIoU
Brain	GE 1.5	0.98	0.93	0.95	0.91	0.33	0.29	**0.92**	**0.87**
	Philips 1.5	0.97	0.95	0.96	0.93	0.41	0.31	**0.94**	**0.85**
	Philips 3	0.95	0.92	0.93	0.89	0.40	0.39	**0.89**	0.80
	Siemens 1.5	0.97	0.95	0.95	0.91	0.39	0.26	**0.90**	**0.81**
	Siemens 3	0.98	0.92	0.96	0.90	0.41	0.32	**0.93**	**0.85**
Lungs	X-ray	0.98	0.95	0.94	0.90	0.47	0.31	**0.83**	**0.76**
Fetal head	Ultrasound	0.95	0.91	0.95	0.91	0.55	0.40	**0.81**	**0.72**

Table 1. Comparison of our method with other baselines. Our method significantly outperforms un-prompted SAM, without using domain expertise or annotated data for prompt engineering. **Note:** GT-SAM uses the prompts extracted from ground truth.

Sidra's chosen leader was **Amir Amini**, a distinguished professor at the University of Louisville with a wealth of experience in **biomedical imaging and electrical engineering**. *"His list of achievements is very long,"* she points out. *"It was a very useful activity to sit in front of such an accomplished person and talk about your personal problems."*

While she was able to attend ISBI, Sidra expressed her frustration at not being able to come to Seattle for CVPR due to visa issues, having been looking forward to presenting her work in person. Nevertheless, she remains optimistic when we turn to talk of the future. *"As I'm in the last year of my PhD and my last semester is about to start, I'm looking for positions,"* she reveals. *"I hope that by sharing this, people might reach out!"*

One thing is clear: Sidra is poised to make a significant contribution to her field. For those seeking a dedicated and insightful researcher, she is an exceptional candidate ready to embark on the next stage of her career!

Did you enjoy this BEST OF CVPR 2024?

37 pages about CVPR in Seattle, 10 days ago!

Will you miss your friends and the brilliant tech?

Do you want to feel every month at CVPR?

Subscribe for free to Computer Vision News!

With us, it's CVPR every month!

Congrats, Doctor Philipp!

Philipp Klumpp completed his PhD just a few weeks ago. He worked with the team for speech processing and understanding of the Pattern Recognition Lab at FAU Erlangen-Nürnberg.

Under the supervision of Elmar Nöth, his research focused on the automated analysis of pathological speech and language using modern ML techniques.

Philipp is now working as a Data Scientist for DATEV.

Speech and language technology has become ubiquitous and significantly more powerful in recent years. These advancements came with highly complex models which demanded enormous amounts of training data. In the domain of pathological speech, such amounts are unheard of. For a robust modern speech recognizer, we are talking about many thousands of hours of transcribed audio samples from at least some 10.000 different speakers. Good luck trying to adopt this for pathological speech!

In his thesis, Philipp proposed a solution to this problem which may appear counterintuitive at first: No pathological data was used during the optimization of any of the large models. Instead, his algorithms rely exclusively on off-the-shelf speech recognition datasets which had been collected from healthy speakers. But how could such a model help to analyze pathological speech? This is where phonetics come into play, the science of speech production, transmission and perception. It helps explain how and why the speech from patients of a particular medical condition deviates from the healthy reference.

Not only could the presented approach solve the problem of data scarcity in the medical domain, but it also yields very explainable outputs which are much easier for a clinical expert to understand and interpret. For example, it was possible to show that Parkinson’s disease (PD) patients lose the ability to reliably produce speech sounds that require a high degree of muscle tension. Compared to a black box model which learned from pathological data how to estimate an articulation score between 0 and 1, an estimate of the speech sound inventory which a patient can still produce reliably carries much more information for the clinical expert and allows them to draw conclusions on their own.

The proposed methods were not limited to the analysis of isolated speech sounds but could also be used to model the articulatory mobility inside the vocal tract. It is well documented that PD causes muscle rigidity, and this also affects the tongue, a complex muscular structure which plays a major role in the articulation of vowels. Philipp proposed a method to estimate the articulatory space – or, in simple terms, the maneuverability of the tongue – inside the vocal tract of an individual. It strongly correlates with the speech quality annotated by a speech and language therapist, as shown in Figure 1.

The model used for the analysis of pathological speech is available here.

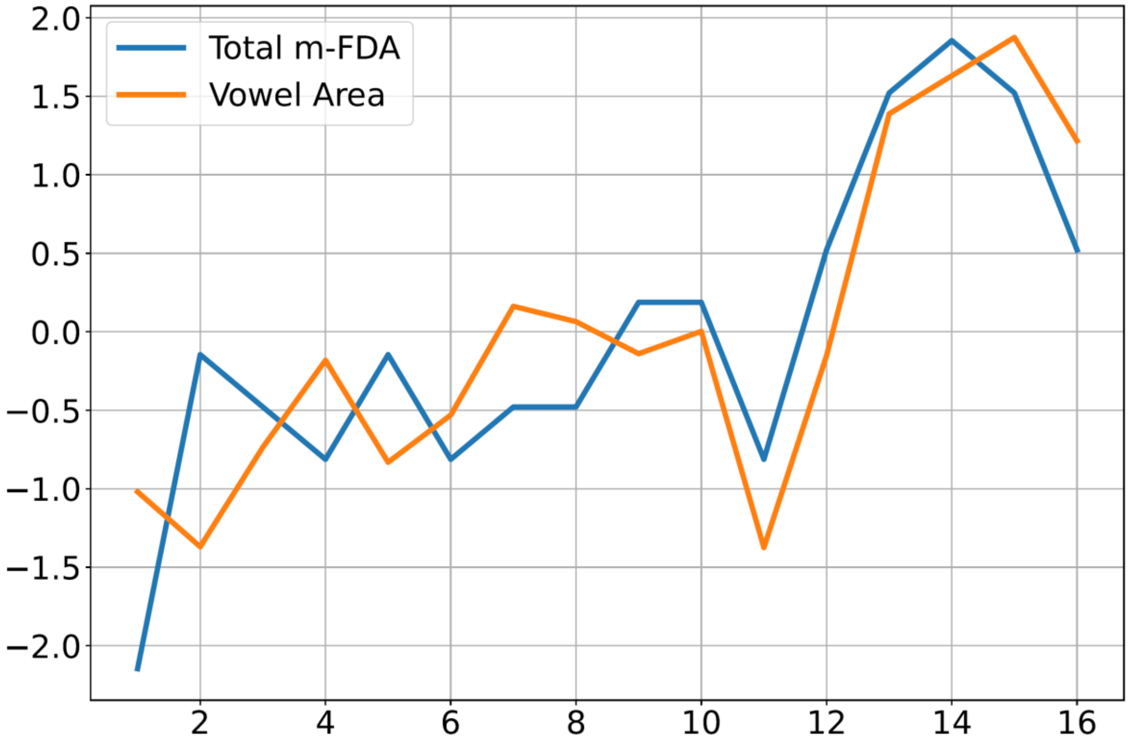


Figure 1: Over the course of 16 recording sessions with a single PD patient, the blue line indicates the (normalized) expert rating of speech impairment (higher score reflects stronger impairment), while the orange line represents the reduction of available articulatory space.

