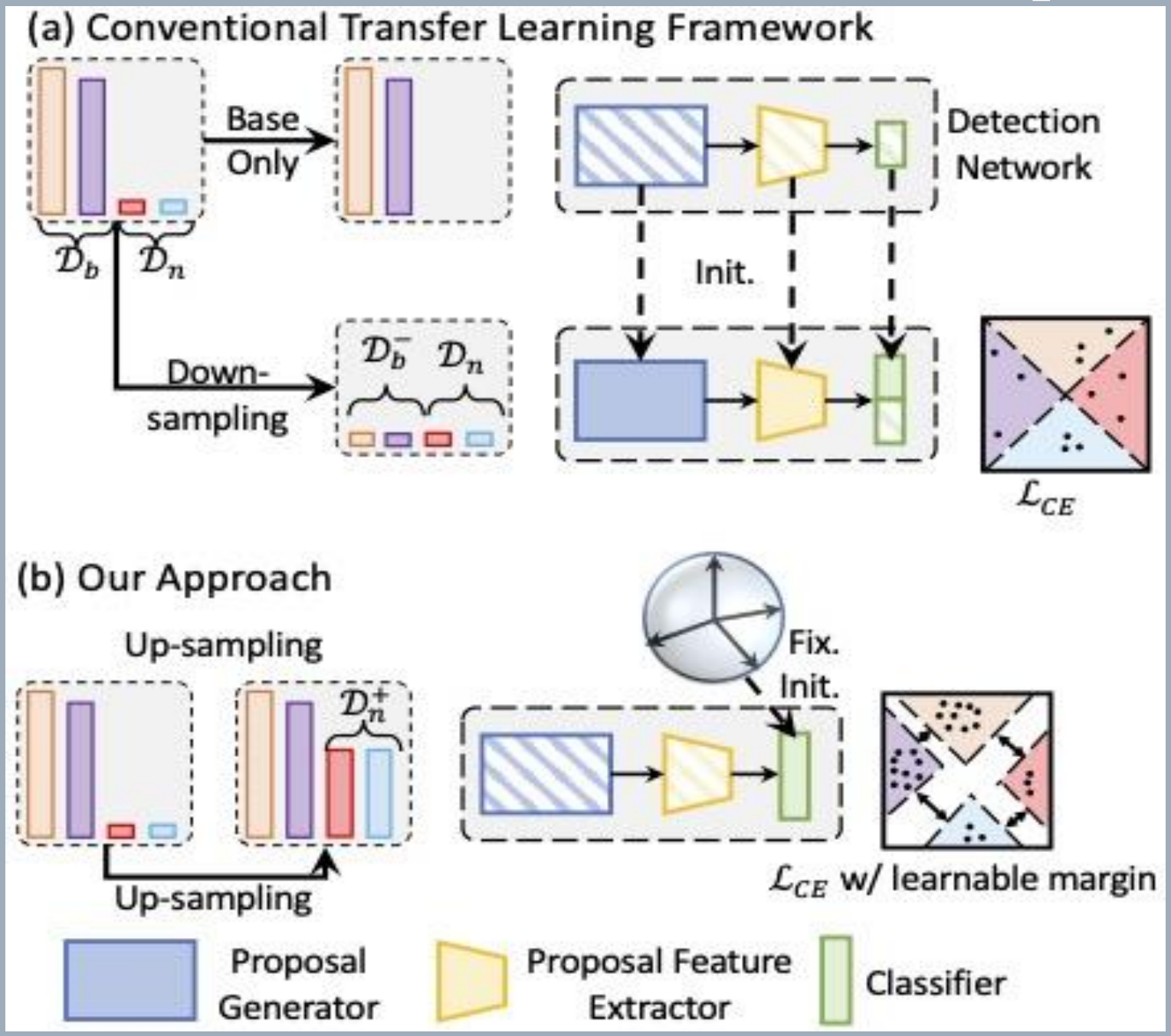
CVPR DAILY 2023

Computer Vision and Pattern Recognition

Tuesday



In cooperation with

Computer Vision News

The Magazine of The Algorithm community

A publication by

RSIP VISION

For today, Tuesday 20

"Hello! I am Maria Sofia, a PhD candidate at Sapienza University of Rome in the Department of Computer, Control, and Management Engineering, and a member of the RSTless group. My research focuses on the theoretical foundations and practical applications of machine learning.

Currently, I am actively investigating the challenging problem of learning in the presence of noisy labels. This research aims to develop robust algorithms that effectively mitigate the impact of label noise during the training of machine learning models."

*"This is the first time I'm attending a big conference in person, and I am thrilled to experience it at **CVPR**. The prospect of being surrounded by experts in computer vision and deep learning, in general, fills me with **enthusiasm**. I look forward to engaging in **stimulating face-to-face discussions, witnessing live presentations, and immersing myself in the intellectually inspiring atmosphere of knowledge sharing**. CVPR offers a valuable opportunity to encounter new ideas and perspectives and be a part of a dynamic community that shares **my passion for deep learning!**"*

Maria Sofia forgot to tell you that she's also presenting her poster #327 today, in the morning session: Leveraging Inter-Rater Agreement for Classification in the Presence of Noisy Labels.

Maria Sofia's picks of the day:

Highlights

(**Award**) DynIBaR: Neural Dynamic Image-Based Rendering
(**Award**) Planning-oriented Autonomous Driving
Tue-PM-035 Canonical Fields: Self-Supervised Learning of Pose-Canonicalized ...
Tue-PM-270 Uncurated Image-Text Datasets: Shedding Light on Demographic Bias

Posters

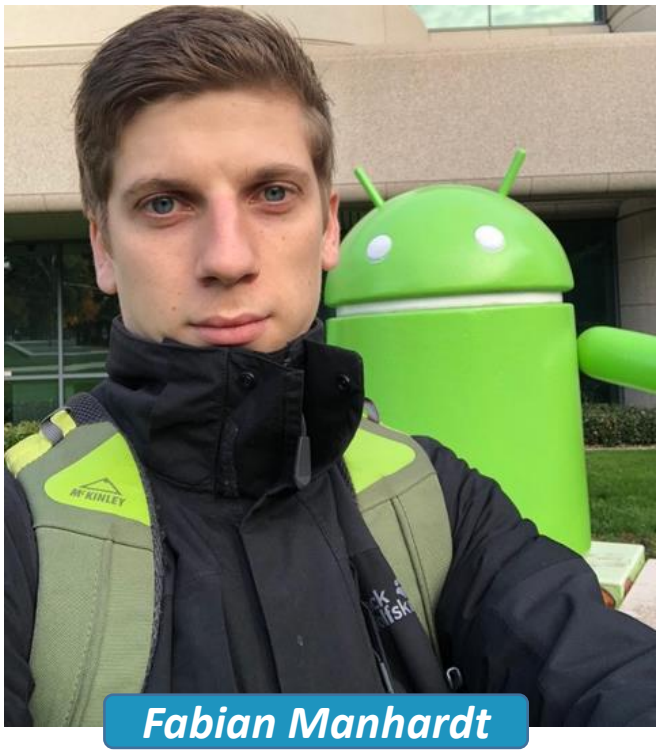
Tue-AM-110 3D Human Keypoints Estimation From Point Clouds in the Wild ...
Tue-AM-325 DivClust: Controlling Diversity in Deep Clustering

SPARF - NeRF from Sparse and Noisy Poses

Prune Truong is a fourth-year PhD student in the Computer Vision Lab at ETH Zurich and an intern in Federico Tombari's team at Google Zurich, where Fabian Manhardt is a Research Scientist.

Prune and Fabian speak to us about their paper proposing a new joint pose-NeRF training strategy designed to be more robust in the real world, which has been accepted as a highlight at CVPR 2023. They speaks to us ahead of their highlight presentation today.

Prune Truong

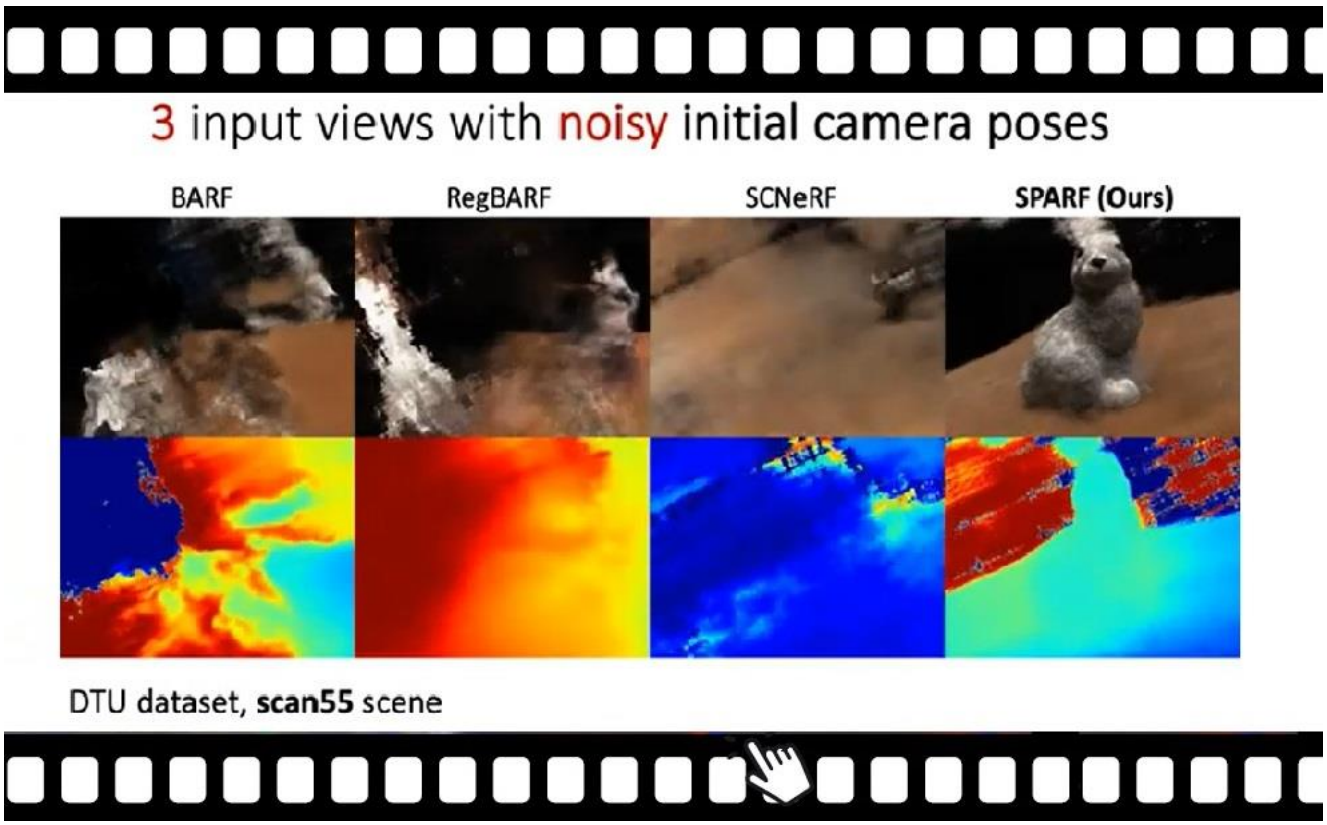


Fabian Manhardt

NeRF (Neural Radiance Fields) is a cutting-edge technology with remarkable potential in generating 3D reconstructions and rendering novel views. NeRF has been shown to work best under two conditions: dense coverage of the 3D space and highly accurate camera poses. This scenario limits its application in the real world, where input views are often sparse, and poses are noisy.