Cracking the Code: Multiclass Classification and Microenvironment Analysis of Brain Tumors Focus on Myeloid Cells

Monika Pytlarz is a PhD student at Sano Centre for Computational Medicine, Kraków, Poland. With a background in clinical diagnostic imaging and bioinformatics, she focuses on integrating deep learning into neurobiology and immunohistochemistry to enhance digital pathology and radiology tools.

Monika is committed to translating computational innovations into practical applications (e.g., by working on the startup). Her interest in this field deepened thanks to the MICCAI BrainLes workshop, which focused on brain lesions and brain tumors and to which she had the opportunity to contribute.

by Monika Pytlarz

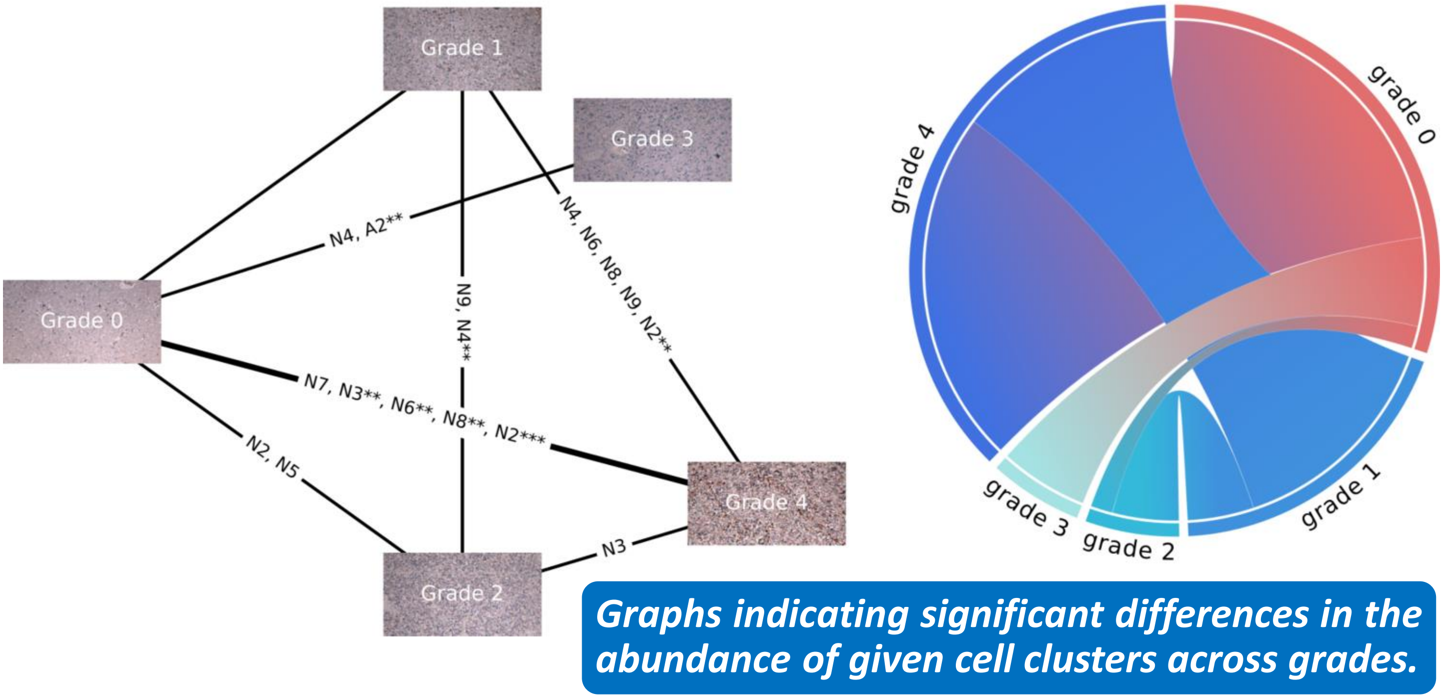
The PhD research project primarily aims **to improve the diagnosis and treatment planning of brain tumors**, particularly gliomas and glioblastomas. These tumors are incredibly challenging to pinpoint and treat due to their **intricate interactions**

within the tumor microenvironment and their notorious **heterogeneity**. Traditional diagnostic methods involve examining slides under a microscope, which is time-consuming and prone to human error. **Even seasoned pathologists can disagree on the same samples**, highlighting the need for precise **computer-aided diagnosis techniques**.

The efforts focused on developing deep learning models that can **automatically grade gliomas by analyzing tissue microarrays with human leukocyte antigen staining**. Recent results presented at **SPIE Medical Imaging** and published in the [Journal of Imaging Informatics in Medicine]() investigate the strategies for glioma multiclass classification - both in supervised and weakly supervised manner with single-cell analysis.

One challenge in this research was the limited data available for training the models. The issue was tackled by augmenting underrepresented classes and evaluating models using k-fold stratified cross-validation. The DenseNet121 architecture with prior image preprocessing outperformed the baseline model increasing accuracy by 9%.

The **study goes beyond identifying the WHO grade (1, 2, 3, 4, or grade "0")** to which the sample should be assigned. The protocol of weakly supervised deep learning with single-cell analysis was applied to the dataset to **learn, discover, and quantify the cell phenotypic neighborhoods across grades**. The model identified **two critical cell clusters, called neighborhood N2 and N4, which showed significant differences in the abundance of cells, most likely microglia and brain**



Graphs indicating significant differences in the abundance of given cell clusters across grades.

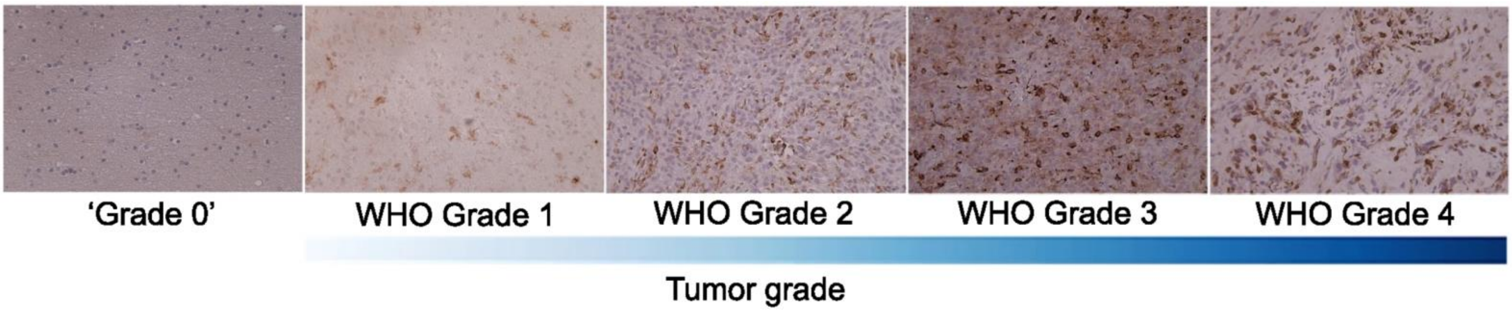
macrophages, across grades. These clusters serve as promising **indicators of tumor progression**, aligning with biological reports on **the pivotal role of myeloid cells in tumor growth** and immunosuppression.

> ***The protocol of weakly supervised deep learning with single-cell analysis was applied to the dataset to learn, discover, and quantify the cell phenotypic neighborhoods across grades.***

Understanding why treatments fail can sometimes feel like solving a mystery novel. In the complex world of brain tumors, **it's often the microscopic details that hold the biggest clues**. Treatments can fail when the cellular interplay is misread. Accurate stratification of brain tumors can enhance personalized treatment strategies.

Looking ahead, the research **interest shifts towards multimodal analysis of glioblastoma subtypes and temporal changes in tumor cell subpopulations**.

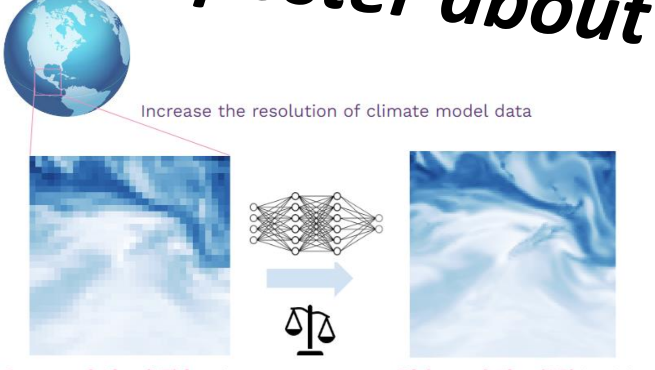
The project's aim is to address critical challenges in glioma research, with potential implications for improving clinical outcomes for patients worldwide. Additionally, the study titled '**Style Transfer between Microscopy and Magnetic Resonance Imaging via Generative Adversarial Network** in small sample size settings' (ICIP 2023) is further explored in the context of high-dimensional brain imaging translation. Stay tuned as this work pieces together the ever-changing puzzle of brain tumor diagnosis and treatment, bringing us closer to a future where beating gliomas is more science and less guesswork!



Examples of annotated glioma tissue microarray cores with human leukocyte antigen staining.

Who is this climber?

And what was her ICLR poster about?


Increase the resolution of climate model data
Low-resolution (LR) input
High-resolution (HR) target

Hard-Constrained Deep Learning for Climate Downscaling

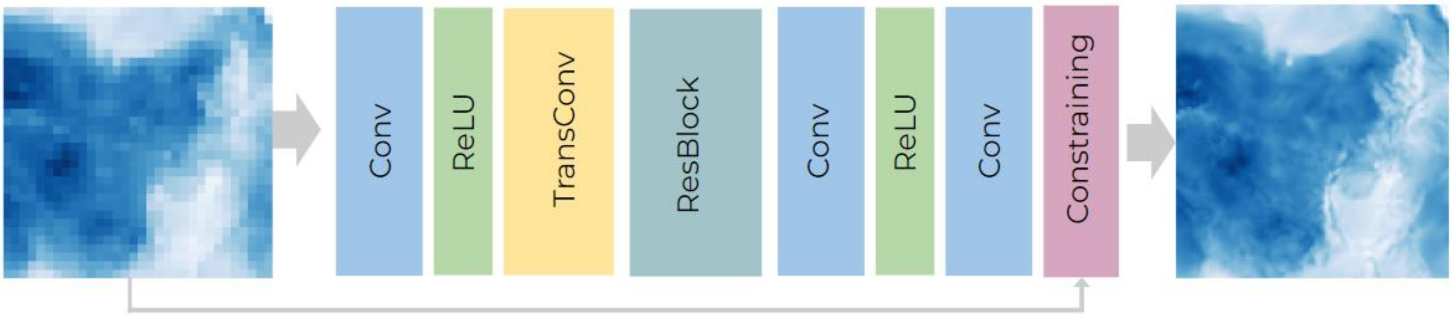
Paula Harder is about to finish her PhD and is an incoming postdoc at Mila Quebec AI Institute.

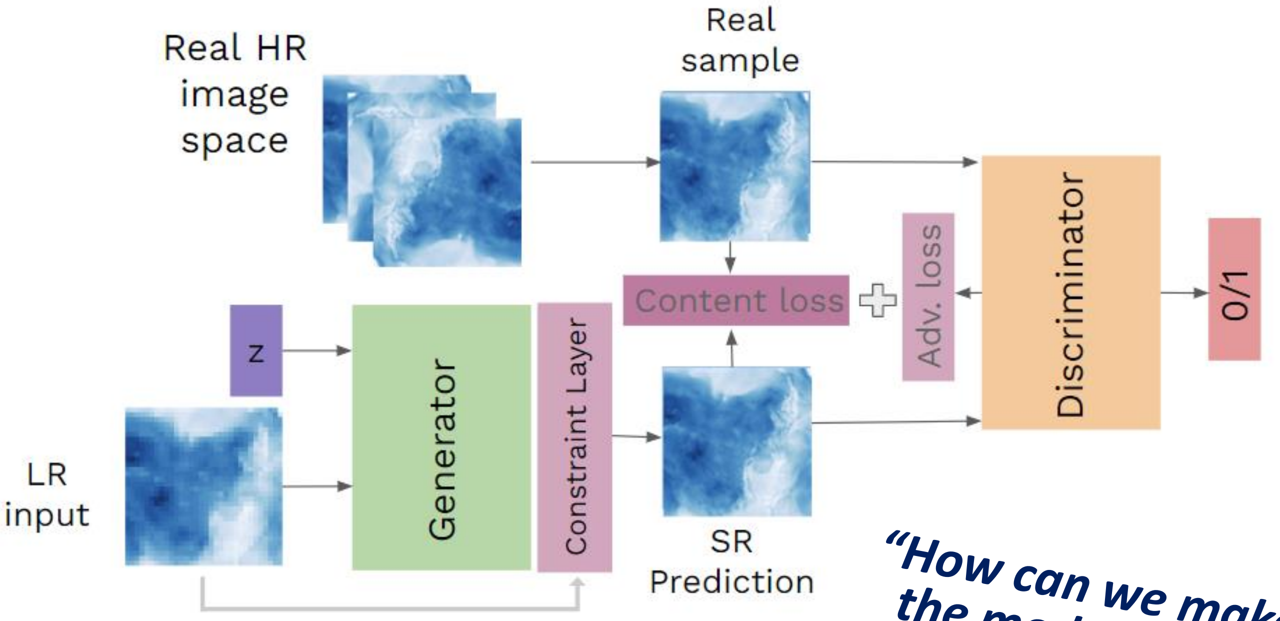
Fresh from her ICLR 2024 poster session in Vienna, she speaks to us about her paper, which uses deep learning to advance climate science.