"When you train a NeRF model on only a few images, it will instantly overfit," Prune explains. "You'll have very nice training renderings, RGB will be good, and the photometric loss will be low, but when you look at the depth renderings, you'll see that the model doesn't learn any meaningful geometry. Therefore, it will be really bad if you try to render a novel view."

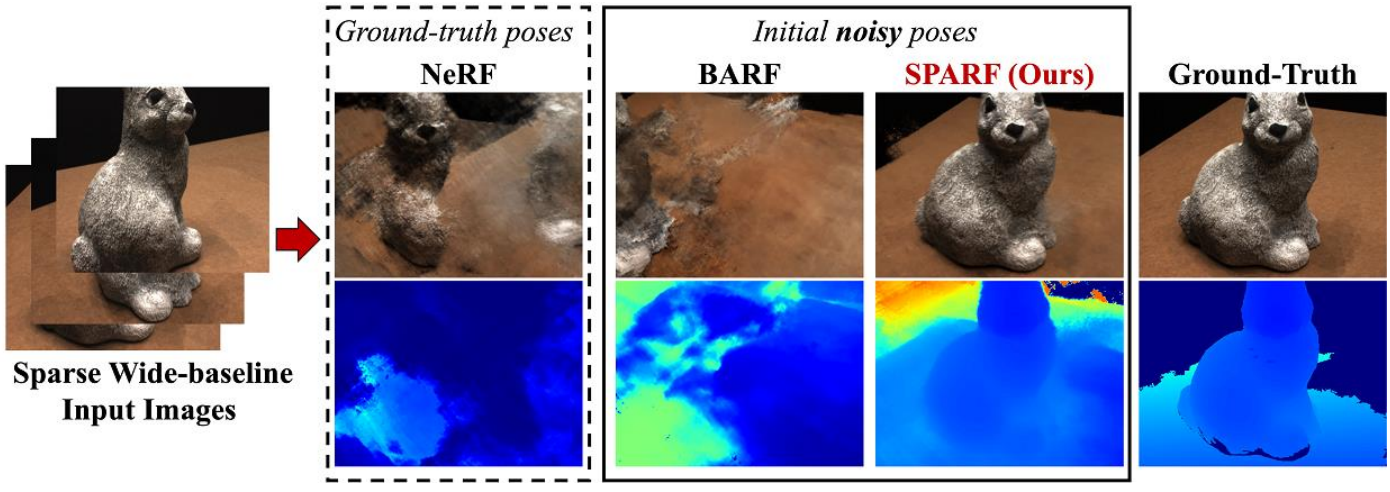
The standard process to estimate per-scene poses is to use a **Structure-from-Motion approach, such as COLMAP**, which works well with many input views. However, with fewer views or an increased baseline between the images, it becomes much more challenging, and the pose estimation results are degraded.

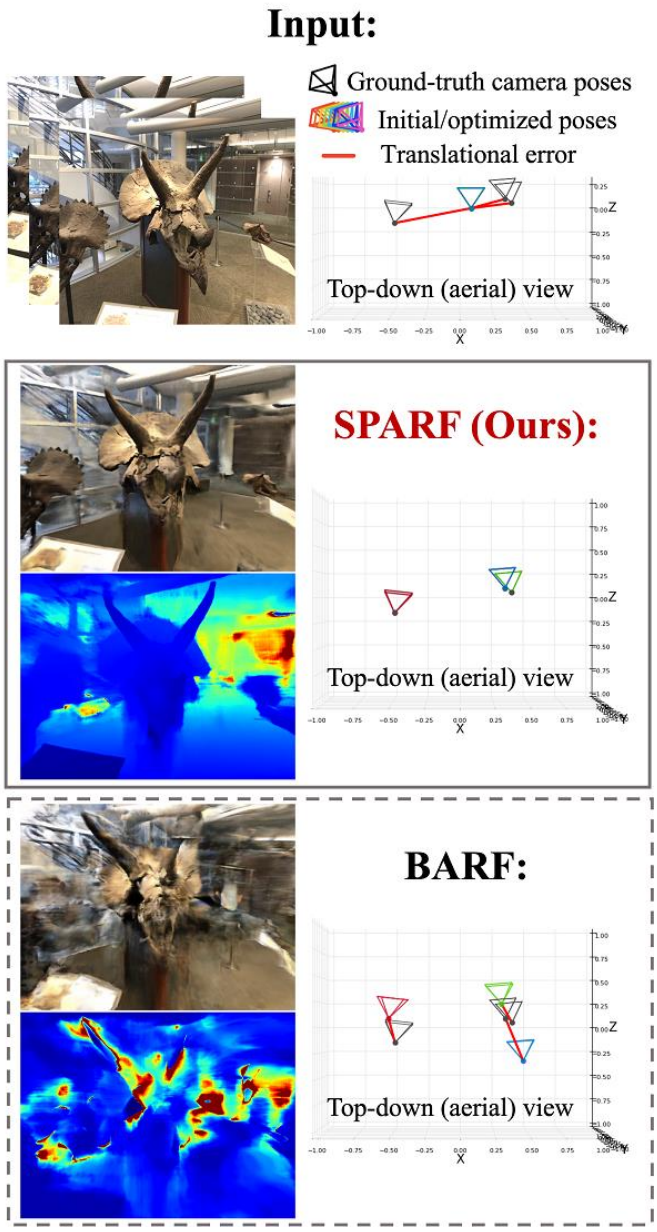


This paper turns the **best-case scenario on its head**, aiming to address the challenge of novel-view synthesis based on a neural field representation using as few as two to three views and noisy poses.

The **potential applications** of this technology are significant. In **robotics**, it could capture 3D reconstructions from a few images, saving time and resources. It could also be used in **AR** or **VR applications**, such as remodeling apartments, with users feeding images of furniture or other objects into the model to generate novel-view renderings and visualize how it would look in their living room.

After completing a literature review, Prune discovered that several works had already tackled sparse input views and found solutions to help with the overfitting problem. However, these works typically used perfect ground truth poses. There had also been works on a joint pose-NeRF refinement, but they assumed dense images.





"All these NeRF papers assume there are hundreds of images, but you don't want to have to take hundreds of images of an object," Fabian points out. *"You take a few, and then this problem becomes severe because the poses you get are usually bad, and you have both problems simultaneously. That's why* ***this is an important research direction for real applications!****"*

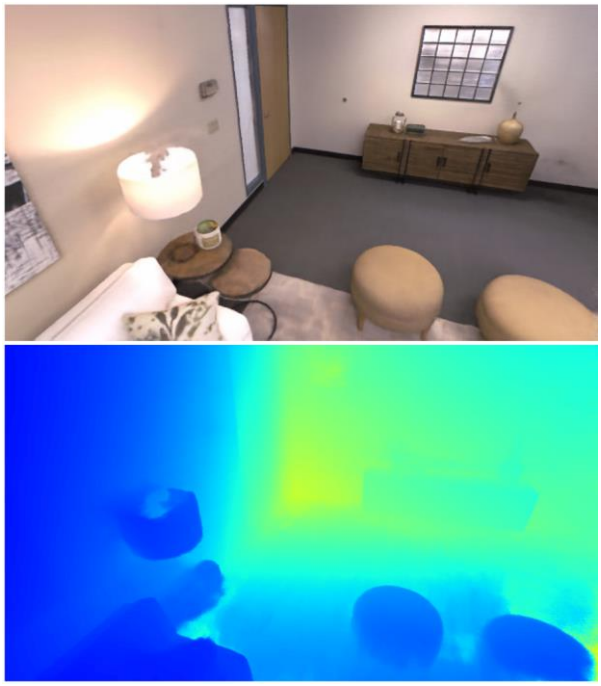
Prune agrees: *"For us, it was still interesting. Even though this joint pose-NeRF refinement didn't work for the sparse scenario, I started from this and worked my way up by adding new constraints."*

For the constraints, Prune took inspiration from **multi-view problems and bundle adjustments**, which are well explored in **classical computer vision and geometry**, and set about integrating them into the NeRF framework.

SPARF can train in a much shorter window, but the output is still far from the rendering quality that could be achieved with dense views. Prune is keen to point out that although they have made great strides, there is still a road ahead.

"There's still a lot of work towards getting perfect rendering from only two or three views," she tells us. *"We can do this joint pose-NeRF refinement much better than before, but we rely on point correspondences between the different views. The problem is getting these point correspondences is itself a research problem.* ***It would be great if there were a way to train or refine everything together.*** *For example, if you're training the NeRF and refining the poses based on the correspondences, could you also update the correspondences based on the NeRF and the poses? Like a system where everything can get better all at once. We haven't reached that point yet!"*

Fabian continues: *"It's a chicken and egg problem. COLMAP uses correspondences to get the poses, but you need the poses to get the correspondences.* ***We need a tool that***



does it all at the same time."

Could this be the seed for their next paper or perhaps open the door for other researchers?

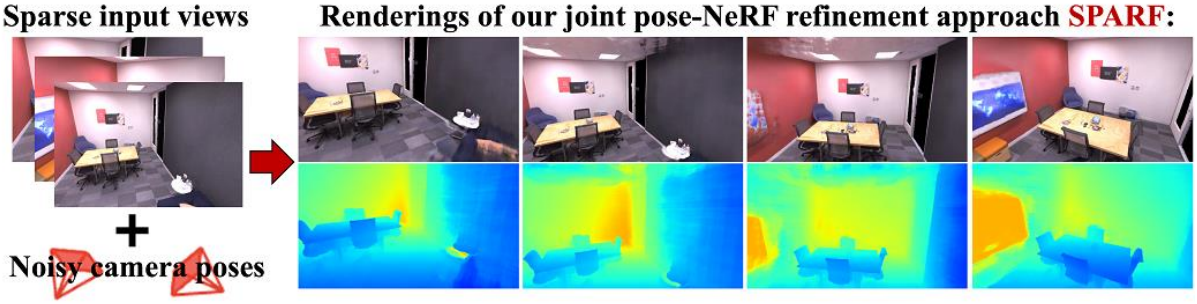
"*That would be pretty nice!*" Prune responds. "*We did a couple of experiments in that direction, and it's turning out more challenging than we hoped, but it would be a very nice output and future direction.*"

As the first author of the work, Prune will present at **CVPR 2023 in June**. She worked on the code during her internship at **Google**, where in the beginning, she was a newbie to the NeRF field and can remember being surprised at the results when the state-of-the-art methods were applied to sparse views.