This article should have been published last month in our BEST OF ICLR special edition. Apologies to awesome Paula and to readers.

Climate downscaling, a technique similar to super-resolution in computer vision, has emerged as an innovative new tool for enhancing the resolution of climate data. Paula and the team behind this work have adapted deep learning architectures to augment climate models, enabling more detailed spatial predictions and incorporating constraints to ensure physical plausibility within the neural networks.

The motivation for this work stems from the limitations of traditional climate modeling, which often provides coarse long-term predictions over large regions, offering a single temperature value for sizable areas. *"Maybe we want to know what's going to happen here in Vienna,"* Paula points out. *"That's where we need higher-resolution climate data.* ***The numerical models are just too***



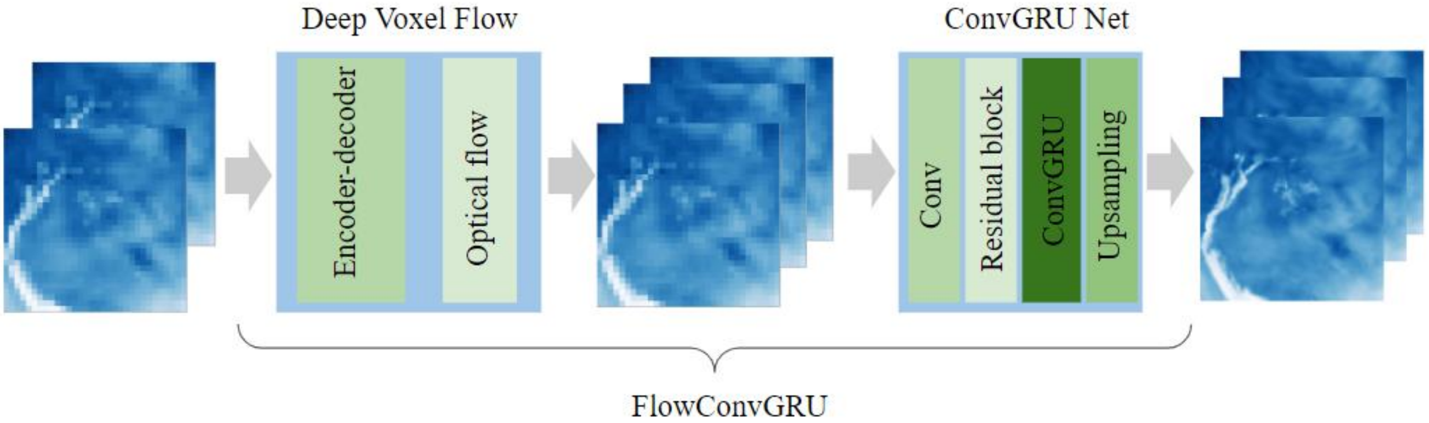


computationally expensive, and that's where deep learning methods from computer vision can help. *The models applied to super-resolve images can also increase the resolution of climate model data!"*

> **"How can we make the models more robust under distribution shifts?"**

Previous attempts at downscaling existed within the climate community but were limited in scope. **Introducing deep learning architectures from computer vision significantly advances the downscaling task and allows for integrating physical constraints**. *"Sometimes we predict rain, for example, but then the neural networks violate the simple physical laws, like conservation of water mass,"* Paula explains. *"We developed a model to ensure that they're physically correct, so we have climate predictions that make sense. Sometimes the models even predict negative masses, and we make sure that doesn't happen."*

Paula's novel approach introduces a new layer at the end of a neural network to act as a constraint, slightly rescaling the output to make it physically accurate. The method can be applied to standard architectures, such as CNNs, GANs, transformer models, and normalizing flows, which are great at super-resolving data. *"The first thing people try is to add another loss term to the loss function,"* she reveals. *"That's very common, but it worked terribly. We did a lot of tuning and training, but it didn't work out. That brought us to the new idea of* ***including this final constraint layer, which worked very well****. After some struggle with the so-called soft constraint, where you have a regularization term in your loss function, we then developed this constraint layer."*

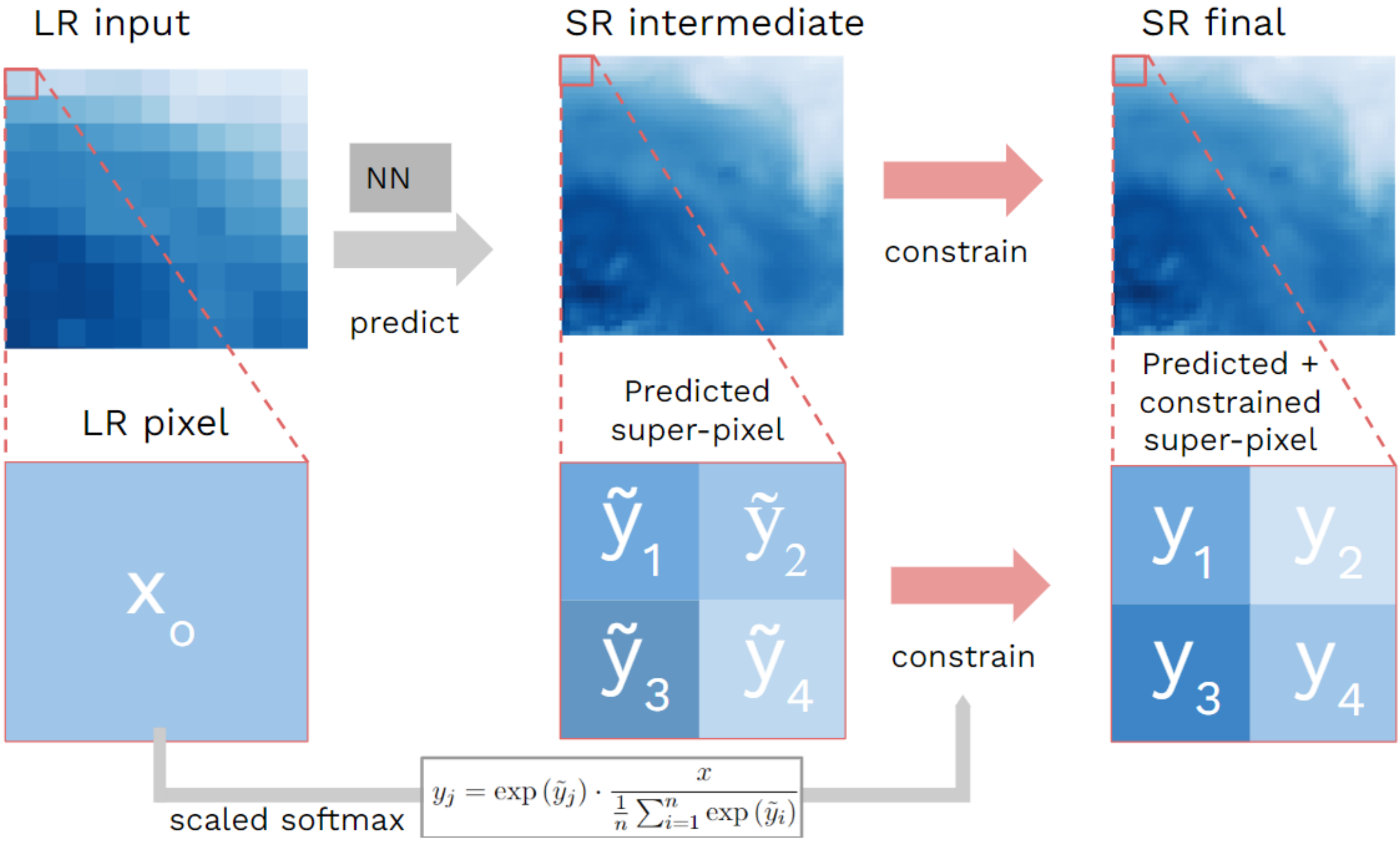


This solution drew inspiration from **the softmax layer used in multi-class classification tasks**, which enforces probabilities that sum up to one. Paula slightly changed this to force it to sum up to the mass. By adapting and refining this, she created a tailored solution that addresses the unique complexities of the climate science problem.

> ***The numerical models are just too computationally expensive, and that's where deep learning methods from computer vision can help!***

Something Paula says the team were insistent on from the very start was aligning the research with the practical needs of climate scientists. **Collaborating closely with domain experts, including climate scientists at IBM, was crucial to understanding the specific requirements and desired outputs within the climate science community**. "*We had some ideas in mind, but as machine learning people, we didn't know what the community exactly needs,*" she recalls. "*It was really important to have people from that domain with domain knowledge be like, okay, these are the kinds of outputs we want to have; these are the kinds of things we care about. We want the model to be something people can use for real-world climate forecasting. We just brought it out, but we're talking with them, and people are very interested in starting to use it. We really hope that it can be built into operational climate and weather forecasts.*"

This work originally appeared in **the Journal of Machine Learning Research** earlier this year, which had a track for ICLR. Regarding the next steps, **Paula envisions broader applications within and beyond**



climate science. She already found that the method they developed could be used in another project involving aerosol modeling with deep learning, where she replaced part of a climate model with a neural network, aiming to conserve the aerosol mass within each time step. The same constraint layer helped solve that problem, and she anticipates that many other problems could use it.

Will Paula be tackling some of these problems herself? *"It's going to be a mix,"* she responds. *"I'm still involved in some projects that use that, but I also know some people who took the idea and are working on it themselves, and I'm very happy about that. I just hope that it can help the community get closer to the goal of having good climate forecasts for the future."*

Paula is keen to continue advancing deep learning models for applications in climate and weather modeling and super-resolution downscaling. In particular, how models can be improved to perform better when applied to new locations. ***"Sometimes, we have great data to train on in one location of the world, and then we apply the deep learning model in a different location, and it completely fails,"*** she tells us. ***"How can we make the models more robust under distribution shifts?*** *I don't have a solution yet, but it's something I'm really curious to explore!"*

Hubert D. Zając holds a PhD in Computer Science from the University of Copenhagen.

His PhD focused on making clinical AI useful in practice.

Now, Hubert is working as a Postdoc in Human-Centered AI at the University of Copenhagen developing a holistic approach to AI innovation encompassing the entirety of the development process: from data creation to onboarding.

Despite recent advancements, e.g., in pathology detection in X-rays, MRIs and CTs, AI-based systems remain notably absent in current clinical practice. One key factor contributing to this gap is the prevalent technology-centric approach to AI innovation, which often results in the limited clinical usefulness of AI-based support systems. To address this issue I conducted research within the framework of an interdisciplinary project aimed at creating a chest X-ray support tool for radiologists in Denmark and Kenya.

Based on a systematic literature review (Zając et al. 2023, TOCHI), I find that challenges afflicting the realization of clinical AI in practice stem not from a single issue but rather from sociotechnical interdependencies present when introducing AI into a clinical context. For example, how a clinical position affects the need for explanations. I argue that addressing these challenges necessitates close collaboration among stakeholders with expertise in HCI, AI, and healthcare throughout the innovation processes.

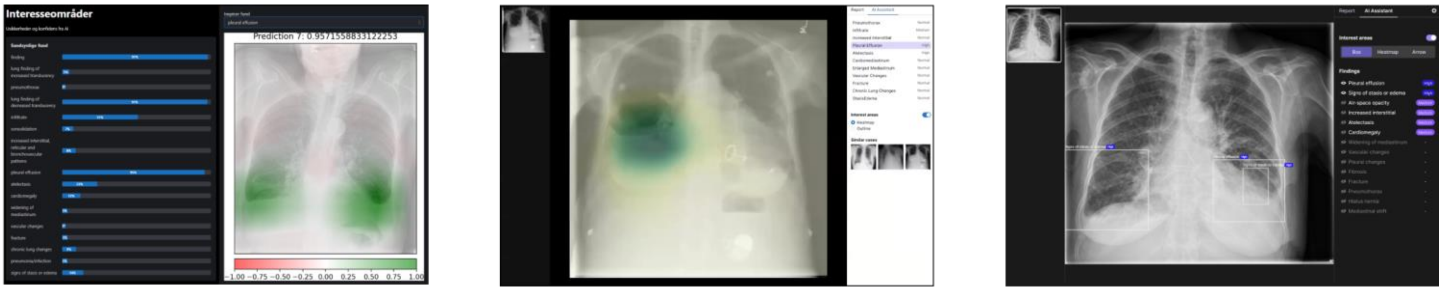


Figure 1: Evolution of the developed chest X-ray support system.