"*I was like, how can it be so bad?*" she recalls. "*I'm most proud of the difference in outputs we can now get. That's because I saw how it was initially, and I thought, how are we ever going to be able to get something reasonable? Now, we get something that looks realistic.*"

Fabian adds: "*The most important thing is that rather than having to record for hours,* ***you can just take a bunch of images or crawl the internet.*** *That opens a lot of doors and makes it more accessible for everybody!*"

To learn more about SPARF, visit Poster Tue-PM-006 this afternoon from 16:30-18:30 in the West Exhibit Hall. It's a highlight paper (check it out).




CVPR DAILY
2022
Computer Vision and Pattern Recognition
Thursday
Workshop:
Computational Cameras and Displays
Presenting Work by:
Ruslan Partsey
Ozan Özdenizci
Dima Damen & team
Exclusive Interview with:
Mathieu Salzmann
Women in Computer Vision:
Anna Rohrbach
Today's Picks by:
Anh N. Thai
In cooperation with
Computer Vision News
The Magazine of The Algorithm community
A publication by
RSIP VISION
CVPR DAILY
2022
Computer Vision and Pattern Recognition
Tuesday
zebra
Workshop:
Medical Computer Vision
Presenting Work by:
Bowen Cheng
Ping Hu
Fabio Cermelli and Dario Fontanel
Editorial with:
Program Chairs
Women in Computer Vision:
Shuran Song
Today's Picks by:
Maria Dobko
In cooperation with
Computer Vision News
The Magazine of The Algorithm community
A publication by
RSIP VISION
CVPR DAILY
2022
Computer Vision and Pattern Recognition
Wednesday
False Positive
CVPR DAILY
2022
Computer Vision and Pattern Recognition
Friday
Don't miss the BEST OF CVPR 2023
in Computer Vision News of July.
Subscribe for free and get it in your mailbox!
Click here

Topology-Guided Multi-Class Cell Context Generation for Digital Pathology

Shahira Abousamra is a senior PhD student at Stony Brook University, advised by Chao Chen and Dimitris Samaras, and working closely with Joel Saltz's group.

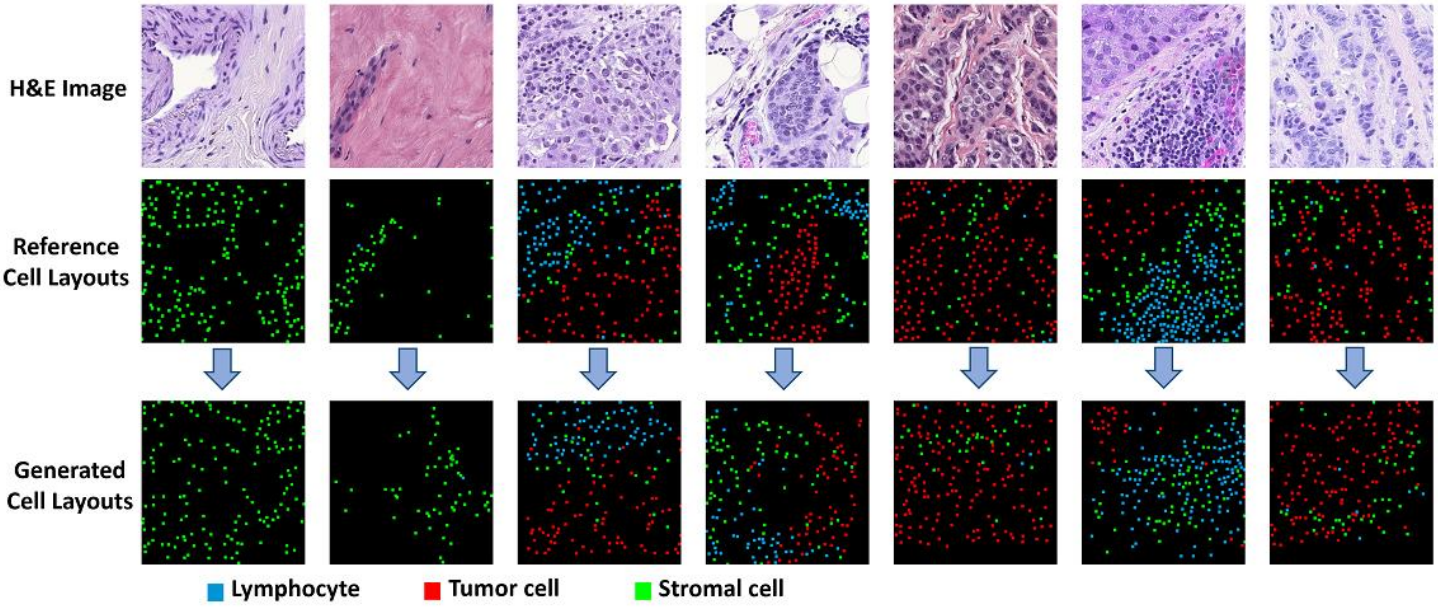
She speaks to us about her work on pathology data ahead of its poster this morning

Pathologists analyze cell samples to determine the presence and type of cancer, assess its stage and aggressiveness, and evaluate patient responsiveness. However, **automating these tasks is challenging** due to the limited data available on the tumor microenvironment. Unlike natural images, which are abundant and anyone can annotate, pathology data, like all biomedical data, requires **patient information and expert annotation**. In overcoming this limitation, this work aims to create a generative model capable of producing synthetic labeled data to augment training and facilitate downstream tasks.

"*We realized that when pathologists look at an image or data from pathology, they look at the distribution of different types of cells, how they colocalize, and the patterns they make, which has not been considered in a generative model before,*" Shahira tells us. "*We're trying to address this by* ***generating realistic cell layouts that capture this biology****. With a reference sample, you can generate new samples that exhibit the same spatial and structural patterns or characteristics and use them efficiently in training for downstream tasks. This paper is about how to model this kind of tumor microenvironment, the cell layout, and these patterns. Given some specific characteristics,* ***how do we train the model to generate data that satisfies these characteristics?***"

Shahira's approach differs from existing work on generating synthetic data in several ways. Previous methods focused on creating visually appealing images, whereas this paper takes it back to the beginning, **focusing on the location and distribution of different cell types**. The spatial organization and

and arrangement of cells are crucial in pathologists' decision-making process. This aspect has not been incorporated into deep learning models before and is usually saved for analysis after data generation.



Shahira introduces a novel approach by integrating these patterns into the deep learning process, developing a model that could **effectively capture the complex structure of the tumor microenvironment** and satisfy the desired conditions.

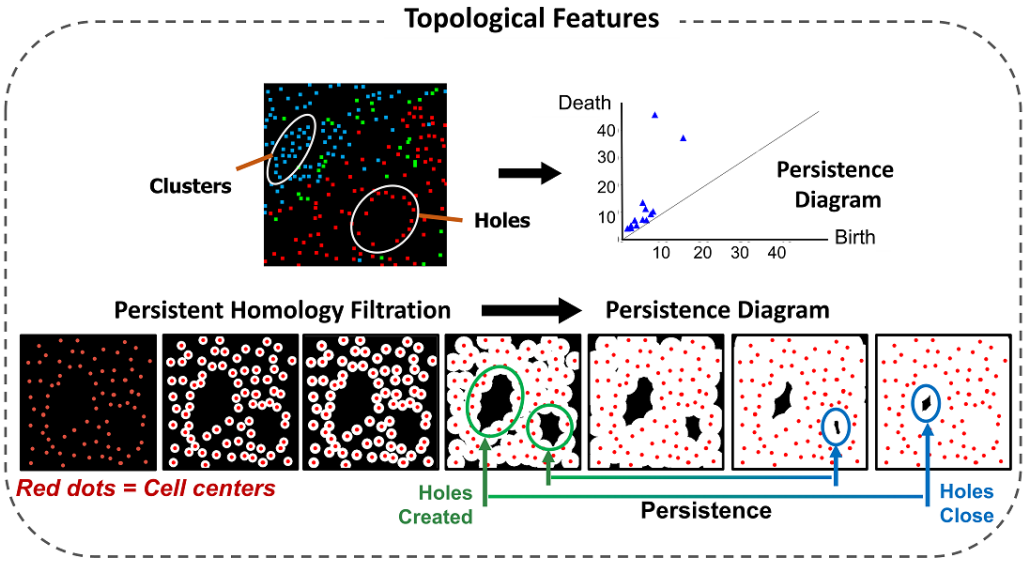
"*First, we look at the cell layout as **a point map of different classes**,*" she explains. "*We see these different distributions that we want to capture. We also see that they form some clusters, holes, and gaps, and want to capture them all. We can capture the spatial colocalization using spatial statistics like cross K-functions, which we've used successfully in a previous cell classification paper. We talked to pathologists, who told us that **when they classify a cell, they don't look at just one but at the whole region**. We didn't want the model to be fixed on the morphology and texture of just one cell but to have a larger context and used the cross K-function to model this spatial context in the image.*"

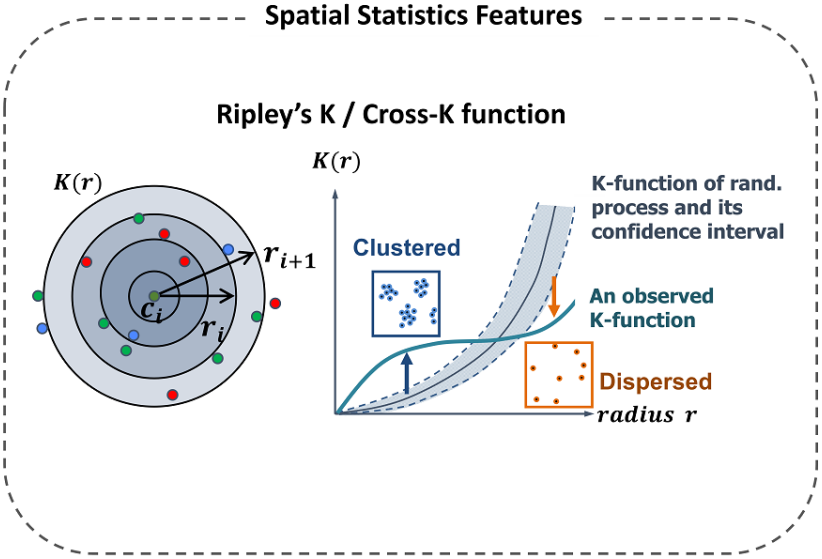
Realizing that clusters, holes, and patterns are like topological features, Shahira used the **persistent homology algorithm**, a topology data analysis algorithm, to capture these characteristics and model the cell layout. Another challenge was how to train the model, given these features, to generate data that satisfies these spatial and topological characteristics.

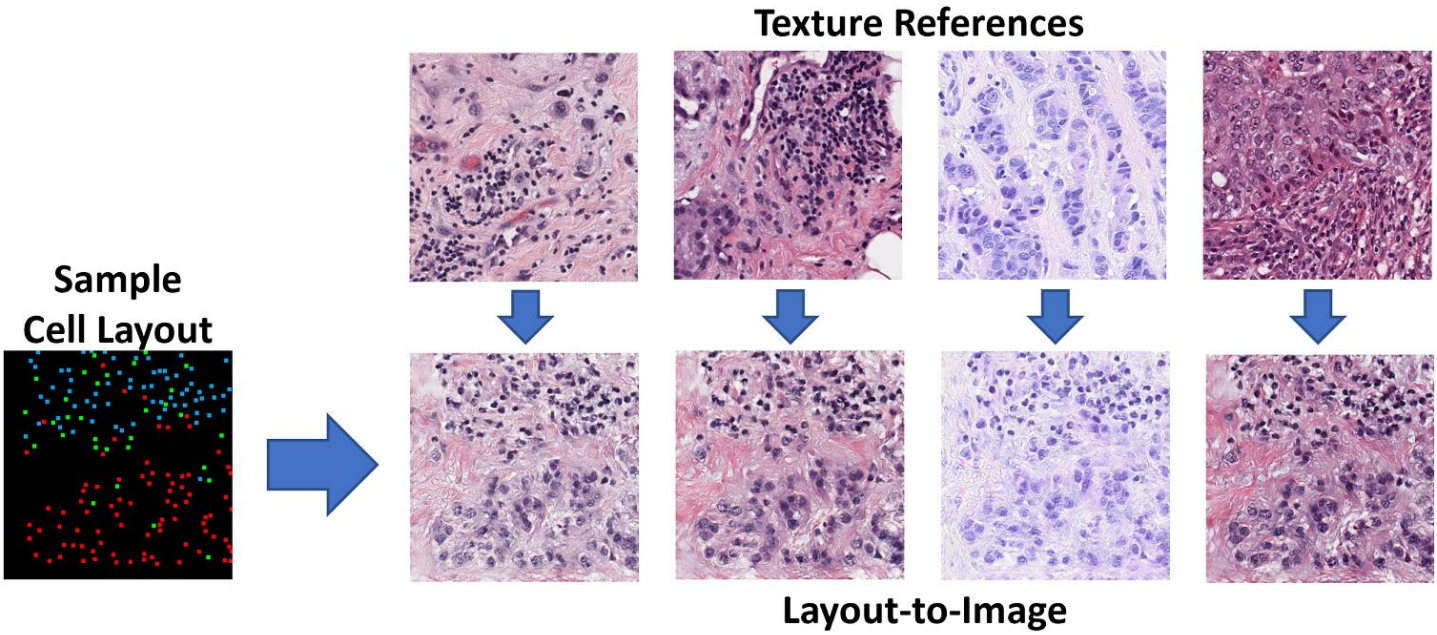
She attempted to use **adversarial learning** and different formulations with little success. Recognizing the need for an alternative approach, she discovered

that directly manipulating the points within the data was crucial to achieving the desired outcomes, and it was necessary to establish a correspondence between the generated data and the reference data to accomplish this. When observing an image composed of points, specific characteristics, such as holes, become apparent. These holes can be matched based on their size, the distribution of different cell types, or the contextual information surrounding them. Shahira employed **a matching technique based on persistence**, which correlated with a cell's size and the cross K-function around the holes. This matching approach established the desired spatial context.

"*Now that we have a correspondence, we want the matched or paired locations to have a similar distribution of cells around them,*" Shahira points out. "*We want **a loss function** that allows us to move points around to affect the points in general. We created **multi-scale density maps** from the point maps for each class and tried to minimize the distance at the matched locations. This way, you will bring points together at the paired locations or move them apart to have the same density values.*"







This process proved a success and was necessary because the training set, the reference data with the cell points classification, was also very small at around 100 images. Shahira required a dataset with this location and classification of cells and with enough context in it. Most datasets have small patches and do not provide the context needed to capture the spatial structures and distributions.

This paper is the first step toward further work to support other downstream tasks, which will help to propel this model into real-world application. Furthermore, it can enhance our understanding of the tumor microenvironment, modeling its complex structures and analyzing and interpreting different characteristics to gain insights into tumor growth. It can also contribute to representation learning.

Shahira hails from Alexandria in Egypt, the largest city on the Mediterranean coast, but has lived in America for seven years. What led her to leave the hot desert climate behind for the frenzy of New York?

"*I did my undergrad, and then I did a master's degree in computer architecture, and I realized I didn't like it!*" she reveals. "*I was working in a software company for a while and then came across* ***computer vision and AI****, and I was so excited. I thought, how do I get to do this? The only way was to go somewhere and start studying it.* ***I'm now in the final year of my PhD and looking for a postdoc position!***"

Shahira is a great catch! To learn more about her work, visit Poster 316 this morning from 10:30-12:30 in the West Exhibit Hall.

DiGeo: Discriminative Geometry-Aware Learning for Generalized Few-Shot Object Detection

Jiawei Ma is a fourth-year PhD candidate at Columbia University, advised by Shih-Fu Chang.

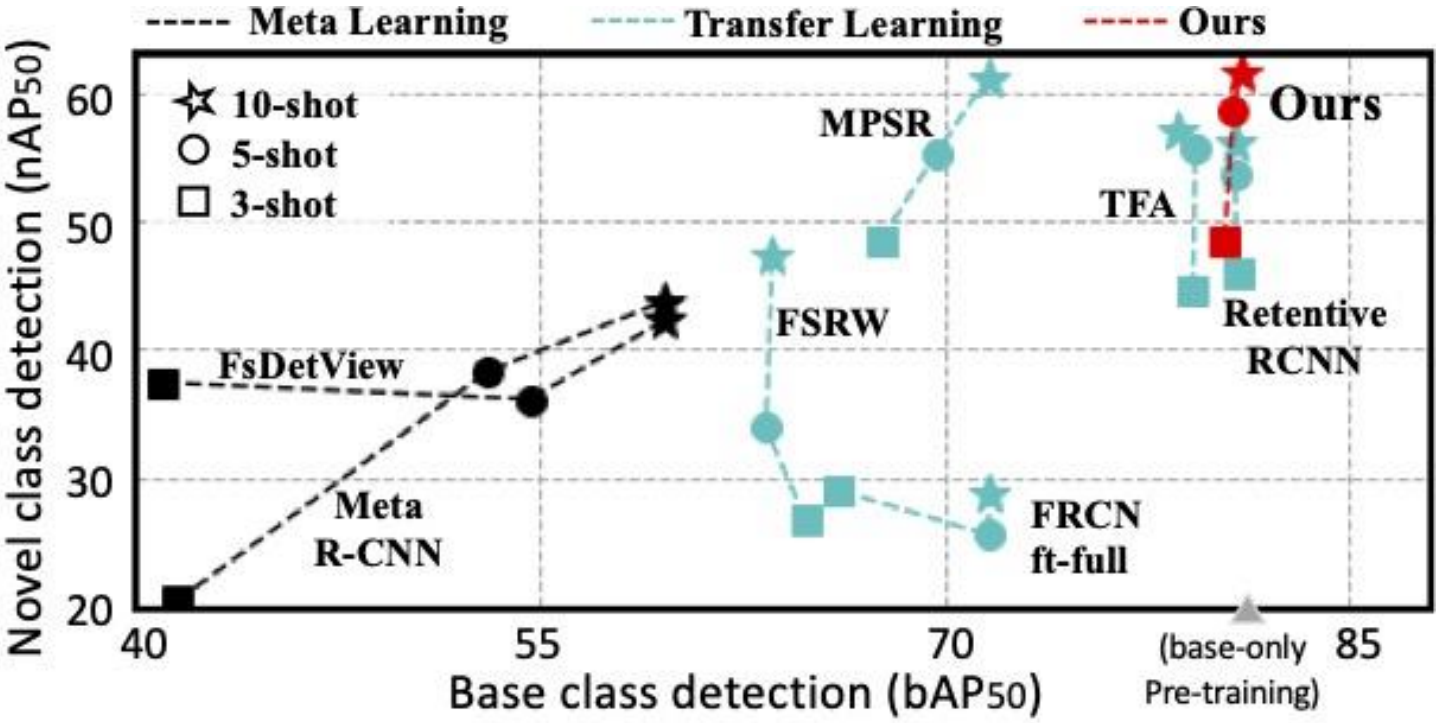
He proposes a new training framework for generalized few-shot object detection tasks and speaks to us ahead of his poster this morning.

The primary purpose of **generalized few-shot object detection tasks** is to train a detector to detect the instances from base classes, which have a lot of training data, and novel classes, which have limited training data.

TFA is a widely-used framework in generalized few-shot object detection tasks. Models are pre-trained on data from base classes and then fine-tuned on a union of base and novel classes. **This fine-tuning is called few-shot adaptation**. An essential part of this second step is aggressively down-sampling the base training set to achieve a balanced training set among the base and novel classes.

"*Few-shot adaptation is key to the success of this framework, but it also causes a problem,*" Jiawei tells us. "*By aggressively down-sampling the training data of the base classes, **it sacrifices the detection precision**. With limited training data, the model will overfit to those few samples of base classes. There is always a trade-off. **We achieve good performance on the few-shot novel classes at the expense of the detection precision on the base classes**.*"

You can only perform a good few-shot novel adaptation or detect the instances of novel classes correctly when you have **well-separated classifier weights**. Down-sampling is crucial because only when the model is trained on a balanced data set will the classifier weights be maximally separated.