Further, I underscore the importance of attending to the pre-labelling phase in dataset creation (Zając & Avlona et al. 2023, AIES ‘23). Particularly, I highlight how external and internal factors: regulatory constraints, the context of creation and use, commercial and operational pressures, epistemic differences, and limits of labelling condition the type of data that could be collected, the purpose for which it could be used, and the design of the labels. These fundamental decisions have consequences for shaping the design space of future AI-based systems that use such datasets.

Next, I propose five visions for AI support grounded in practical challenges of chest X-ray practice faced across clinical contexts. The visions include distributing examinations by user's expertise, detecting medical emergencies, providing decision support on subtle and difficult cases, measuring visual features and comparing changes across historical examinations, and double-checking reports against radiographs for missed or misinterpreted findings. These visions transcend functionalities traditionally emerging from technology-centred innovation processes and offer nuanced insights into potential AI applications in radiology.

Finally, I delineate how AI-based systems should be configured both before and in use to realize previous visions in practice (Zając et al. 2024, DIS ‘24). The purpose of the configuration is to align the technical dimensions of AI-based systems with clinical needs that depend on social dimensions of clinical practice. The social dimensions span medical knowledge, clinic type, user expertise level, patient context, and user situation. The technical dimensions of AI comprise medical focus, functionality, decision threshold, and explainability methods. By ensuring alignment between these dimensions, AI-based systems can deliver value in concrete situations for concrete medical professionals in clinical practice.

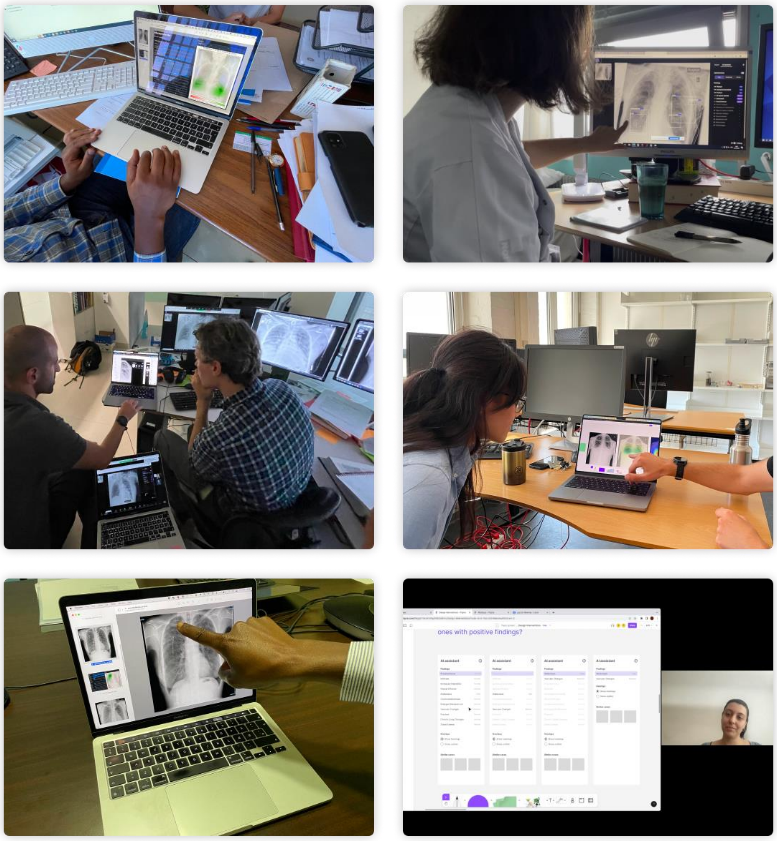
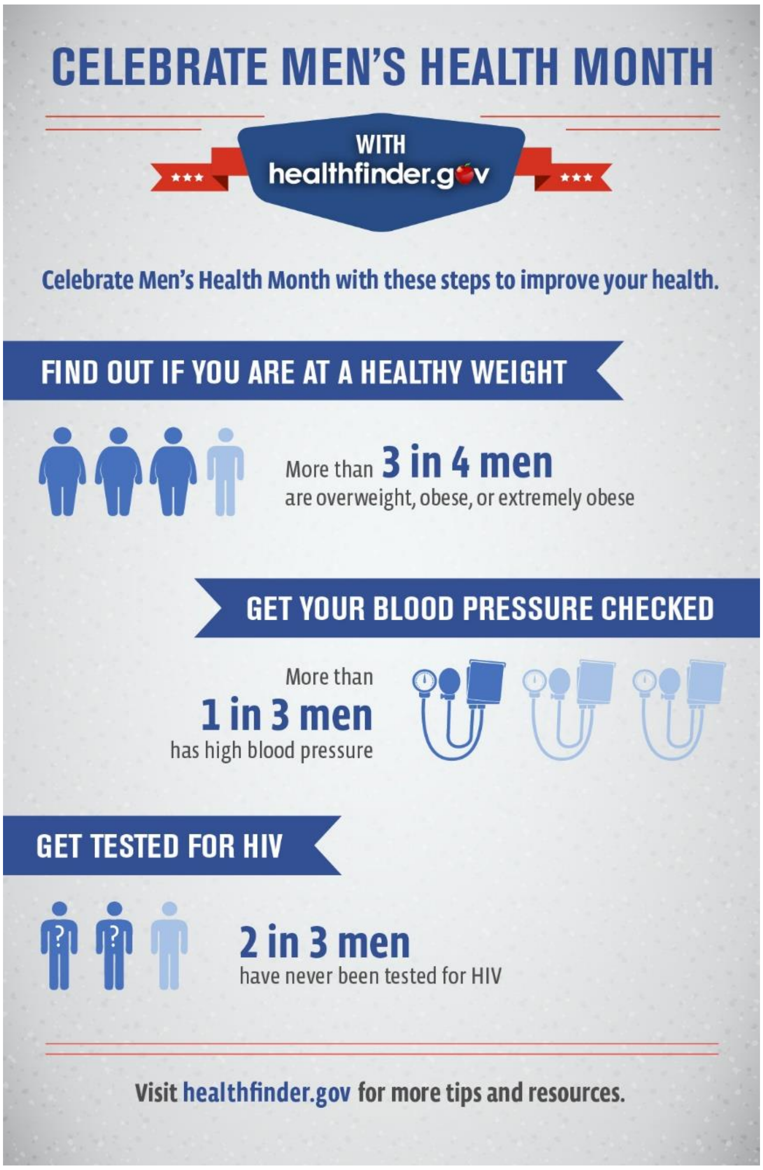


Figure 2: A collage of design-focused fieldwork in Denmark and Kenya.



CELEBRATE MEN'S HEALTH MONTH

WITH healthfinder.gov

Celebrate Men's Health Month with these steps to improve your health.