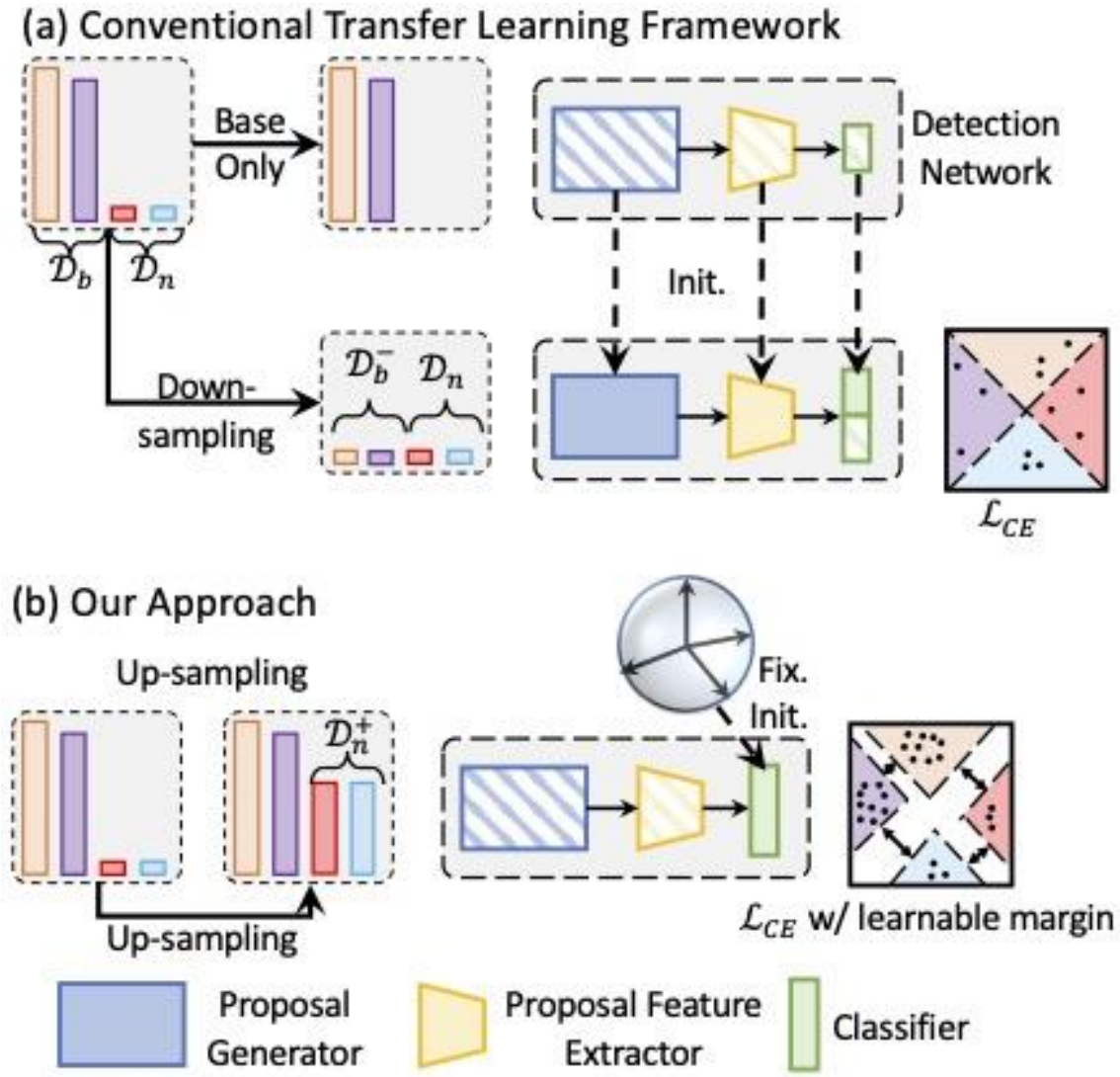


Generalized few-shot object detection is a powerful technique that can improve the performance of object detection models. For example, in **self-driving cars**, the algorithm is trained to detect a wide range of objects and environments. However, in the real world, there is not enough training data to model every object and environment the car may encounter, so **users may meet instances that the model has not been trained on and needs to adapt for**. After the adaptation, the model must retain the ability to precisely detect the original pre-trained objects.

Long-tail object detection is another relevant scenario. For example, in the cat species, some object classes, such as domestic cats, are particularly prevalent. In contrast, others are rare, such as tigers. In these cases where collecting enough training data for all classes may not be possible, generalized few-shot object detection can help.

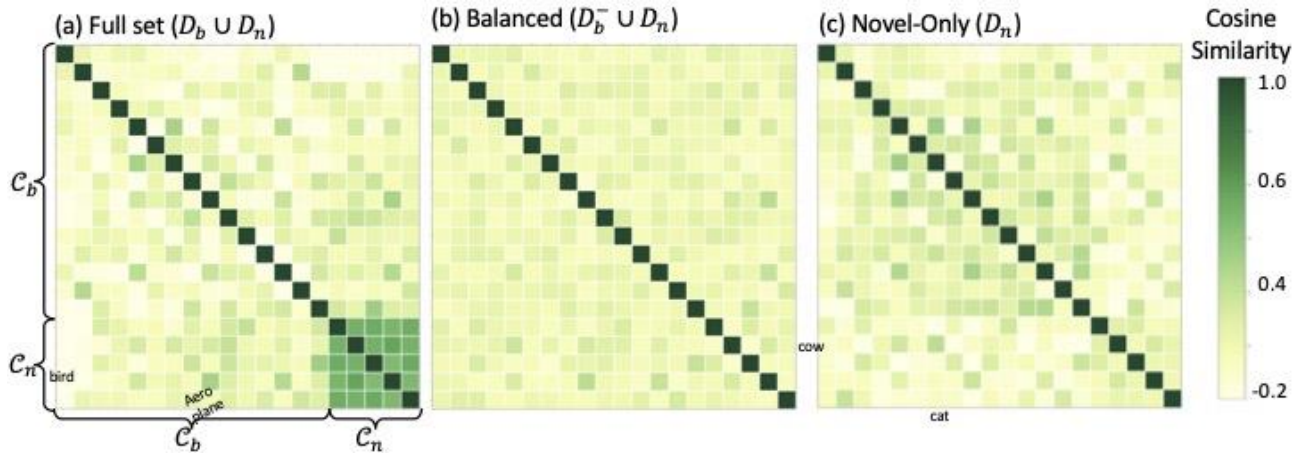
Jiawei tells us this work has not been without its challenges, which have covered two aspects: the technical and the conceptual.

"*On the technical side, how can we ensure the well-separated classifier weights after those experimental observations?* ***How can we ensure a perfect geometry of the expected feature distribution?*** To have a good object detector for the base and novel classes, we must carefully design the feature space's geometry", he poses. "*On the conceptual side, because this TFA framework is popular and no one has ever studied it, there is little related literature.*"



Regarding the computer vision techniques involved in this work, Jiawei identifies that the aim was to **maximize both the inter-class separation and the intra-class compactness**, also called the feature distribution's geometry.

"*For maximizing the inter-class separation, we use a classifier called* ***simplex equiangular tight frame****,*" he explains. "*Simplex ETF is offline-derived and fixed during the training. One property of this classifier is that the weights are all maximally and equally separated in the entire feature space. From the perspective of maximizing intra-class compactness, many different approaches can be used, but we use the simplest one to highlight the importance of this conceptual idea. We do that by adding margins during the training. For conventional cross-entropy loss, we directly minimize the difference between the probability distribution predicted by the model and the ground-truth vector. We add margins in the conventional cross-entropy loss to force the features close to the corresponding class center.*"

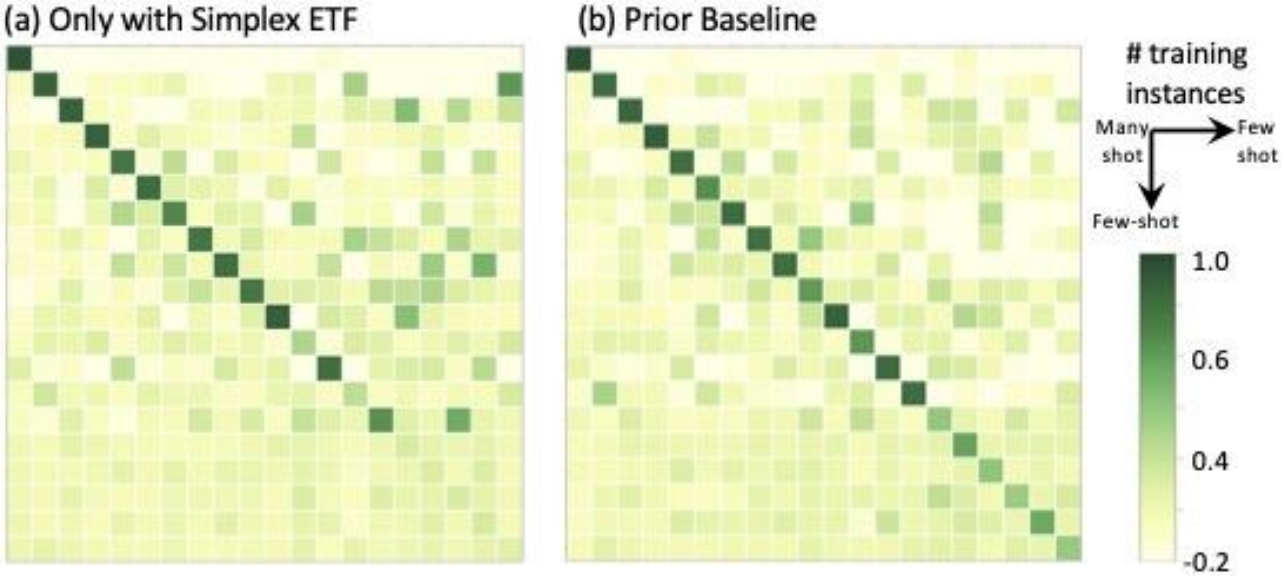


Jiawei hopes this work will encourage researchers to **understand feature distribution better** when performing any task. It is essential because any operation should be mathematically analyzed and clearly understood as to why it is helpful for the task. In this paper, he analyzes inter-class separation and intra-class compactness to discover why certain aspects of the conventional framework are essential.

He would encourage researchers to focus on **bias theory in conventional training, especially deep learning**.

"*We always encounter **long-tail datasets**, and this imbalance can be interpreted differently,*" he adds. "*It can be the number of training data, it can also be the times that this environment happens, and we would like to encourage the researchers to focus on the bias toward this understanding and then use the analysis on the feature geometry, which is also called the distribution of features, to develop their projects.*"

To learn more about Jiawei's work, visit Poster 305 this morning from 10:30-12:30 in the West Exhibit Hall.

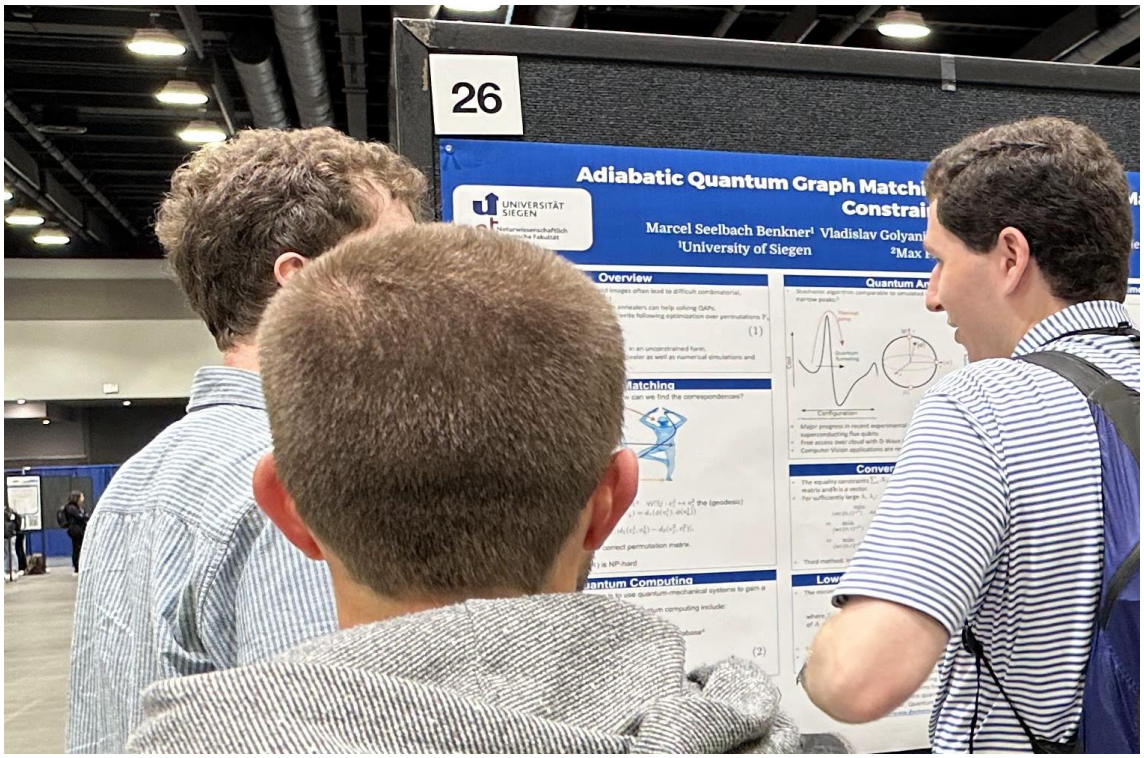


by Tolga Birdal

Tolga Birdal is an assistant professor (Lecturer) in the Department of Computing of Imperial College London.

Today's computer vision involves processing large data volumes to satisfy combinatorial optimization objectives requiring excessive computational power. To date, the GPU hardware has fulfilled these high data processing demands and fueled rapid advancements within the field. Yet, **training widespread deep learning models using GPUs requires extensive amounts of resources** due to the prolonged GPU-utilization periods stretching from weeks to months.

Confronted with such challenges, the tiny yet evolving community of **Quantum Computer Vision and Machine Learning (QCVML)** has asked: **Is there a more sustainable path forward for computer vision?**



Acknowledging the need for a radical shift in computational paradigm, we have looked into recent breakthroughs in **quantum computing**. Modern **quantum computers (QCs)**, capable of leveraging quantum phenomena like superposition, entanglement, and tunneling, are no longer limited to simulations. This opens the door to the exciting field of **Quantum Computer Vision (QCV)**, which aims to transpose existing computer vision problems into a framework suitable for quantum computation. However, leveraging quantum hardware to tackle complex vision tasks presents unique challenges:

- How can we **adapt current CVML algorithms** to function on QCs?
- How can we **create hybrid solutions** maximizing the strengths of CPUs, GPUs, and QPUs?
- How can we devise **scalable divide-and-conquer algorithms** for optimal QPU utilization?
- How can the process of **quantum implementation inform CPU implementations**?

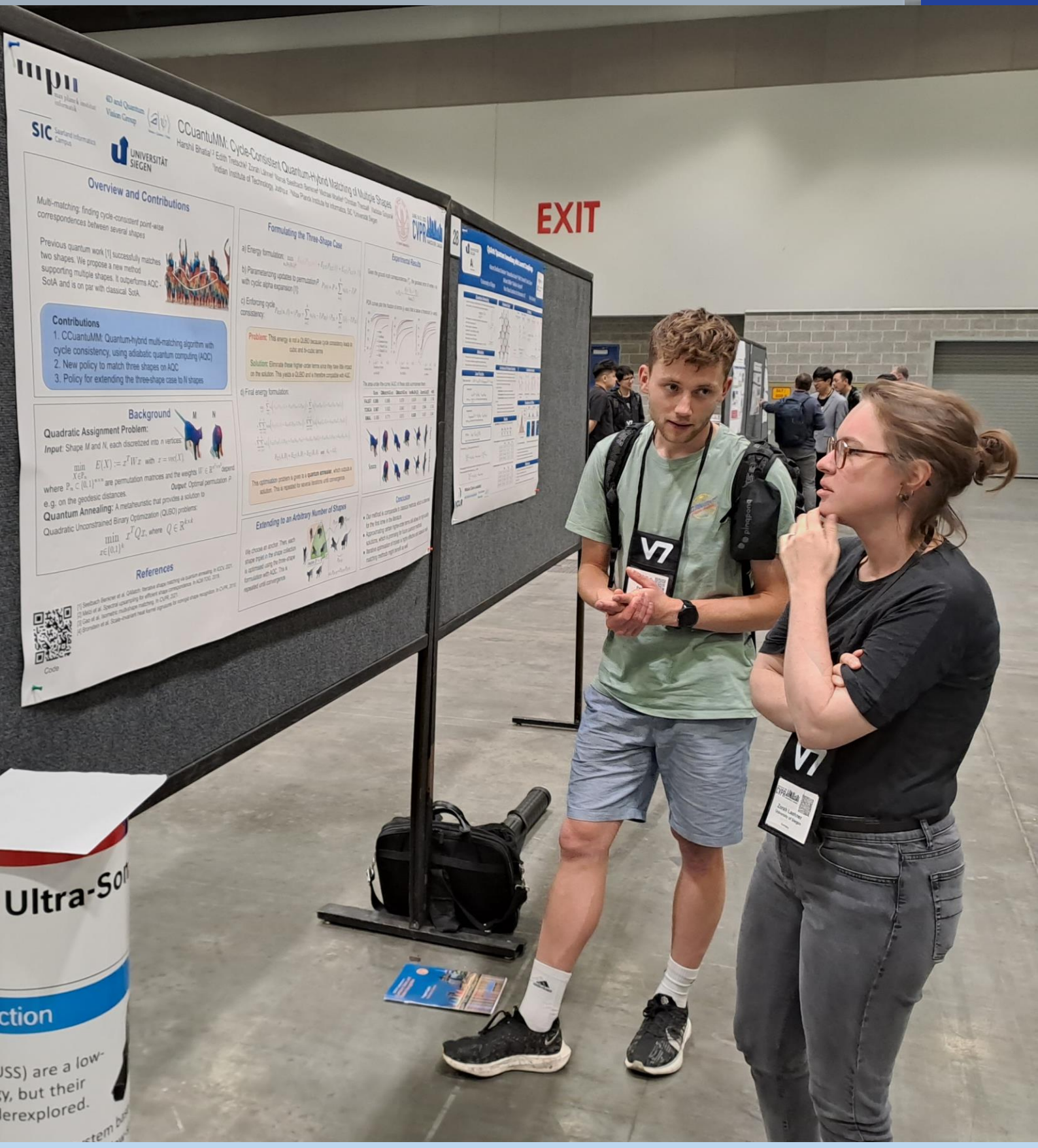
During our **CVPR workshop**, Q-CVML, three core themes were discussed through excellent invited talks: (i) Adiabatic quantum computing, presented by **Michael Möller** from **MPI** and **Victoria Goliber** from **D-Wave**, (ii) Gate-based quantum computing, discussed by **Roberto Bondesan** from **Imperial College** and **Tat-Jun Chin** from **University of Adelaide**, and (iii) Quantum-inspired computer vision, presented by **Anand Rangarajan** from **University of**

Florida. Given the pace of development in these areas, we anticipate a synergetic co-evolution that will mutually inform and inspire each other.

The panel discussion revealed that both our speakers and the organizers viewed QCV as more than just an efficient computational approach. By 'quantum-izing' existing algorithms, we are inevitably crafting new perspectives on classical computation. Such a mindset liberates us from being solely bound by the constraints of current hardware. As we've seen in the past, hardware technology will inevitably advance. And even if it doesn't keep pace, the process will yield fresh insights into problem-solving. This intellectually stimulating journey, driven by curiosity, not only propels us forward but also encourages us to **question and reassess our current methods of solving vision problems**.

We asked **Federica Arrigoni**, an Assistant Professor at the **Politecnico di Milano in Italy** a comment about the event: "*The organisers have done an excellent job in putting together many inspiring talks from renowned researchers in the field,*" she said "**covering both relevant applications of quantum computing in computer vision as well as an introduction to the field with high educational value!**"

This was the first workshop on Q-CVML. We believe it was timely and the participation proved us right. The workshop hosted some intriguing discussions and excellent questions. A set of resources such as poster slides could be found on the workshop website.



Zorah Lähner (right), a Postdoctoral Researcher at Universität Siegen, explaining her work on how to match multiple shapes cycle-consistently with quantum annealing, as part of the QCVML workshop. Research was led by first author Harshil Bhatia. To learn more about this work, visit Poster 123 this morning from 10:30-12:30 in the West Exhibit Hall.

Andrei Bârsan, a PhD student at the University of Toronto under the supervision of Raquel Urtasun, and a Senior Research Scientist at Waabi, presenting at the tutorial All You Need To Know About Self-Driving.

JUNE 2023

Computer Vision News

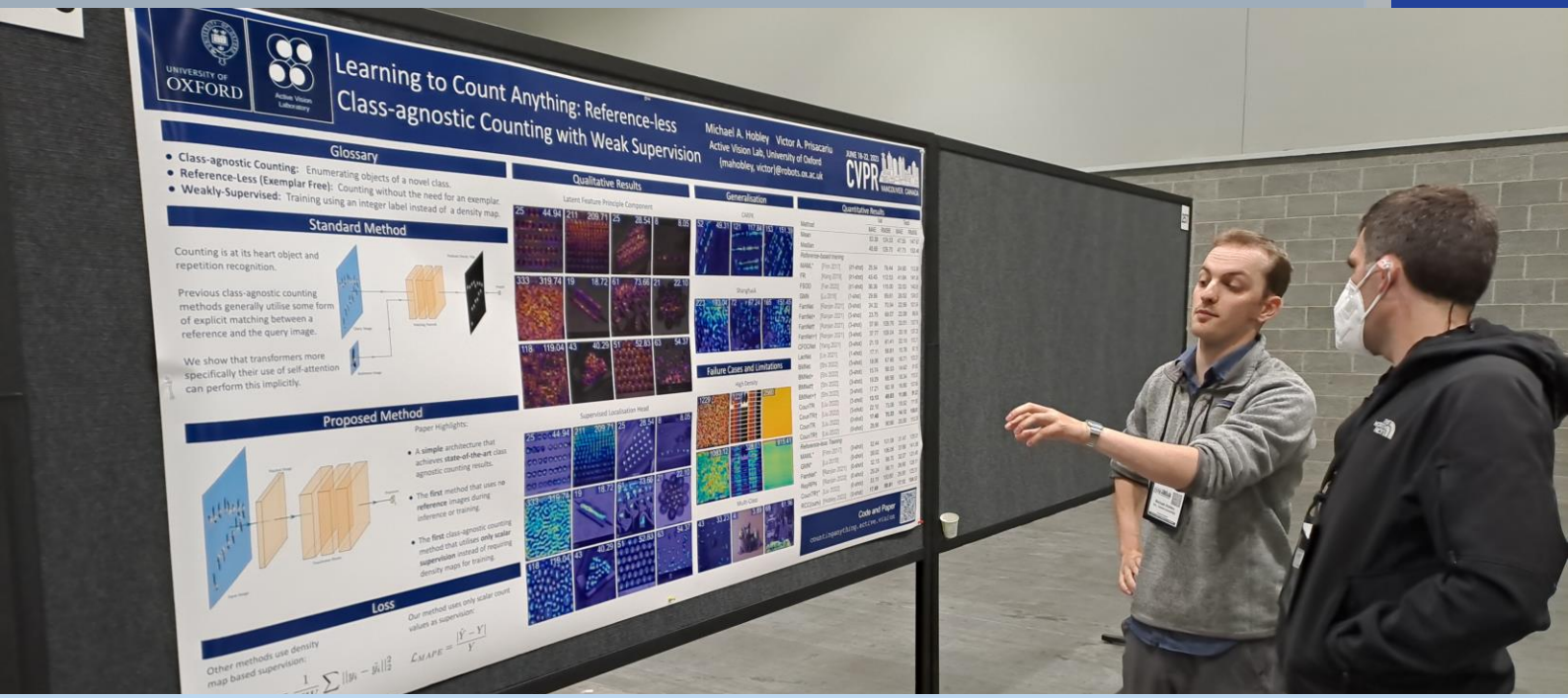
& Medical Imaging News

The Magazine of the Algorithm Community

Did you read Computer Vision News of June?