FIND OUT IF YOU ARE AT A HEALTHY WEIGHT

More than **3 in 4 men** are overweight, obese, or extremely obese

GET YOUR BLOOD PRESSURE CHECKED

More than **1 in 3 men** has high blood pressure

GET TESTED FOR HIV

2 in 3 men have never been tested for HIV

Visit healthfinder.gov for more tips and resources.

Last month was Men's Health Month in the US, a chance to raise awareness about healthcare issues affecting men, with prostate health high on the agenda.

In light of this, RSIP Vision's Arik Rond speaks to senior urologist MD Arnon Lavi about the latest developments in prostate treatment with AI.

Benign prostatic hyperplasia (BPH) is a non-cancerous enlargement of the prostate gland that can cause urinary problems in men. While **open surgery** is still a standard treatment for BPH in certain parts of the world, particularly for larger prostates, modern medical advancements have led to **less invasive endoscopic procedures**. *"Today, we have less invasive and more patient-friendly options, enhancing patient satisfaction and recovery,"* Arnon tells us. *"There are a lot of emerging technologies and energies being used surgically, such as **water vapor, aquablation, GreenLight laser, the traditional transurethral resection of the prostate (TURP), and holmium laser enucleation of the prostate (HoLEP)**."*

Despite these advances, minimizing the negative impacts of surgery, such as **urethral stricture** and the need for **prolonged catheterization**, remains challenging. Intraoperatively, surgeons must carefully manage bleeding risks, particularly for those patients taking anticoagulant or antiplatelet therapy, while accurately measuring prostate size and identifying anatomical features. Additionally, selecting the most appropriate surgical energies or techniques based on individual patient characteristics adds another layer of difficulty. Experience plays a crucial role. *"With every technique, the higher caseload you do, the better the outcomes,"* Arnon points out. *"That's something across the surgical field, and I think it's true for benign prostate surgery."*

Artificial Intelligence has the potential to revolutionize prostate treatment by addressing these challenges. AI enhances preoperative planning by predicting bleeding risks **and identifying anatomical features such as big median lobes, high bladder necks, and proximity to the urethral orifices that are difficult to assess with conventional imaging**. *"Better patient selection is the main challenge for benign prostatic surgery,"* Arnon notes. *"AI can help us select and do that finetuning to choose the best treatment, energy, or technique for each patient with his unique characteristics."*

For **localized prostate cancer**, treatment has traditionally involved either **surgery or radiation therapy combined with androgen deprivation therapy**. *"As far as we know, the oncological outcomes are equivalent, but there are certain characteristics or adverse effects that differ between these two treatments,"* Arnon explains. *"When we talk about prostate cancer surgery, the standard of care throughout the Western world today is probably* ***robotic surgery****. However, the fact that there's* ***no tactile feedback with the available robotic consoles*** *is a major drawback – although the newer systems are trying to incorporate these things."*

Studies have shown that around 5% of men experience **persistent urinary incontinence** one year after undergoing prostate cancer surgery, while incidence rates of erectile dysfunction vary from 30% to 50%. Predicting these outcomes is complex, and AI tools to better assess individual patient risk are still under development. Surgeons face

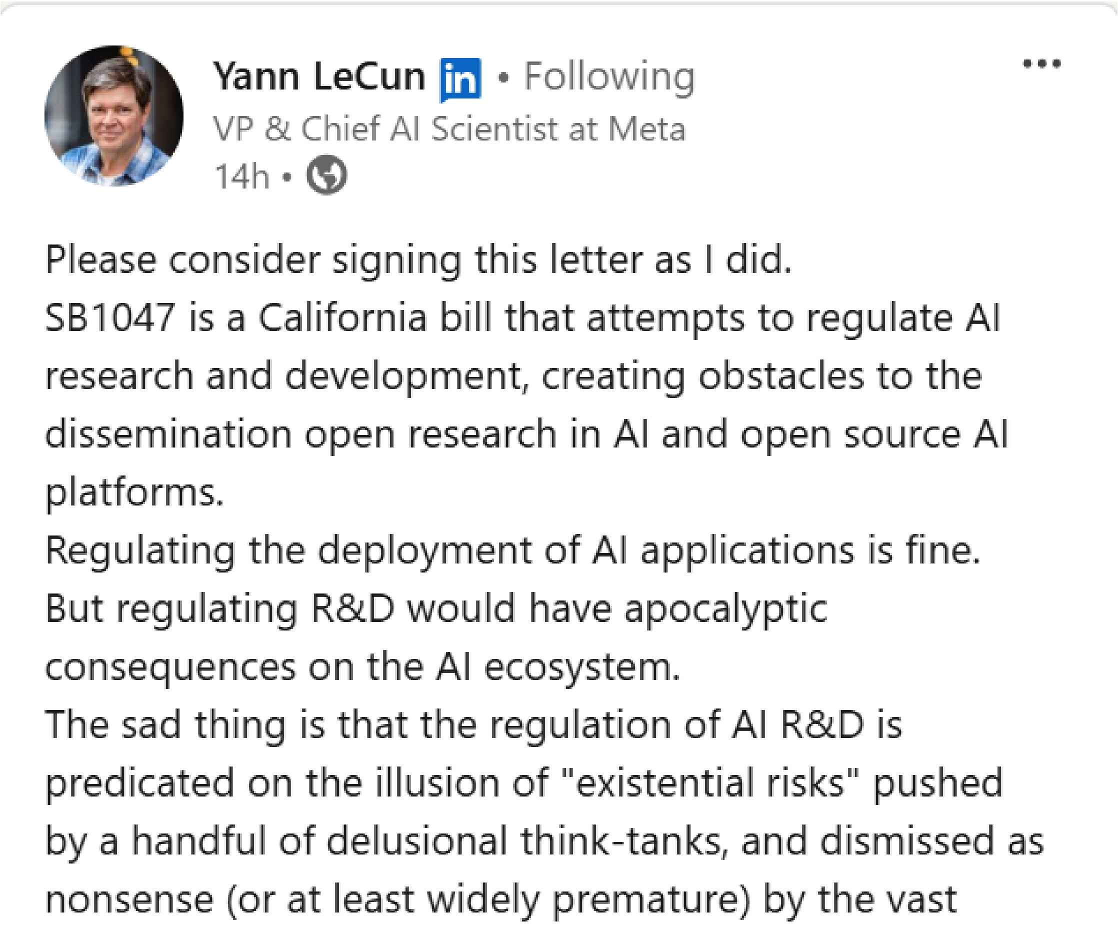
the task of **preserving the neurovascular bundle**, which is crucial for erectile function, and the **pelvic floor**, which is vital for urinary control. However, anticipating the extent of surgery or necessary margins is not always possible. *"When we counsel patients before surgery, we can tell them the general numbers, but when the patient in front of you is hesitating or trying to decide, these numbers don't say a lot,"* Arnon remarks. *"The main challenge is understanding which patients we can operate on with good functional outcomes. We don't yet know that in a majority of cases."*

Today, **multiparametric MRI** makes it easier to locate, categorize, and treat the focus of the cancer rather than the whole prostate. **Focal therapies like high-intensity focused ultrasound (HIFU), cryotherapy, and electrical energy** are proving more popular due to their promising functional outcomes. They **reduce damage to surrounding tissues**, and patients face **fewer side effects** than those caused by surgery and radiation. Furthermore, **AI has the potential to enhance the precision and safety of radiation therapy by better localization and minimizing toxicity to adjacent structures**.