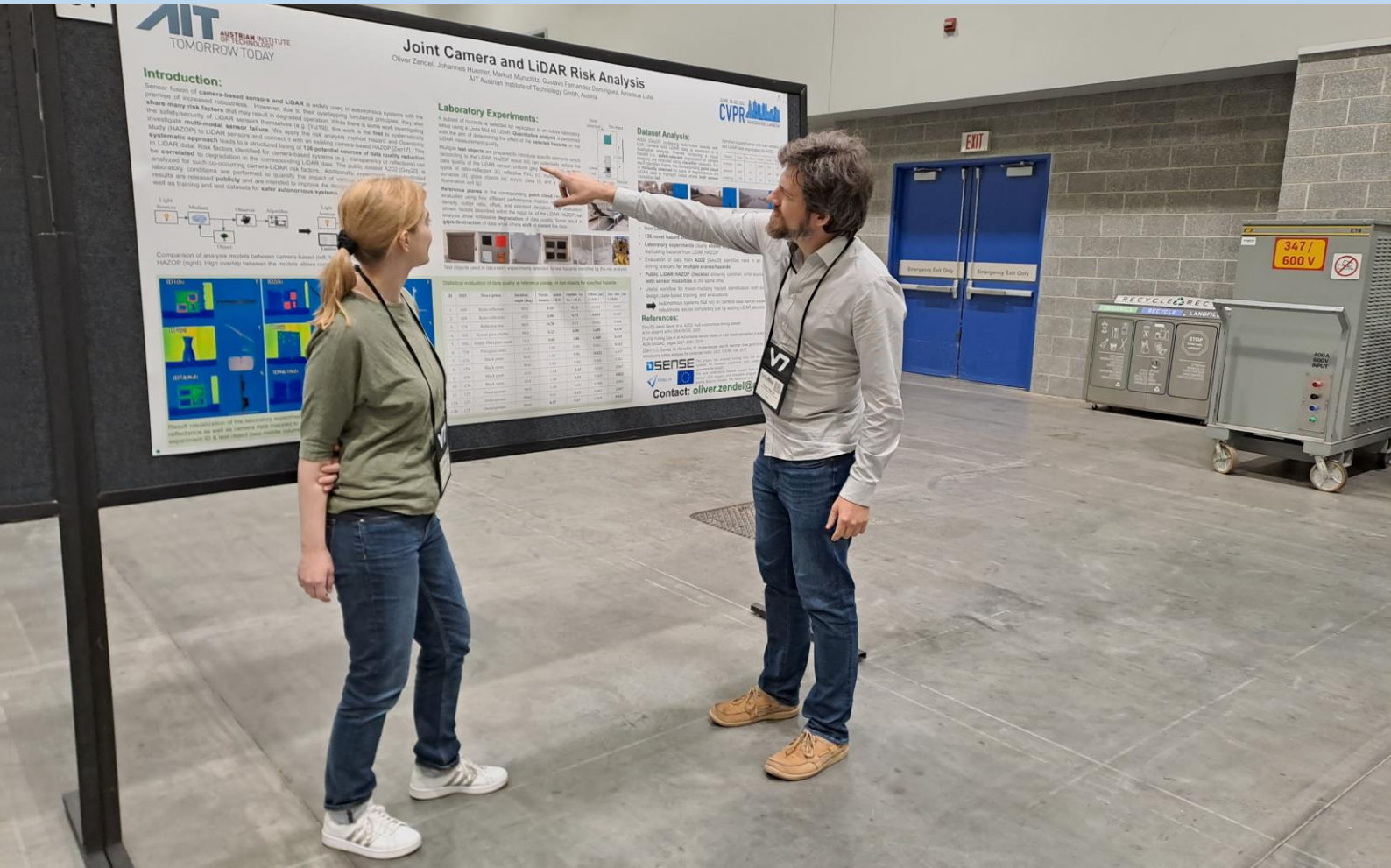
Read it here

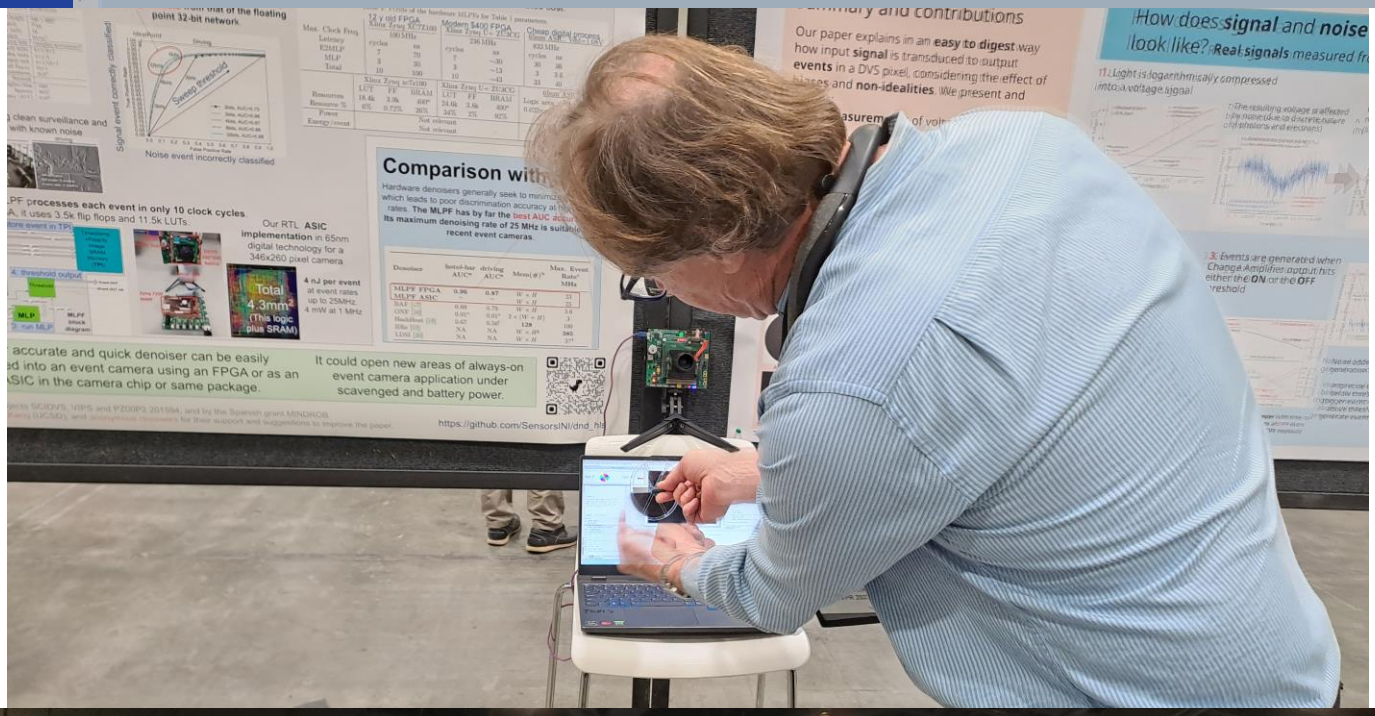
What an impressive panel at the Generative Models for Computer Vision workshop: from left, Angjoo Kanazawa, Björn Ommer, Angela Dai and Andrea Tagliasacchi.

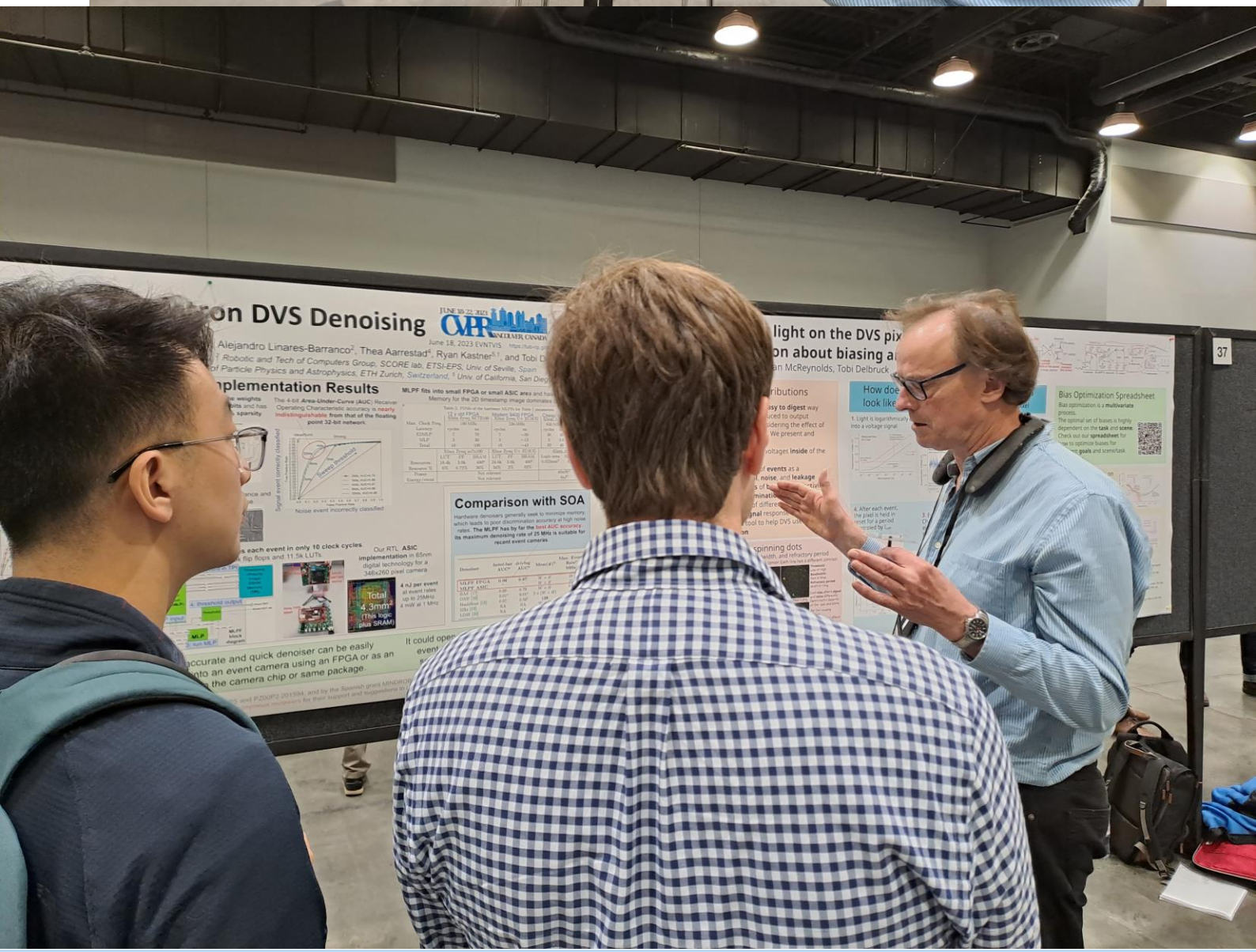


Michael Hobley (top left), a PhD student at the University of Oxford, is presenting his poster on counting objects of previously unseen classes without the need for examples of type or spatial supervision.

Oliver Zendel (bottom right), a scientist at AIT - Austrian Institute of Technology, is pointing at his work "Joint Camera and LiDAR Risk Analysis“ during the Workshop on Autonomous Driving (WAD).







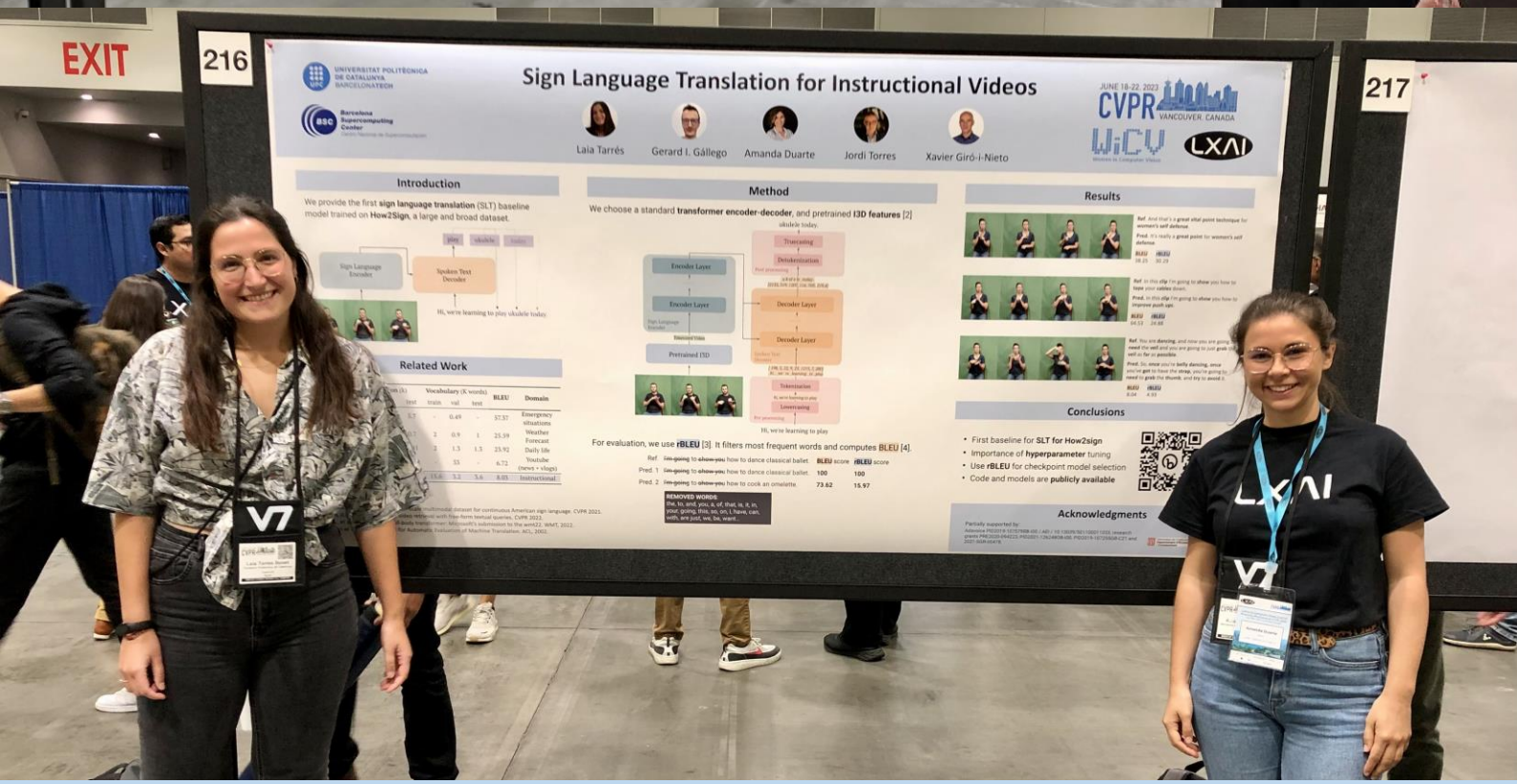
Tobi Delbrück (top and right), a professor of physics and electrical engineering at the University of Zurich and ETH Zurich, presenting his posters and giving a live demonstration.

Russian Invasion of Ukraine

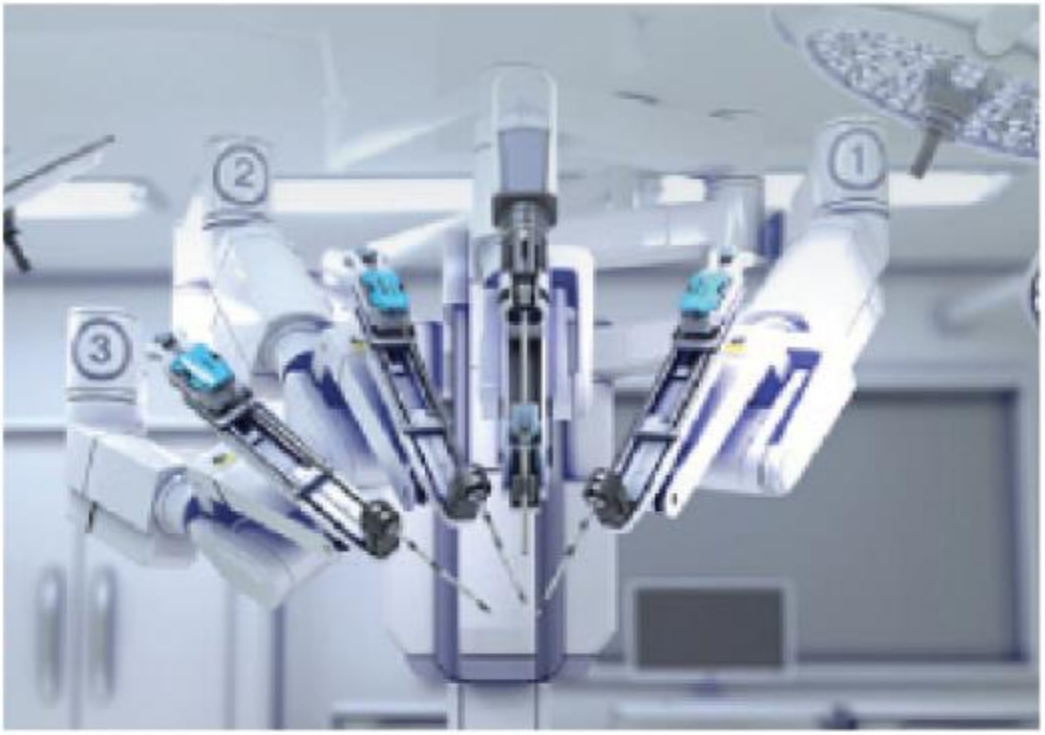
CVPR condemns in the strongest possible terms the actions of the Russian Federation government in invading the sovereign state of Ukraine and engaging in war against the Ukrainian people. We express our solidarity and support for the people of Ukraine and for all those who have been adversely affected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

Denys Rozumnyi is a final-year PhD student at ETH Zurich and a researcher at CTU in Prague. His main research topics are 3D reconstruction and deblurring, especially of highly motion-blurred objects. If you're interested in his work or want to talk to him, please come to the poster 122 this afternoon. Photo: Matias Valdenegro

Lucas Ventura, a PhD student in the IMAGINE computer vision team at Ecole des Ponts ParisTech (ENPC) and Inria (Paris), supervised by Cordelia Schmid and Gül Varol, presented his poster about text-to-video retrieval without access to manually-labeled videos. The model is trained using automatic frame captions, which constitute free labels for supervision. Work was presented at the Workshop on Learning with Limited Labelled Data for Image and Video Understanding.



Amanda Duarte (right), a Postdoctoral Researcher at the Barcelona Supercomputing Center in Spain, and Laia Tarrés (top and left), a PhD student at Universitat Politècnica de Catalunya, presented their work on Sign Language Translation at the Women in Computer Vision workshop and at the LatinX in AI workshop.



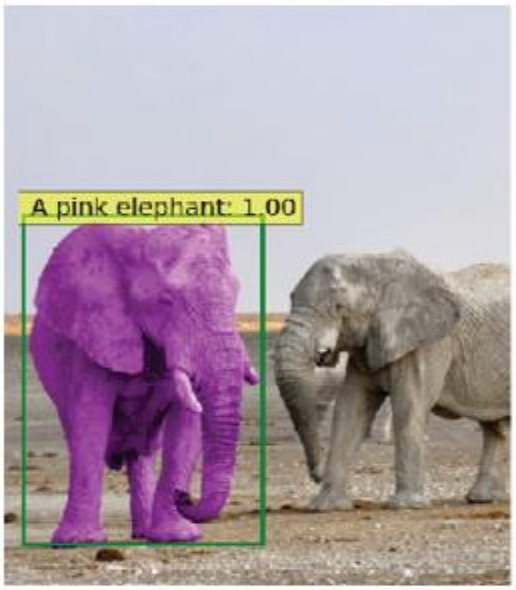
IMPROVE YOUR VISION WITH

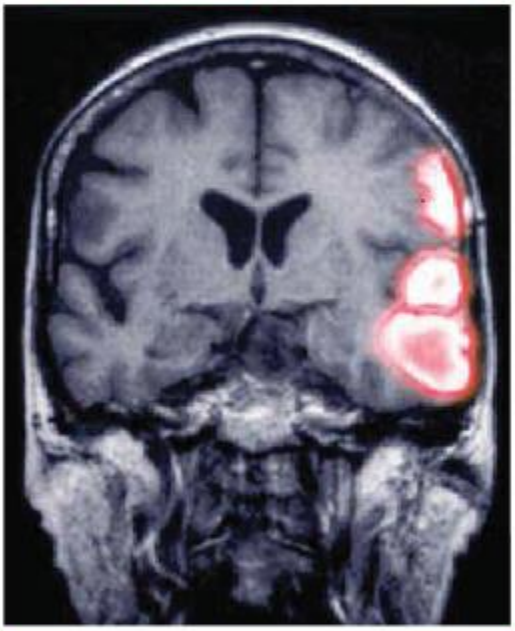
Computer Vision News

SUBSCRIBE

to the magazine of the algorithm community

and get also the new supplement Medical Imaging News!





RSIP VISION