As technology evolves, the ability to tailor treatments to individual patient's needs and minimize undesirable outcomes will significantly enhance the quality of prostate healthcare. *"The tools to handle many of these challenges are constantly getting better and better,"* Arnon emphasizes. *"AI is making progress, and **things that weren't an option a few years ago are now becoming good options**. There's still a lot to be done, but the field can be significantly improved with these kinds of tools."*

If you think we could help with your project, contact RSIP Vision today for an informal discussion about what AI can do for your work.

RSIP VISION
Visual Intelligence
for MedTech



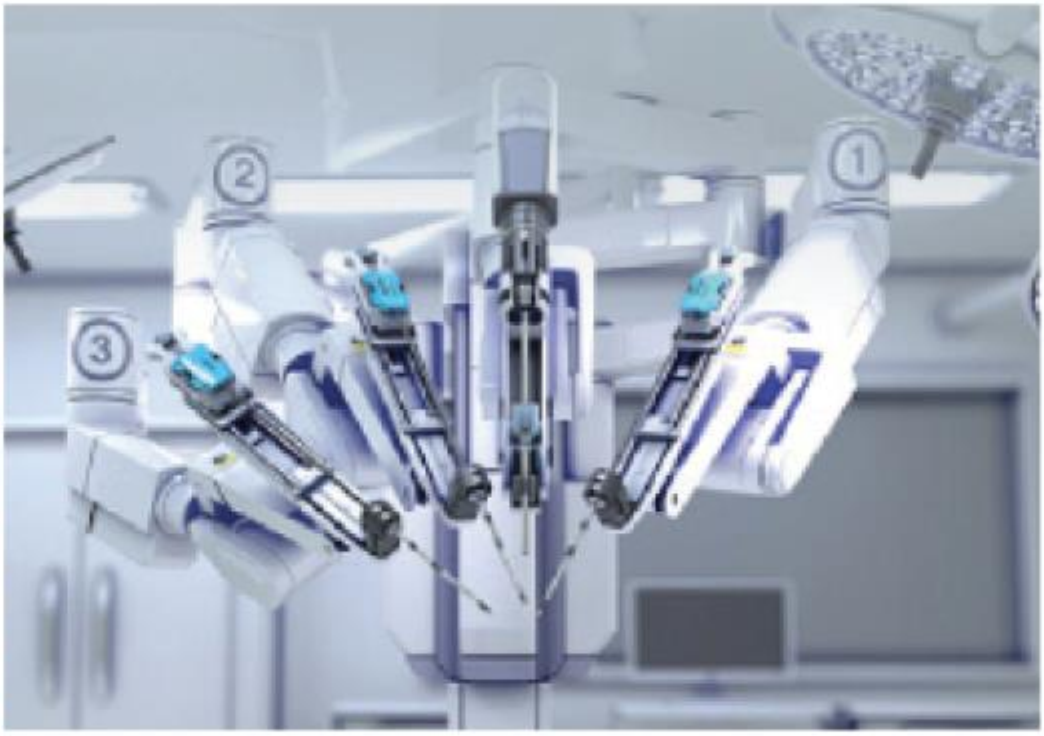
Yann LeCun in • Following
VP & Chief AI Scientist at Meta
14h •

Please consider signing this letter as I did.
SB1047 is a California bill that attempts to regulate AI research and development, creating obstacles to the dissemination open research in AI and open source AI platforms.
Regulating the deployment of AI applications is fine.
But regulating R&D would have apocalyptic consequences on the AI ecosystem.
The sad thing is that the regulation of AI R&D is predicated on the illusion of "existential risks" pushed by a handful of delusional think-tanks, and dismissed as nonsense (or at least widely premature) by the vast majority of researchers and engineers in academia, startups, larger companies, and investment firms.

https://stopsb1047.com/

An interesting post by **Yann LeCun**, a couple of days ago on LinkedIn. Worth having a look.

If you missed my latest interview with him a few months ago, read it here!



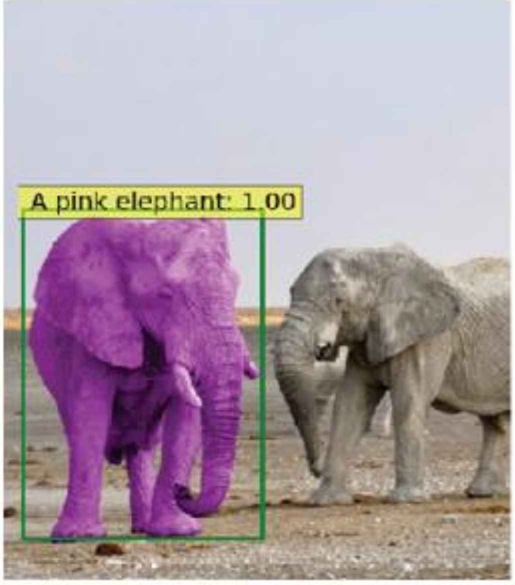
IMPROVE YOUR VISION WITH

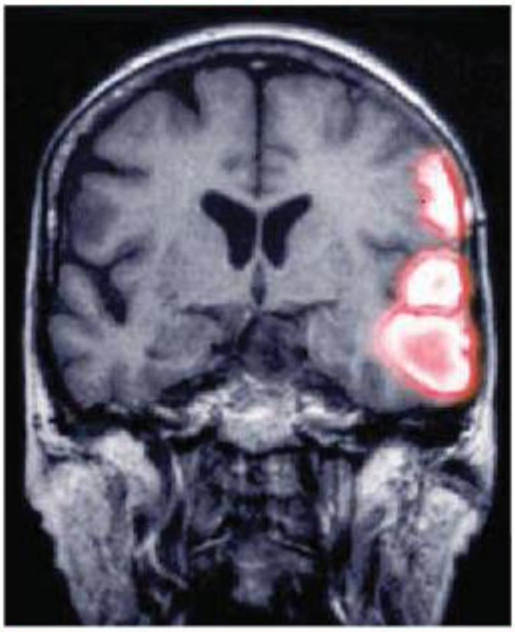
Computer Vision News

SUBSCRIBE

to the magazine of the algorithm community

and get also the new supplement Medical Imaging News!





RSIP VISION